Vrije Universiteit Brussel

Faculteit Wetenschappen
Departement Natuurkunde

Sharif University of
Technology

Physics Department

Measurement of the b-tagging efficiency in the CMS experiment with the first LHC collisions

Abideh Jafari

Promotors
Prof. Dr. Jorgen D'Hondt
Prof. Dr. Farhad Ardalan

Thesis submitted in order to obtain the
academic degree Doctor of Science

September 2011

Contents

Introduction

The Standard Model of particle physics describes the fundamental interactions other than gravity between the elementary particles. The model has successfully achieved to describe the physics observations as well as to predict the existence of particles such as the top quark, discovered in 1995.
An important yet undiscovered particle introduced by the model is the Higgs boson which is believed to be responsible for the electroweak symmetry breaking, the mechanism by which the particles become massive. Besides, there are also observations, not explained by the model, that lead to the development of other theories beyond the Standard Model. The Large Hadron Collider (LHC) is built at CERN aiming to search for the Higgs boson and to investigate new theories at the TeV scale.
Different experiments including the Compact Muon Solenoid (CMS) experiment are established to analyze the data produced by colliding proton beams at the LHC. For many data analyses, the jets originating from b-quarks play a key role in the identification of the physics signal as well as in rejecting background processes. Therefore, different b-jet identification algorithms exploiting the distinct experimental signatures of the b-quark jets are developed in CMS. It is crucial however to investigate the performance of these algorithms using data driven methods.
In this thesis, a data driven method is applied to measure the performance of the b-jet identification algorithms using top quark events. The very high production rate of top quarks at the LHC and the almost exclusive decay of the top quark into a b quark and a W boson provides a rich sample of b-quark jets which is well suited for the performance measurement purposes. At the LHC the top quark is produced mostly in pair where in this thesis the $t\bar{t} \rightarrow qq'b\bar{b}e\nu_e$ decay channel is considered. The method is applied on the first proton collision data collected in 2010 by the CMS experiment.
In Chapter 1 the Standard Model of particle physics is reviewed where the role of the top quark in the Standard Model and beyond is briefly discussed. The method detailed in this thesis is applied on the LHC collision data collected by the CMS. Hence Chapter 2 is dedicated to an introduction on the Large Hadron Collider and the Compact Muon Solenoid detector. The analysis strategies are designed and developed using simulated collision events. A review of the simulation is given in Chapter 3. The reconstruction of the physics objects interacting with the CMS detector material, with the focus on the electrons and jets, is explained in Chapter 4. It is also detailed in Chapter 4 how the electron selection efficiency is measured in data. For the data collected in 2010, the efficiency ratio between the data and simulation is reported. The ratio is to be used in the $t\bar{t}$ cross section measurement for which the presence of an additional systematic uncertainty is also discussed.

Chapter 5 is devoted to the selection of $t\bar{t}$ events and the event topology reconstruction together with the detailed discussion about the b-jet identification efficiency estimation in the selected event sample. In the same chapter, the method is applied for the first time on the LHC collision data collected by the CMS detector in 2010 and the results are reported.
Chapter 6 contains the summary where prospects for an extension of the method towards a simultaneous $t\bar{t}$ cross section and b-jet identification efficiency measurement are proposed.

Chapter 1

The top quark of the Standard Model

1.1 The theoretical background

Developed by Glashow in 1961, Weinberg in 1967 and Salam in 1968, the Standard Model of particles physics [1, 2] describes the electromagnetic and the weak interactions in a unified picture. Mathematically, the model is a gauge theory explained by the direct product of $SU(2)_L \times U(1)_Y$ where the L index indicates that the weak interaction occurs between the left handed particles. The $U(1)_Y$ part accommodates the electromagnetic-like interactions where the index Y is the so-called hypercharge of the interacting particle. Only the massless particles contributes in this picture. Hence, the model is complemented with a Spontaneous Symmetry Breaking (SSB) mechanism known as the Higgs mechanism[1], to describe the real world with massive particles. The SSB mechanism leads to the presence of a new massive particle, called the Higgs boson, which has not been discovered yet. The color interaction between quarks which is described by a non-Abelian $SU(3)_c$ gauge theory can also be added to the electroweak interactions resulting in a gauge group structure of $SU(3)_c \times SU(2)_L \times U(1)_Y$ before symmetry breaking. Table 1.1 contains the list of the Standard Model elementary particles together with their masses where the mass units are expressed in Planck units for which $\hbar = c = 1$.

1.1.1 The electroweak model

In theory, the particles are represented by fields, $\phi(x)$, where x is the space-time coordinate. This notation helps to describe the physics processes mathematically. The amplitude of processes like the weak muon decay,

$$\mu \rightarrow e + \nu_e + \nu_\mu$$

[1] Proposed and developed by Higgs, 1964, 1966; Englert and Brout, 1964; Guralnik, Hagen and Kibble, 1965; Kibble 1967.

	fermions			bosons	interaction
	I	II	III	(force carriers)	type
quark sector	u 2.5 MeV	c 1.29 GeV	t 172.9 GeV	γ massless	electromagnetic
	d 5 MeV	s 100 MeV	b 4.19 GeV	g massless	strong
lepton sector	ν_e < 2 eV	ν_μ < 0.19 eV	ν_τ < 18.2 eV	Z^0 91.19 GeV	weak
	e 0.511 MeV	μ 105.6 MeV	τ 1.777 GeV	$W^\pm$ 80.4 GeV	weak

Table 1.1: Elementary particles in Standard Model and their masses [3]. The fermions (spin-half) are listed in three generations (I-III) where the mass is the main difference between generations. Fermions in the quark sector are influenced by all forces. In the lepton sector, neutrinos only interact weakly where the rest are affected by the weak and electromagnetic forces. The interaction type of each gauge boson (spin-one) is indicated in the last column.

carried out only by the left-handed leptons or right-handed anti-leptons, can be written in terms of charged currents,

$$J_\alpha(x) \equiv J_\alpha(x)^\pm = \nu_{e,L}(x)\gamma_\alpha e_L(x), \tag{1.1}$$

where[2] $e_L(x)$ and $\nu_{e,L}(x)$ are the fields of the electron and its neutrino, respectively. The γ_α element is α'th member of the Dirac matrices, used to transform the spinors (see Appendix A). To describe the whole interaction, this current is further coupled with a gauge field, A^α, that represents the gauge boson mediating the weak force. Equation 1.1 suggests that $e_L(x)$ and $\nu_{e,L}(x)$ could be arranged in such away to make a doublet

$$L = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L$$

associated with the $SU(2)$ group. In this "isospin" notation, the electron (neutrino) receives an isospin charge of $-(+)\frac{1}{2}$ (see Appendix A). The current in Equation 1.1 is charged and coupled with a charged gauge boson. One can define the $\tau^\pm$ operators using the 2×2 Pauli matrices, τ^i's, which are the generators (see Appendix A) of the $SU(2)$ group and rewrite Equation 1.1 in a compact format,

$$J_\alpha(x)^\pm = \bar{L}\gamma_\alpha\tau^\pm L, \tag{1.2}$$

where

$$\tau^\pm = \frac{\tau^1 \pm i\tau^2}{2}. \tag{1.3}$$

[2] The real combination is $\bar{\nu}_{e,L} - e_L$ or $\bar{e}_L - \nu_{e,L}$. This is not indicated to avoid too many signs in the equation.

The third generator of the group, τ^3, induces a neutral current hence an interaction between the left-handed fermions [3] via a neutral gauge boson. Although the electromagnetic force is mediated by photons which are neutral, the neutral current cannot represent this interaction because unlike the weak force, the electromagnetic interaction treats the left- and right-handed fermions equally. One can add the fact that the neutrinos are blind to the electromagnetic force while they contribute in the weak neutral current interaction. Hence, moving towards the unification of the weak and the electromagnetic forces, a $U(1)$ component needs to be added to $SU(2)_L$. This component has to preserve the $SU(2)_L \times U(1)$ symmetry thus cannot be exactly the same as $U(1)_{em}$. The $U(1)_Y$ introduces a hypercharge of Y to fermions which relates to electric, Q, and isospin, C_{IS}, charge as follows

$$\frac{Y}{2} = Q - C_{IS}. \tag{1.4}$$

This gives a hypercharge of $Y = -1$ to the left-handed fermions while the right-handed fermions are given a hypercharge of $Y = -2$. The lepton sector of the electroweak model contains

$$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L, \quad e_R; \quad \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_L, \quad \mu_R; \quad \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_L, \quad \tau_R;$$

where the right-handed leptons are $SU(2)$ singlet. According to experimental results, almost all produced and observed neutrinos are left-handed, hence no ν_R is in the model.

1.1.2 The Higgs mechanism

The presented model includes the massless fermions and 3+1 massless gauge bosons. To represent the real world with massive particles, the $SU(2)_L \times U(1)_Y$ symmetry has to break spontaneously,

$$SU(2)_L \times U(1)_Y \rightarrow U(1)_{em} \ ,$$

where $U(1)_{em}$ stands for the electromagnetic symmetry which actually exists. This is achieved by introducing a new complex scalar field doublet of neutral, ϕ^0, and positively charged, ϕ^+ fields,

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$$

that interact with the fermions and the gauge bosons and this is why such particles receive masses after the symmetry breaking. The ϕ field evolves under the potential of

$$V(\phi^\dagger\phi) = -\mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2$$

for which μ^2 and λ are positive. Figure 1.1 illustrates this potential for two given fields, ϕ_1 and ϕ_2. The potential is minimized at many points symmetrically located around

[3] The spin-half particles which obey the Fermi-Dirac statistics, in contrast to bosons which obey the Bose-Einstein statistics.

the $V(\phi)$ axis in the (ϕ_1, ϕ_2) plane. It is often said that the potential has a degenerate vacuum state. In quantum field theory, the vacuum state is the quantum state with the lowest possible energy and particles are considered as the excitations of the vacuum. In the notation of the electroweak model, the symmetry or degeneracy is spontaneously broken if the scalar ϕ doublet develops a particular vacuum expectation value of e.g.

$$\phi_0 = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix}$$

where

$$\phi^\dagger\phi = |\phi|^2 = v^2/2, \quad \text{with} \quad v = \sqrt{\mu^2/\lambda}.$$

It can be shown that

$$C_{IS}\,\phi_0 = -\tfrac{1}{2}\phi_0, \quad \text{and} \quad Y\phi_0 = \phi_0,$$

but $Q\phi_0 = 0$, where C_{IS}, Y and Q are the charge operators given in Equation 1.4. Since the vacuum state is not physical, it has to be annihilated by physical operators. It means that the ϕ_0 vacuum state preserves the electromagnetic symmetry, $U(1)_{em}$, while it destroys the $SU(2)_L$ and $U(1)_Y$ symmetries. To see how particles become massive under the SSB process, the scalar field ϕ is written in terms of the fields denoting the shift from the vacuum state ϕ_0,

$$\phi_0 = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = e^{i\vec{\tau}.\vec{\xi}/2v} \begin{pmatrix} 0 \\ (v+H)/\sqrt{2} \end{pmatrix} \tag{1.5}$$

where $\vec{\tau}$ denotes the three Pauli matrices and ξ_i are real fields known as massless Goldstone bosons [4]. The H field stands for the so-called Higgs boson. Using this

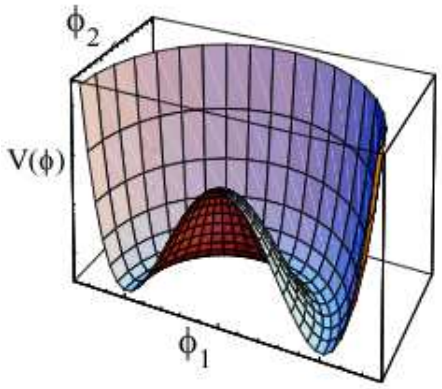


Figure 1.1: The Higgs potential with a degenerate vacuum state for two given fields, ϕ_1 and ϕ_2.

parametrization, three massless Goldstone bosons disappear under the unitary transformation of $e^{-i\vec{\tau}.\vec{\xi}/2v}$. There, with a lengthy calculation that can be found in the text

[4] The SSB mechanism with the appearance of massless fields was originally developed for O(2) symmetry, known as Goldstone Theorem (Goldstone, 1961; Goldstone, Salam and Weinberg, 1962; Buldman and Klein, 1962).

books of particle physics, such as [2], three massive gauge bosons, Z^0 and $W^{\pm}$, together with a massless photon, γ, appear in the symmetry-broken Lagrangian. The disappearance of the Goldstone bosons, ξ_i, via the $e^{-i\vec{\tau}.\vec{\xi}/2v}$ transformation may be thought of as missing degrees of freedom. This is however not true since the physical degrees of freedom are conserved by the SSB mechanism. The three Goldstone bosons disappear as they are "eaten" by the massive gauge bosons to serve as a new degree of freedom, the longitudinal polarization. The mass of the W and Z bosons can be predicted by the model using the experimental data. The relation

$$m_W = m_Z \cos\theta_W$$

is held between the masses of the weak bosons where $\theta_W \approx 0.231$ [3] is the weak mixing angle used to decouple the electromagnetic current from the weak neutral current. The W boson mass relates to the Fermi constant, $G_F = 1.16637 \times 10^{-5}\,\mathrm{GeV}^{-2}$ [3], in the Fermi theory of β decay via

$$m_W = \tfrac{1}{2}\left(\frac{e^2}{\sqrt{2}G_F}\right)^{1/2} \frac{1}{\sin\theta_W}.$$

Hence masses of $m_W \approx 80\,\mathrm{GeV}$ and $m_z \approx 90\,\mathrm{GeV}$ are expected for the W and Z boson, respectively.
The fermions with the exception of the neutrinos become massive due to their interaction with the Higgs particle. It can be shown that the strength of Higgs-fermion interactions is proportional to $|\frac{e}{2\sin\theta_W}\frac{m_f}{m_W}|^2$ where m_f is the fermion mass. Thus the Higgs coupling to ordinary leptons (e, μ and τ) or quarks (u, d, s, c and b) is extremely small while for heavy fermions like the top quark or possible exotic fermions the interaction cannot be neglected [4, 5].

1.1.3 Extention to quark sector

The observation of left-handed charged weak currents of hadrons such as

$$\pi^- \rightarrow \mu^- + \nu_\mu$$

suggests that like the lepton case, the left-handed components of the quark fields can be constructed into a doublet. Given the currently known three quark generations,

$$\begin{pmatrix} u \\ d \end{pmatrix}_L, \quad u_R,\, d_R; \quad \begin{pmatrix} c \\ s \end{pmatrix}_L, \quad c_R,\, s_R; \quad \begin{pmatrix} t \\ b \end{pmatrix}_L, \quad t_R,\, b_R;$$

one can immediately realize the differences with the lepton sector: the existence of the right-handed quarks with both isospin charges, $\pm\frac{1}{2}$. One has to add the interaction of all quarks with the electromagnetic field, too. Like the lepton sector, quarks interact with the Higgs boson field and receives mass after spontaneous symmetry breaking.

CKM matrix

The fields present in the electroweak Lagrangian are the eigenstates of the electroweak interaction. In reality however, the mass eigenstates which can be observed are related

to the electroweak eigenstates by a unitary transformation. Since each electroweak eigenstate is a linear combination of the mass eigenstates, the transformation can be illustrated as

$$\begin{pmatrix} u_w \\ c_w \\ t_w \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} u_m \\ c_m \\ t_m \end{pmatrix},$$

where the m and w indices represent the mass and electroweak eigenstate. A direct result of this transformation is the presence of interactions that are mixing different flavors. Imposing the unitarity requirement, the transformation matrix, known as the CKM (Cabbibo-Kobayashi-Maskawa) matrix, will have three real parameters and one phase that cannot be absorbed by the field redefinition [5].
The measured magnitudes of the CKM elements are [3]

$$\begin{pmatrix} 0.97452 \pm 0.00022 & 0.2252 \pm 0.0009 & (3.89 \pm 0.44) \times 10^{-3} \\ 0.230 \pm 0.011 & 1.023 \pm 0.036 & (40.6 \pm 1.3) \times 10^{-3} \\ (8.4 \pm 0.6) \times 10^{-3} & (38.7 \pm 2.1) \times 10^{-3} & 0.88 \pm 0.07 \end{pmatrix}.$$

The diagonal elements are close to one while the off-diagonal elements are smaller. In particular, the couplings of the top quark with other flavors than the b quark are rather small, leading to an almost exclusive $t \to Wb$ decay.
The phenomenon of flavor-mixing is not observed in the lepton sector since the neutrino masses are taken as $m_\nu = 0$. Hence, redefining the lepton fields, they can be the mass and the electroweak eigenstates at the same time.

1.2 Quantum Chromodynamics

In addition to the electromagnetic and weak interactions, quarks are subjected to the strong interaction, explained by Quantum Chromodynamics (QCD), due to their "color". There are several evidences such as the π^0 decay to photons, confirming that the "color" degree of freedom appears in fact in 3 species. The strong interaction is modeled by the non-Abelian SU(3) gauge group reflecting the existence of three colors. The model introduces gluons as the gauge bosons mediating the strong force. In addition to the interaction with quarks, gluons can interact among each other.

While the interaction between the gauge bosons and matter fields, is common between the electromagnetic, weak and the strong interactions, the gauge boson self interaction, is special for gluons. As will be discussed later, this property leads to the so-called "asymptotic freedom" in the strong interaction which is the reason that free quarks are not observed in the actual world. Quarks in nature are in the form of the color-singlet (colorless) hadrons including mesons ($q\bar{q}'$) and baryons ($qq'q''$).

[5] Proposed in [6], the extention of the quark model to 3 generations with the non-resolvable phase in the CKM matrix accommodates the weak processes containing CP violation, like the decay of the K_L hadron.

1.2.1 Running coupling constant

The strength of the electromagnetic interaction is governed by the electron charge, e, which depends on the Q^2 of the interaction due to the quantum effects.The quantity Q^2 is the momentum transfer of the electromagnetic interaction in which the electron participates. In the formalism of quantum electrodynamics, QED, this effect is described by the quantum corrections. Given an electron-electron scattering process, the first order correction comes from the electron loop, the left diagram in Figure 1.2. This leads to an effective strength of the QED interaction ,

$$\alpha_{eff}(Q^2) = \frac{\alpha_0}{1 + \frac{\alpha_0}{3\pi}\log\left(\frac{\Lambda^2}{Q^2}\right)}, \tag{1.6}$$

where $\alpha(Q^2)$ and α_0 represent the effective (renormalized) and the bare value of $\frac{e^2}{4\pi}$. The quantity Λ is a cut-off scale to avoid the ultra violet divergences in the scattering amplitude. Assuming that for some value of $Q^2 = \mu_R$ the experiment results in $\alpha(\mu_R^2) = \alpha_0$, the quantity $\alpha(Q^2)$ at any momentum transfer can be obtained accordingly,

$$\alpha_{eff}(Q^2) = \frac{\alpha_{\mu_R^2}}{1 - \frac{\alpha_{\mu_R^2}}{3\pi}\log\left(\frac{Q^2}{\mu_R^2}\right)}$$

where it can be seen that the non-physical Λ cut-off has been removed.
The effective $\alpha_{eff}(Q^2)$ is called the "running coupling constant", increasing at higher momentum transfers. The μ_R parameter is called the "renormalization scale" since it is renormalizing the electron charge in order to avoid ultraviolet divergences. It should be noted that the final physics results do not depend on the choice of the renormalization scale.

Similar to QED, the coupling constant of QCD, α_s, evolves as the momentum trans-

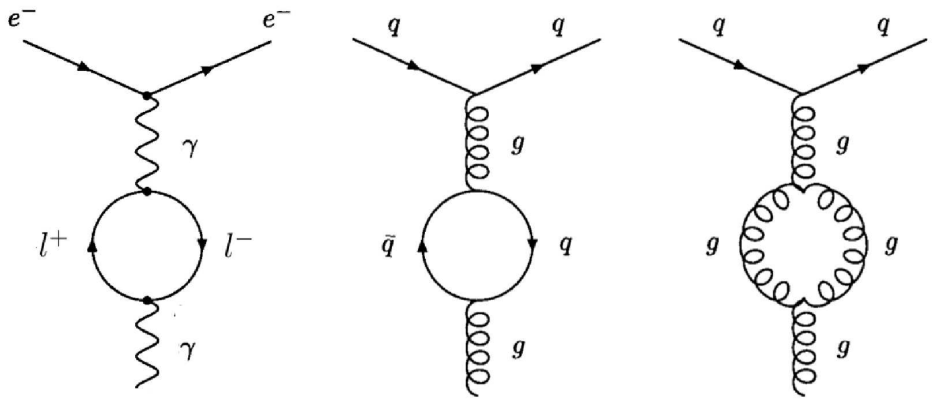


Figure 1.2: The Feynman diagrams of the first order loop corrections to the photon (left) and gluon (middle and right) propagators. The gluon self-interaction introduces extra terms in the quantum corrections which eventually leads to a different behavior between the QED and QCD coupling constants as a function of the momentum transfer, Q^2.

fer changes. The middle and the right diagrams in Figure 1.2 depict the one-loop corrections to the quark scattering which have an extra contribution from the gluon self-interaction. This leads to

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu_R^2)}{1 + \frac{(33-2\,n_f)\alpha_s(\mu_R^2)}{12\pi} \log\left(\frac{Q^2}{\mu_R^2}\right)} \tag{1.7}$$

for the running α_s where n_f, currently equal to 6, is the number of quark flavors contributing in the fermion loop of Figure 1.2. Since unlike QED, the running α_s decrease as Q^2 is enhanced, quarks are free particles at very high energies (asymptotic freedom). This means the perturbative QCD is valid at high energies and can be used to calculate the physical amplitudes. At low momentum transfers on the other hand, the running α_s becomes to large for the perturbative QCD to be valid. Phenomenological models are used instead in this energy regime where the quarks are confined in hadrons (confinement). A free parameter, Λ_{QCD}, is introduced in the evolution of $\alpha_s(Q^2)$ for which the denominator in Equation 1.7 becomes zero. This parameter is extracted from the experiment and depends on the momentum transfer.

1.2.2 The shortcomings of the Standard Model

The Standard Model of particle physics has achieved a great success in explaining, with a remarkable precision, almost all known aspects of elementary particles behavior. The model is actually well-established but, in addition to the Higgs particle which has not been discovered yet, it poses a number of well-defined questions to be addressed by forthcoming experiments [7]. Some of these questions are presented here:

- While the model partially unifies the weak and electromagnetic forces, the gravitation as a fundamental interaction is not considered in such a unification.
- The model describes the visible matter very well. However in cosmological physics, there are indications of a new type of weakly-interacting matter, dark matter, which constitutes up to $\sim 25\%$ of the energy density of the Universe [8].
- Another evidence [9] from cosmological physics is the accelerating expansion of the Universe, attributed to the dark energy. The dark energy accounts for $\gtrsim 70\%$ of the total energy density of the Universe [3] but it is not explained by the Standard Model of particle physics.
- Within the Standard Model framework, the quantum corrections to the Higgs mass are divergently large while for the theory to be perturbative, the Higgs mass has to be $\lesssim 1\,\text{TeV}$. To resolve this problem, known as hierarchy problem, an alternative which is disfavored by physicists is to fine tune the relative magnitudes of the tree level and loop contributions in order to get a small net correction. The other way is to introduce a new symmetry such as "supersymmetry" [10] to protect the Higgs mass.
- The observation of a non-zero neutrino mass [11] implies that the Standard Model must be extended to incorporate this fact.

Different models have been developed to explain (resolve) such observations (problems) consistently. These models can be tested in interactions with a very large momentum transfer accessible at high energy colliders like the Large Hadron Collider, the LHC [12], which is described in Chapter 2.

1.3 Top quark physics

The heaviest quark of the Standard Model, the top quark, has eluded the experiments until it was discovered in 1995 by the CDF [13] and DØ [14] collaborations [15, 16] at the Fermilab Tevatron [17]. The observation was at $\sqrt{s} = 1.8\,\text{TeV}$ center-of-mass energy for the top quark pair production via the strong interactions [6] while this particle is also produced as single top quark by electroweak processes. The single top quark process was observed quiet recently in 2009 by the CDF and DØ experiments [18, 19]. A great precision is achieved in the most recent result of the top quark mass measurement, $m_t = (173.2 \pm 0.9)\,\text{GeV}$ [20], which is a combination of the CDF and DØ analyses performed by the Tevatron Electroweak Working Group [20]. In 2010, the top quark has been rediscovered at the LHC. Soon after the first $p-p$ collisions, the ATLAS [21] and CMS [22] experiments reported their first observations of top quark pair production at $\sqrt{s} = 7\,\text{TeV}$ center-of-mass energy [23, 24] where more robust results were obtained using the whole LHC data collected in 2010 [25, 26]. The electroweak production of the top quark is also measured at the LHC [27].

1.3.1 The role of the top quark in the Standard Model and beyond

Most of the interesting properties of top quark are connected to its large mass. It induces a very short life time for the particle, $\tau \approx 3.3\times10^{-25}\,\text{s}$ [28], which is shorter than the hadronization time scale. Hence, the top quark is the only quark for which valuable information such as the spin polarization will not be destroyed by hadronization [29]. Since the top quark mass is close to the electroweak SSB scale and it has a large coupling with the Higgs boson, it is believed to play an important role in the electroweak symmetry breaking process [30]. The mass of the top quark is also important in the Higgs boson searches via the electroweak precision measurements. It contributes in the accuracy of the W and Z boson mass measurements through quantum loop corrections where the leading m_t dependence is quadratic, e.g. [7]

$$s_W^2 \equiv 1 - \frac{m_W^2}{m_Z^2} \;\ni\; -\frac{2\alpha}{16\pi\sin^2(\theta_W)}\frac{m_t^2}{m_Z^2}.$$

It can be seen that the contribution of such terms from other quarks is negligible. These relations were used to constrain the top quark mass by the electroweak measurements before its discovery [31]. The Higgs boson has also contributions in the quantum corrections of the electroweak parameters. Figure 1.3 illustrates the contour of 68% confidence level (CL) in m_t and m_H for the fit to all high momentum transfer data [32].

[6] See Section 3.2.1 for more details.

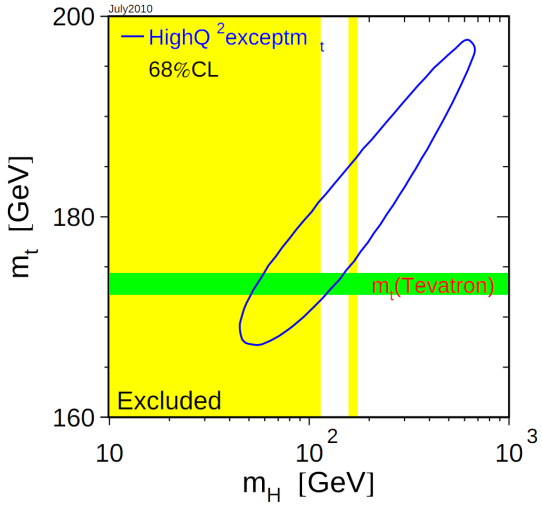


Figure 1.3: The 68% CL contour in m_t and m_H for the fit to all high-Q^2 data except the direct top quark mass measurement. The horizontal band illustrates the top quark mass measurement with the $\pm 1\sigma$ width. The vertical band is the 95% CL exclusion limit on the Higgs boson mass, up to 114 GeV, obtained from the direct searches at LEP-II. Calculated by the fit, the 95% CL upper limit on m_H is 158 GeV. It has been increased to 185 GeV considering the LEP-II direct search. The Higgs boson mass from 158 GeV to 175 GeV is excluded by Tevatron. The plot and the information are taken from [32].

Since the top quark mass is a free parameter, the data of the direct m_t measurement is not included in the fit. It is however indicated on the plot by the shaded horizontal $\pm 1\sigma$ width. The 68% CL contour shows stronger constraints on m_t compared to the Higgs boson mass because of the quadratic and logarithmic dependences of the electroweak parameters on the top quark and the Higgs boson mass, respectively.
For many theories beyond the Standard Model, top quark pairs are considered as background which need to be understood and suppressed in order to discover the phenomena of the new theory. On the other hand, the top quark is a key object in various new physics models. Numerous extensions to the Standard Model predict gauge interactions with an increased coupling to the third-generation quarks, specially top quarks. New particles are expected from these theories which can decay to top quark pairs, appearing as resonances in the top quark pair production [33]. For the wide resonances that may be hidden in the $t\bar{t}$ mass spectrum, the measurement of the charge asymmetry in top quark pair production can be studied instead [34]. In addition, the models introducing the fourth quark generation that can decay to top quarks have recently become interesting again [35]. One has to consider supersymmetry models with more Higgs boson doublets in which depending on the mass of these new Higgs bosons, they may decay to top quarks or may be produced via the top quark decay.

1.3.2 Commissioning and calibration using top quark

As presented in Section 1.1.3, the decay of the top quark to down-type quarks of other generations is highly suppressed. Therefore the top quark decays exclusively to a b-quark and a W boson which is real due to the large top quark mass. The final

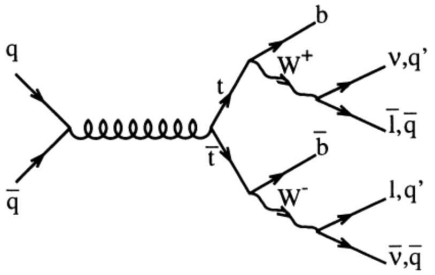


Figure 1.4: The Feynman diagram of the top quark pair production via $q\bar{q}$ annihilation together with the decay of top quarks. Different combinations of the W boson decays from each top quark result in different topologies in the $t\bar{t}$ final state.

state of the top quark pair production at hadron colliders features different topologies, categorized by the W boson decay modes which are either leptonic or hadronic. Figure 1.4 illustrates the top quark pair production via quark annihilation (explained in Section 3.2.1) together with its decay to all possible channels. A graphical view of different topologies is shown in Figure 1.5 where the decay rate is reflected by the size of the boxes. The diversity of particles found in the $t\bar{t}$ decay means that all physics

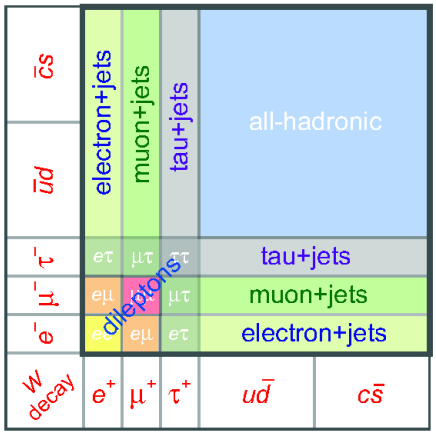


Figure 1.5: Different categories for the $t\bar{t}$ final states [36]. While the all-hadronic channel include the hadronic decays of the W boson from the top and anti-top quark, the lepton+jets categories indicate the final states in which one of the W bosons decay leptonically. In the dilepton channels, both of the W bosons decay leptonically. Regarding the possible hadronic decays of the W boson to ud or cs quark doublets, the color degree of freedom for each quark, and three lepton flavors for the leptonic decay, the decay rates are approximately $\frac{4}{9}$ for the full-hadronic, $\frac{4}{27}$ for lepton+jets per lepton flavor and $\frac{1}{9}$ for the inclusive dilepton final states. The rates are indicated by the size of the boxes.

objects relevant for the detector commissioning including leptons and jets of hadrons are available in the top quark pair final state. Considering the high rate of the $t\bar{t}$ production at the LHC, top quark pairs are well suited for commissioning purposes. Moreover since the top quark properties are mostly measured accurately at the Tevatron, the jets of hadrons in the $t\bar{t}$ final state can be exploited to calibrate the jets energy measured by the detector [37]. The presence of two b-quarks which are hadronized to b-flavored jets in the top quark pair final states provides a rich sample of b-flavored jets by which the efficiency of the b-flavor identification algorithms can be estimated with a data driven approach.

Chapter 2

The CMS Experiment at the LHC

The Standard Model of particle physics has successfully undergone many precision tests at high energy colliders. The model has also succeeded to predict the existence of for example top-quark which was discovered at the Tevatron [17] in Fermilab. However, as discussed in the previous chapter there are still missing elements in the model for which colliders at even higher energies are needed.

The Large Hadron Collider, LHC [12], is built up to make these new territories accessible in high energy physics. This proton-proton collider and accelerator, with the planned 14 TeV center-of-mass energy and with the highest rate of collisions, is the latest and the most powerful in a series of particle accelerators that allows scientists to probe the structure of matter at its tiniest dimension. More about the design and the physics motivations behind the LHC, the performance of the machine and its current experiments is addressed in Section 2.1.

Beside the powerful collider, excellent detectors with the highest possible technology are required to collect the collision data with a good quality. The Compact Muon Solenoid, CMS [22], is one of the experiments at the LHC described in Section 2.2. The excellence of data also relies on the computing infrastructure and facilities. Section 2.3 is dedicated to the CMS data taking and monitoring in addition to its computing environment.

2.1 The Large Hadron Collider

The Large Hadron Collider is installed in the existing 26.7 km tunnel constructed between 1984 and 1989 for the Large Electron Positron (LEP) collider. LEP was the flagship accelerator of CERN (European Organization for Nuclear Research) between the years 1989 and 2000 when it was shut down and dismantled to be replaced by the LHC. The approval of the LHC project was given by the CERN Council in December 1994. The plan was to accelerate the proton beams to 7 TeV so the center-of-mass energy reaches to 14 TeV in the collisions. It is also designed to accelerate and collide the lead ions, the so-called heavy ion collisions.

The construction of the machine has been a monumental effort spanning almost 15

years and involving scientists and engineers from all over the world. The accelerator has been in operation since fall 2009. After the commissioning phases with lower energies (0.9 and 2.36 TeV), finally in March 2010 the machine started to work at 7 TeV center-of-mass energy. Performing effectively in 2010, LHC delivered about $47\,pb^{-1}$ data to its experiments. In November 2010, LHC experienced its first heavy ion collisions. This data was also analyzed by different experiments at the LHC. The 2011 run has already started in March at the same center-of-mass energy. The 2011 goal is to collect more than $1\,fb^{-1}$ of collision data.

The physics motivations behind the construction of the LHC is discussed in Section 2.1.1. In Section 2.1.2 the design and the layout of the machine are described while the different experiments and their physics program are introduced in Section 2.1.3.

2.1.1 Physics motivation

The aim of the LHC is to reveal the physics beyond the Standard Model and to search for the Higgs particle which is believed to be responsible for Electroweak Symmetry Breaking (EWSB). The experimental study of the Higgs mechanism can shed light on the mathematical consistency of the SM at energy scales above about 1 TeV. Moreover, the BSM discoveries could take the form of supersymmetry or extra dimensions, the latter requiring modification of gravity at the TeV scale.

The Higgs boson

The Higgs boson production mechanism at the LHC is the same as at the precedent hadron collider, Tevatron. However, due to the higher energies accessible at the LHC, the production rate is enhanced by up to two orders of magnitued. While at Tevatron, the Higgs boson total cross section for $M_H \sim 160\,\text{GeV}$ is $\sim$0.495 pb [38], at the LHC it reaches to 36.6($\sim$10.4) pb for 14(7) TeV center-of-mass eneregy [39]. Taking into account the higher collision rate at the LHC, the number of Higgs processes becomes considerable. Gluon fusion through a heavy-quark loop [40] (see Figure 2.1 a) is the

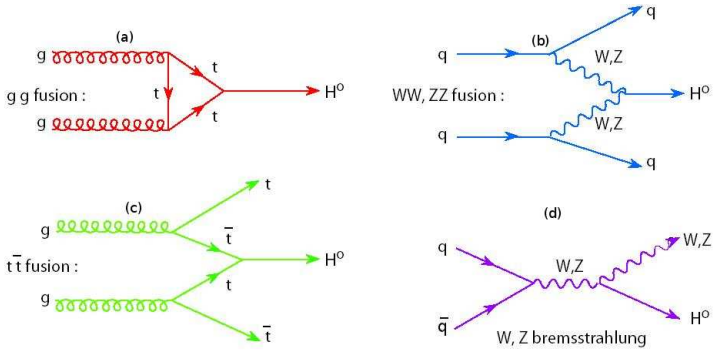


Figure 2.1: Feynman diagrams for the Higgs boson production at the LHC.

main production mechanism of the Standard Model Higgs boson at hadron colliders. Another mechanism is the production of a Standard Model Higgs boson in association with two quarks (Figure 2.1 b), frequently quoted as the vector-boson fusion (VBF)

channel. Higgs boson production in the VBF channel plays an important role in the determination of Higgs boson couplings at the LHC (see e.g. [41]). Bounds on non-standard couplings between the Higgs boson and electroweak (EW) gauge bosons can be imposed from precision studies in this channel [42]. This channel contributes in a significant way to the inclusive Higgs boson production specially in high mass ranges. The production of the Higgs boson can also be in association either with W/Z-bosons, Higgs-strahlung processes, or a $t\bar{t}$ pair (Figure 2.1 d and c). Each of these processes has their own physics interests. While the measurement of the $t\bar{t}H$ production rate can provide relevant information on the top-Higgs Yukawa coupling, the Higgs-strahlung processes are interesting in low-mass Higgs discovery and W-H coupling measurements. Figure 2.2 shows the Higgs boson cross section versus its mass for different mechanisms at Tevatron and the LHC. The total Higgs boson production rates at the LHC in 14 and 7 TeV are plotted versus the Higgs boson mass in Figure 2.3. The Standard

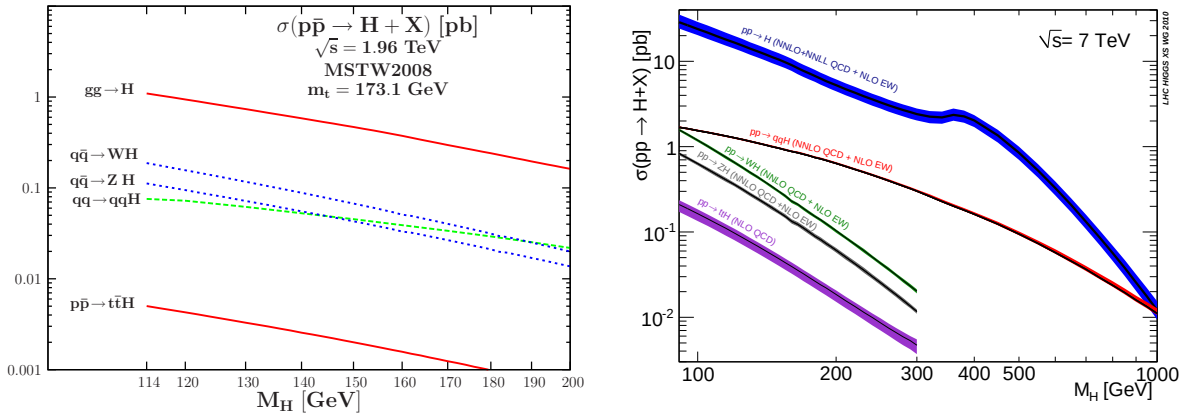


Figure 2.2: The conribution of different Higgs boson procuction modes in Tevatron (left) and at the LHC (right).

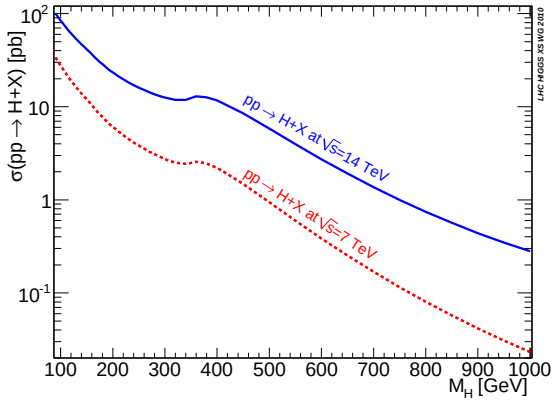


Figure 2.3: Inclusive SM Higgs boson production compared in 14 TeV and 7 TeV at the LHC.

Model Higgs particle decays in different channels with different rates depending on the

Higgs boson mass range, Figure 2.4. When the gluon fusion production mechanism is combined with the decay channels $H \rightarrow \gamma\gamma$, $H \rightarrow WW$, and $H \rightarrow ZZ$, it becomes one of the most important channels for Higgs boson searches and studies over the entire mass range, 100 GeV$< M_H <$1 TeV, to be investigated at the LHC. Although

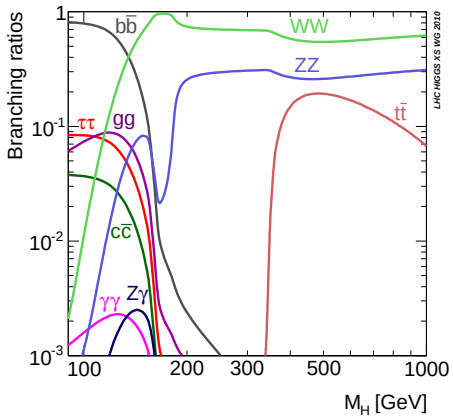


Figure 2.4: Standard Model Higgs boson decay branching ratio.

there is still no indication of the Higgs boson existence at the LHC or elsewhere (see for example [43]), in the near future with more accumulated data an answer to the question of spontaneous symmetry breaking is expected.

The Physics Beyond the Standard Model

With the high center-of-mass energy in collisions, LHC provides the opportunity for a wide range of theories beyond the Standard Model to be investigated [44]. Looking for a candidate for dark matter, supersymmetric models in addition to scenarios with universal extra dimensions and composite models are studied at the LHC. In SUSY, the search for the super-particles, finding new constraints on the SUSY parameters together with the global supersymmetric fit studies [45, 46] are also in the basket. In a search for excited quarks q that might have been manufactured in $q + g$ collisions and decay via $q \rightarrow q + g$, one of the first LHC results has already set limits on physics beyond the Standard Model that are stronger than those set by previous experiments [47, 48]. In some string scenarios, the scattering of quarks and gluons in the channels $q + q$, $q + g$ and $g + g$ may reveal resonances at indistinguishable masses. The same LHC results have also excluded this possibility up to a mass of 2.5 TeV [45], a limit that is also much stronger than previous constraints. In some theories with large extra dimensions, gravity may become strong at the TeV scale, in which case the high-energy collisions of quarks and gluons might produce microscopic black holes [49, 50]. The theories that predict such a possibility also predict that these microscopic black holes would decay very rapidly through Hawking radiation [51]. The production and decay of microscopic black holes at the LHC has now been excluded over a wide range of masses [52].

2.1.2 LHC design and performance

Each physics process with a cross section of σ_{process} is produced N_{event} times per second at the LHC

$$N_{\text{event}} = \mathcal{L}\sigma_{\text{event}}. \tag{2.1}$$

In Equation 2.1, $\mathcal{L}$ denotes for the machine luminosity, the quantity which is proportional to the number of collisions per second and depends on the beam properties. The beams at the LHC have a bunched structure of n_{b} bunch per beam, each bunch has $N_{\text{p}} = 10^{11}$ protons at the nominal conditions. In general

$$\mathcal{L} \propto f_{\text{rev}} \frac{N_{\text{p}}^2\, n_{\text{b}}}{\sigma_1\, \sigma_2}. \tag{2.2}$$

To increase the probability of a physics process, the luminosity needs to be enhanced by the increament of either the number of colliding particles (N_{p} and/or n_{b}) or the rate of the collision (f_{rev}: revolution frequency). If the beams are well-focused, the proton density is higher and there is more chance for the hard interactions to happen (σ_{i}: beam cross section). The integral of the delivered luminosity over time is called integrated luminosity

$$L = \int \mathcal{L} dt. \tag{2.3}$$

Expressed in inverse of cross section units (i.e. 1/nb, 1/pb or 1/fb), Integrated luminosity is a measure of the collected data size, and it is an important quantity to characterize the performance of an accelerator. In the nominal design condition of the LHC, it is foreseen to collect $100\, fb^{-1}$ of data per year. Figure 2.5 is the integrated luminosity delivered by the LHC during the year 2010 at 7 TeV center-of-mass energy. Working at high energies, E, the De Broglie wavelength of the colliding particles goes down like $\frac{1}{E}$ and hence the particle "cross section" decreases like $\frac{1}{E^2}$. Therefore in order to maintain an equally effective physics programme, the luminosity of a collider should increase in proportion to E^2. Wherease in the past and present colliders the luminosity is around $\mathcal{L} = 10^{32}\, cm^{-2}s^{-1}$, the LHC has succeefully reached this value only at the its first year of operation in 2010. With 2825 bunches of protons per beam, the LHC is designed for $\mathcal{L} = 10^{34}\, cm^{-2}s^{-1}$ where even very rare physics intractions are probable to be created. To make this high luminosity accessible, it was decided for the LHC to be a particle-particle collider because the production of anti-protons was not efficienct enough. However, it was not possible to accelerate two particle beams in the same ring like what was the case for the LEP and Tevatron.

Having the LEP tunnel "as built", a serious limitiation was the small space for installing two completely separated rings. This finally resulted in the adoption of the "twin-bore" superconducting magnet design, Figure 2.6 , since the use of the LEP tunnel was still cost-saving. The necessity of superconductivity is driven by the high energy demand. The large magnetic fields require very large currents and are efficiently achivelable using superconducting magnets. The superconducting magnets at the LHC are at the edge of the present technology. Using superfluid Helium, magnets are cooled down to 2 K and they function at a field strength of 8.33 T.
The core of the dipole magnet is the "dipole cold mass" that reffering to Figure 2.6,

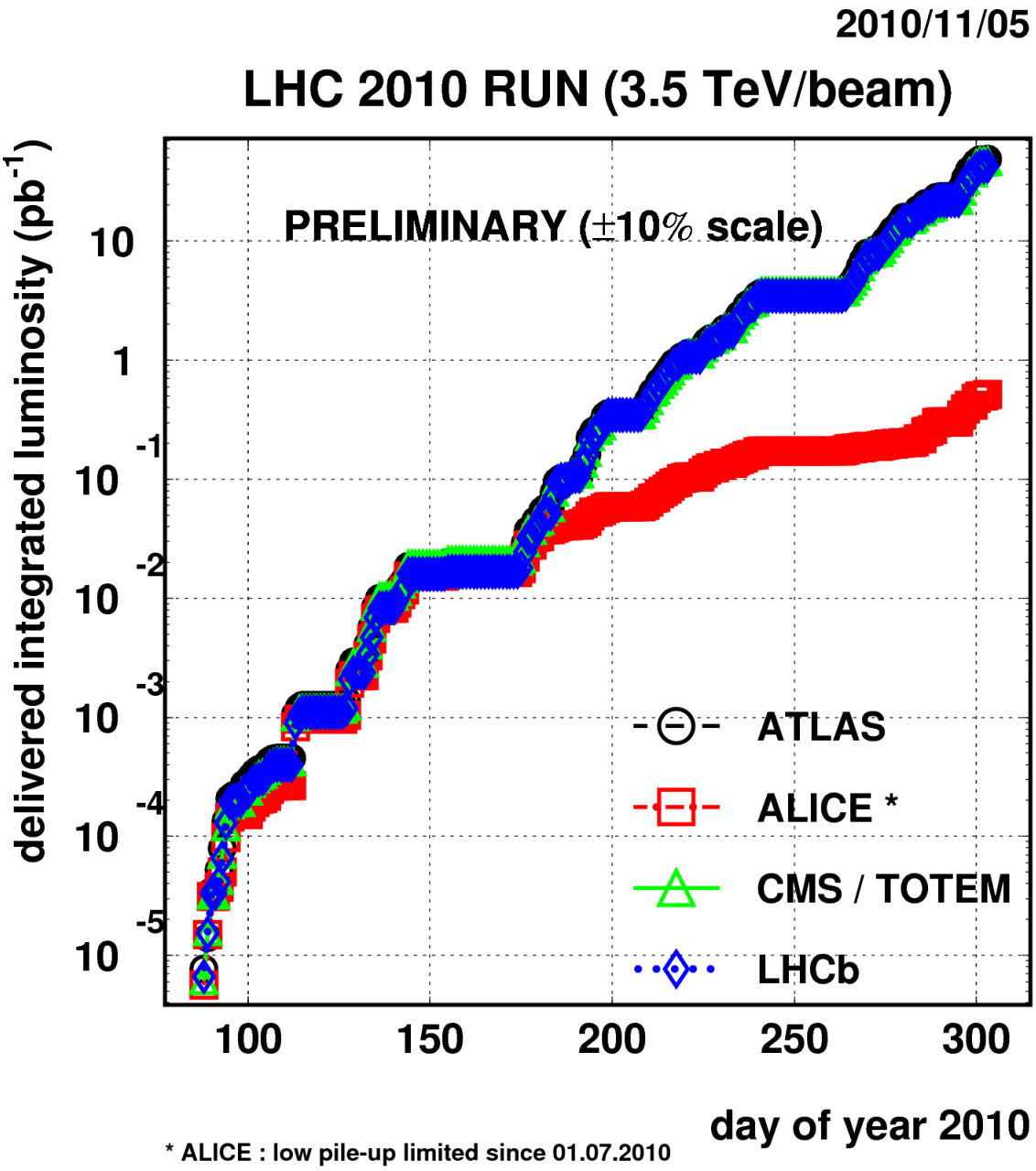


Figure 2.5: LHC delivered luminosity to different experiments in 2010.

is the part inside the shrinking cylinder/He II vessel. Hence it contains all the components cooled by the superfluid Helium. The cold mass provides two apertures for the pipes, the tubes in which proton beams move around. In the superconduction coils, currents of $\sim 12\,\mathrm{kA}$ circulate in such a way that the magnetic field has opposite directions in the two cold bore tubes. Each magnet dipole has an overall length about 16.5 m (connections and supports included), a diameter of 570 mm and a mass of about 27.5 t. The LHC main ring consists of 1232 dipole magnets.

Besides the superconducting dipoles, the LHC uses other types of magnets. To keep

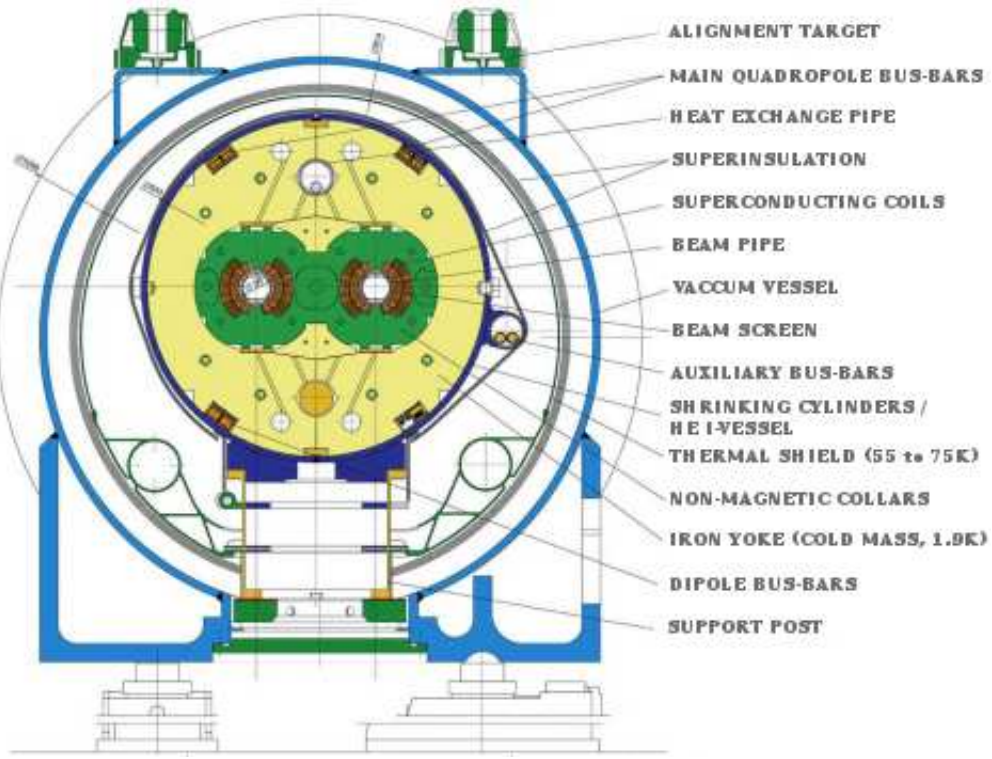


Figure 2.6: Schematic cross-section of cryodipole (lengths in mm).

the beam focused specially against the repulsive electromagnetic force among protons, quadrupole magnets are used at the LHC. They concentrate the beam both vertically and horizontally along its track. The focusing changes with energy of the incoming particle and due to the slight difference in the energy of protons, different tunings are needed (sextupole magnets). Other magnetic multipoles act to help in beam focussing and counteracting other interactions that each beam suffers (e.g. electromagnetic interactions among bunches, electron clouds from the pipe wall, etc). Depending on their functionality, these magnets are positioned in the LHC ring in a predefined sequence [53].

Regarding the stored beam energy of up to 350 MJ at 7 TeV, the beams at the LHC are highly destructive. The superconducting magnets in the LHC would quench at 7 TeV if small amounts of energy (on the level of $30\,mJ\,cm^3$, induced by a local transient loss of 4×10^7 protons) is deposited into the superconducting magnet coils [54]. Any significant beam loss into the cold aperture must therefore be avoided. However, beam losses cannot be completely suppressed. A so-called primary beam

halo will continuously be filled by various beam dynamics processes and the beam current lifetime will be finite [55]. The handling of the high intensity LHC beams and the associated high loss rates of protons requires a powerful collimation system [56].
Figure 2.7, depicts the layout of the machine. Following the LEP tunnel geometry, the LHC has eight arcs and eight straight sections. The straight sections (528 m each), serve as an experimental or utility insertion. Two high luminosity experimenta insertion, ATLAS and CMS, are located at Point 1 and Point 5 respectively. ALICE at Point 2 and LHCb at Point 8, are the two more experimental insertion stations. In these four locations, as shown in Figure 2.7, the beams cross from one magnet bore to the other i.e. the collisions take place. There, the β function indicated as "Low" at the interaction points is a quantity determind by specially the quadropole magnet conficuration of the accelerator. Low β implies that the beams are narrow enough to make a high luminosity collision.
Two of the remaining insertion points, Point 3 and Point 7, are equiped with the collimation system. Two independent RF systems are installed at Point 4, one for each LHC beam. To make a fast extraction of the circulating beams from each ring of the collider with minimal losses, the LHC beam dumping system is designed. When it is time to get rid of the beams, also in case of emergencies, the beams are extracted from the ring by a system of kicker magnets, they are diluted to reduce the peak energy density and then they are absorbed in a dedicated system. The straight section at Point 6 is devoted to the beam dump insertion where the beam has an independent abort system. The CERN accelerator complex is started by proton production [57]. Accelerated up to 100 keV, protons are sent to a Radio Frequency Quadrupole which both speeds up and focuses the particle beam. They leave the quarupole system with an energy of 750 keV and enter the linear accelerator, LINAC2 [58]. The linac tank is a multi-chamber resonant cavity tuned to a specific frequency which creates potential differences in the cavities that accelerate the particle up to 50 MeV. Protons pass the linac and reach the 157 m circumference circular accelerator Proton Synchrotron Booster (PSB) in a few microseconds. The PSB is a circular four ring accelerator with a magnetic field to bend the particles. The energy of particles increases in time so as the magnetic field, hence a constant orbit is maintained during acceleration. The PSB accelerates particles to 1.4 GeV in 530 ms, then after less than a microsecond they are injected in the 628 m circumference circular accelerator Proton Synchrotron (PS). In the PS protons reach the energy of 25 GeV. The PS is responsible for providing 81 bunch packets with 25 ns spacing for the LHC. Triplets of 81 bunches formed in the PS are injected into the 7 km circumference circular accelerator Super Proton Synchrotron (SPS), where they are accelerated to 450 GeV in 4.3 seconds, and sent to the LHC. The LHC finally enhances the beam energy to the nominal 7 TeV value in about 20 minutes, considered as the minimum needed time. Filling the LHC requires 12 cycles of SPS and each SPS fill requires 3 to 4 cycles of PS accelerator. Considering the cycling time of PS and SPS, the LHC filling time is approximately 4 minutes per beam. In 2010, with the same sequential procedure the proton beems were accelerated up to 3.5 TeV at the LHC. Thanks to the successful operation, the beam energy will increase toward the nominal value, 7 TeV. Depending on the beam lifetime, the beams circulate up to 10 hours. The dipole magnets are then

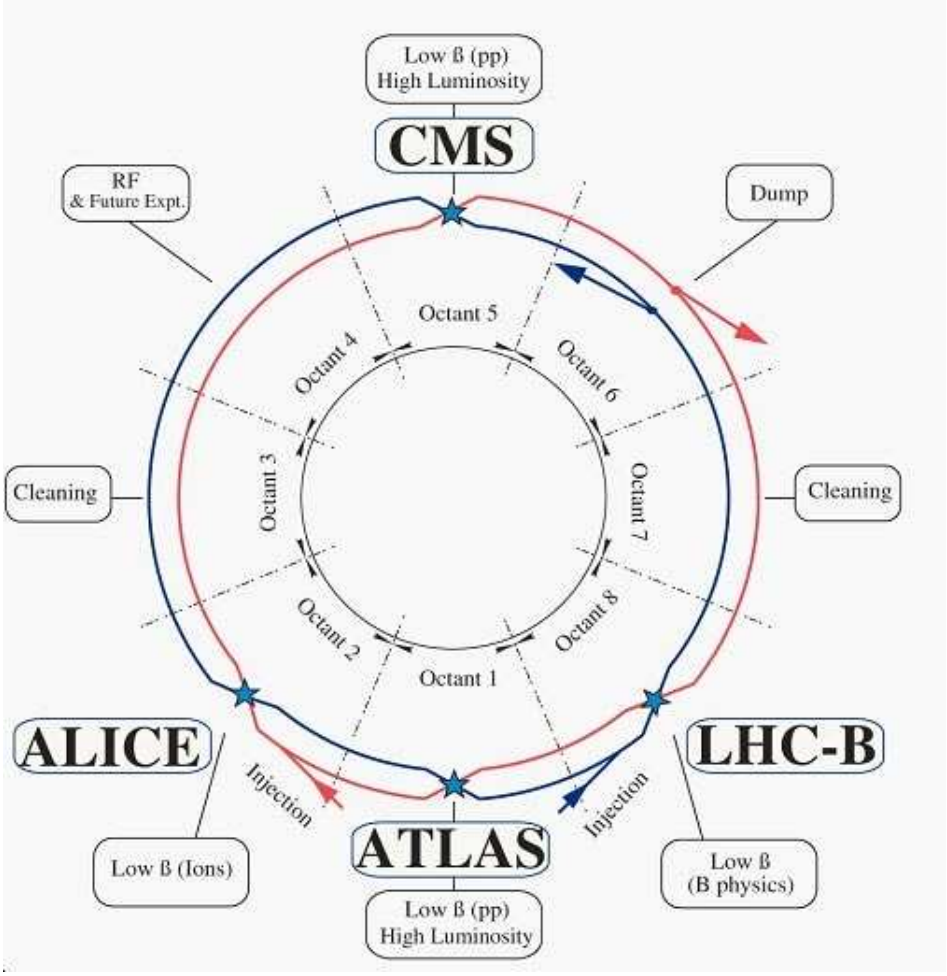


Figure 2.7: Schematic layout of the LHC.

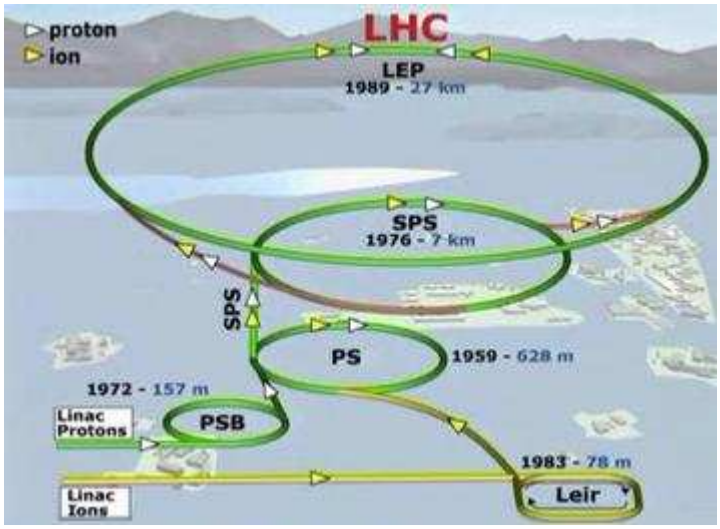


Figure 2.8: The CERN accelerator complex.

ramped down to 0.54 T (450 GeV for the dumped beam energy) and they remain so for some 20−40 min. Meanwhile beam injection is repeated before the magnets are ramped up again to 8.3 T for another cycle of high energy collisions. Including the time needed for the readjustment of the machine settings and the check of all the main systems, the turnaround time for the LHC is about 70 minutes which is the theoretical minimum. The heavy ion injector chain has the PS, SPS and the LHC ring in common with the proton beams while it owns a dedicated pre-acceleration system. Figure 2.8, shows the LHC accelerating complex for both protons and lead nuclei.

2.1.3 Current Experiments at the LHC

The LHC has 6 experiments in total, all are run by international collaborations and bring together scientists from institutes all over the world. Each experiment is distinct and characterized by its unique particle detector. The search for a Standard Model Higgs boson has been a benchmark in the design of the ATLAS (A Toroidal LHC ApparatuS) [21] and CMS [22] detectors, the two general-purposes experiments at the LHC, which should either discover or exclude it over all the mass range up to ∼1 TeV. Supersymmetry and/or extra dimensions are features of unified theories, and may also lie within the reach of the ATLAS and CMS experiments. Having two independently designed detectors is vital for cross-confirmation of any new discoveries made.

LHCb [59] is a dedicated medium-size experiment, studying CP violation and rare decays of heavy quarks, looking for new physics beyond the dominant Cabibbo-Kobayashi-Maskawa paradigm within the Standard Model. Another medium-size experiment, ALICE [60] is designed to exploit the unique physics potential of nucleus-nucleus interactions at LHC energies in the heavy-ion collisions. The aim is to study the physics of strongly interacting matter at extreme energy densities, where the formation of a new phase of matter, the quark-gluon plasma, is expected.

Two experiments, TOTEM [61] and LHCf [62], are much smaller in size. They are designed to focus on forward particles (protons or heavy ions). These are particles that just brush past each other as the beams collide, rather than meeting head-on. The TOTEM (Total Cross Section, Elastic Scattering and Diffraction Dissociation) experiment aims to obtain a measurement of the total and elastic p-p cross sections, with an uncertainty of about 1%, over a large range of 4-momentum transfers. Modest in size, TOTEM has been installed near the point where protons collide in the center of the CMS detector. It uses silicon sensors installed in the LHC tunnel approximately 200 m away from CMS. The LHCf experiment is intended to measure the energy and numbers of neutral pions (π^0) produced by the collider. This study will give important information for understanding the development of atmospheric showers induced by ultra-high-energy cosmic rays hitting the Earth atmosphere. The results will complement other high-energy cosmic ray measurements from the Pierre Auger Observatory in Argentina [63], and the Telescope Array in Utah [64].

The ATLAS, CMS, ALICE and LHCb detectors are installed in four huge underground caverns located around the ring of the LHC. While detectors used by the TOTEM experiment are positioned near the CMS detector, those used by LHCf are near the ATLAS detector.

LHC experiments in 2010

The experiments at the LHC, have been in operation since the beginning of the LHC beam collisions (see for example [65] and [66]). The two general-purposes experiments, ATLAS and CMS, have been working with their rich physics program in common areas. They both have performed outstanding scientific research and have provided excellent results. The full list of the journal articles for ATLAS and CMS is available on the CERN Document Server [67]. In the followings, some of the physics output of CMS in 2010 is highlighted.

Brief review of 2010 physics in CMS

The good data-taking performance of the CMS detector in 2010 is shown in Figure 2.9 where it managed to collect $43\,pb^{-1}$ out of $\sim 47\,pb^{-1}$ data delivered by the LHC and handled an increase of more than 5 orders of magnitude in instantaneous luminosity over 7 months. The CMS collaboration has started to work with the collision data since

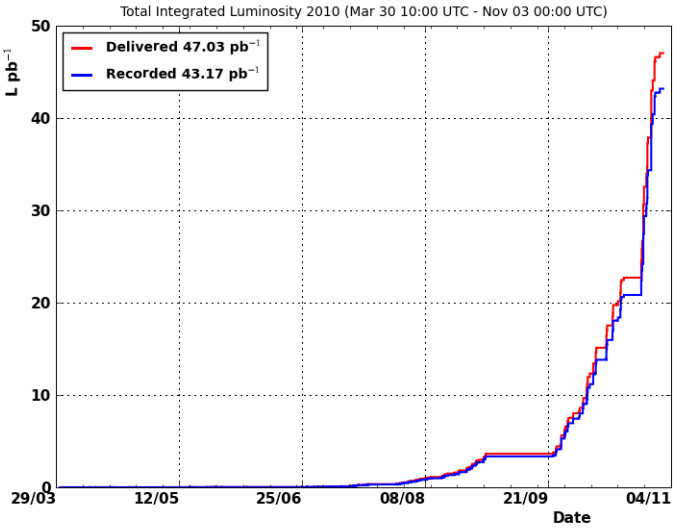


Figure 2.9: Integrated luminosity versus time delivered to (red), and recorded by CMS (blue) during stable beams in 2010 at 7 TeV centre-of-mass energy.

the very first runs at a center-of-mass energy of 900 GeV. Besides the big effort to check the functionality of all sub-detectors, one of the first attempts was the measurement of the underlying event [ref to section] activities at this energy scale [68]. Increasing the center-of-mass energy to 2.36 TeV at the LHC, CMS had the occasion to put the data of different energies in collision together and perform divers measurements including the Bose-Einstein correlation [69] where the signal was observed in the form of an enhancement of pairs of same-sign charged particles with small relative four-momentum. Part of these analyses were repeated with 7 TeV data however, this data was mainly utilized to ascertain the Standard Model and to search for the new physics. A vast area of Standard Model physics, from the production of strange particles [70] and b-hadrons [71] to the cross section measurements of the Electroweak gauge bosons [72] has been covered by CMS. In this effort, many analyses were dedicated to rediscovery of top-quark, produced via both the strong and Electroweak interaction [ref to section].

Following its original plan, CMS has conducted several studies in the quest for the Higgs boson of the Standard Model. In $H \rightarrow WW$ channel, limits were set on the production of the Higgs boson in the context of the Standard Model and in the presence of a sequential fourth family of fermions with high masses. In the latter context, a Higgs boson with mass between 144 and 207 GeV is ruled out at 95% confidence level [73]. The outcome of the LHC in 2010 had also surprises amongst one can point to the observation of Long-Range Near-Side Angular Correlations, which was seen in the high multiplicity events [74].
Looking for evidences of new physics, various experimental signature were studied [75–77]. As a hint for extra dimensions, the existence of W'-boson in very high mass ranges was investigated. Combining both electron and muon final states, masses below 1.58 TeV were excluded at 95% confidence level for a sequential SM-like W'-boson [78]. Another analysis set lower limits of 3.5-4.5 TeV on the mass of microscopic black-holes, the phenomena which could be expected as a consequence of the extra dimensions [52]. With only $35\,pb^{-1}$ of data searches for squarks and gluinos expanded the excluded range established during the last 20 years by a factor of 2 [45]. In neutral MSSM Higgs boson studies, new bounds have been obtained in the MSSM parameter space [79].
The physics effort of the CMS collaboration was not limited to the pp collisions but continued with the heavy-ion runs where the jet quenching phenomenon was studied and observed [80].

2.2 The Compact Muon Solenoid experiment

As a general-purpose experiment, the design of the CMS detector has been optimized to meet the requirements of its physics goals [81]. The detector is characterized by its strong superconducting magnet and its powerful muon system. However, it is also able to well identify other physics objects and to measure their properties with a good accuracy. Depending on their functionality, sub-detectors are made up of different materials and are arranged in different dimensions and configurations. To make precise measurements of position and momentum and to reach an accurate object reconstruction, the alignment of the whole apparatus is crucial. It means that the exact position and orientation of modules in sub-detectors together with the relative orientation of sub-detectors with respect to each other, need to be known precisely in the CMS coordinate system. The coordinate system adopted by CMS has the origin centered at the nominal collision point inside the experiment, the y-axis pointing vertically upward, and the x-axis pointing radially inward toward the center of the LHC. Thus, the z-axis points along the beam direction toward the Jura mountains from LHC Point 5. The azimuthal angle ϕ is measured from the x-axis in the x-y plane. The polar angle θ is measured from the z-axis. Pseudorapidity is defined as $\eta = -ln(tan(\theta/2))$.
In addition, calibration is necessary for those parts of the detector that are meant to measure the energy of particles. Both alignment and calibration are the subjects of a very big effort in CMS either before data-taking with the test beam observations [81] and cosmic muons [82] or during the operation with the collision data [83].
An overall view of the detector, its layout and its subsystems is described here while more details about subdetectors and their performance are given in the following sub-

sections.

Figure 2.10 is a schematic view of the CMS detector. The overall dimensions of

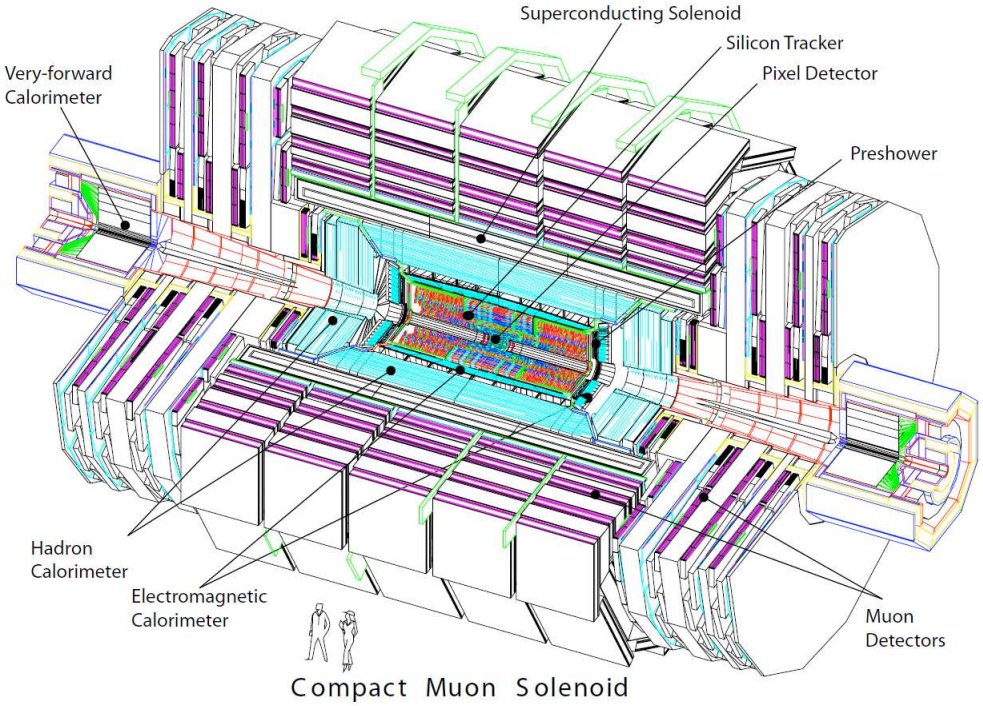


Figure 2.10: An schematic view of the CMS detector.

the CMS detector are a length of 21.6 m, a diameter of 14.6 m and a total weight of 12500 tons. It can in principal be decomposed into the magnet which holds the core of the detector and the muon stations covering the magnet. The superconducting solenoid with a designed strength of 4 T has a length of 13 m and an inner diameter of 5.9 m. The magnetic field configuration, has been an important aspect driving the detector design. The strong magnet with its large bending power is essential for the precise measurement of the momentum of muons within a compact spectrometer. In fact in the presence of such a powerful magnet, the high momentum resolution is achieved without making stringent demands on the muon-chamber resolution. The magnetic flux is returned through a 1.5 m thick iron yoke and makes it saturated. Therefore, the four muon stations which are surrounding the magnet can be integrated and provide the full geometry coverage. The muons are identified by the muon system where their momentum is measured with a good resolution over a wide range of momenta while their charge is also unambiguously determined.

As it can be seen in Figure 2.10, the bore of the large magnet accommodates the calorimetry apparatus and the tracking system. A good charged particle momentum resolution and reconstruction efficiency is obtained in the inner tracker. Efficient on-line selection and off-line tagging of τ's and b-jets are achieved by the pixel detectors close to the interaction region. The tracking volume is defined by a cylinder of length 5.8 m and diameter 2.6 m. For electromagnetic particles like electrons and photons, the

electromagnetic calorimeter, ECAL, stands with a good energy resolution and wide geometric coverage, $|\eta|$ <3.0. The preshower system helps rejecting the undesirable π^0 particles in the ECAL endcap region. The ECAL is surrounded by a brass/scintillator sampling hadron calorimeter, HCAL, with coverage up to $|\eta|$ <3.0. With its fine lateral segmentation, the hadron calorimeter, HCAL, is responsible for the measurement of the hadron activities in jets and is a key element for the reconstruction of the missing transverse energy, $E_{\mathrm{T}}^{\mathrm{miss}}$.
Coverage up to a pseudorapidity of $\eta = 5.0$ is provided by an iron/quartz-fiber calorimeter, HF. The Cerenkov light emitted in the quartz fibers is detected by photomultiplier. The forward calorimeters ensure full geometric coverage for the measurement of the transverse energy in the event.

2.2.1 Inner tracking system

The reconstructed tracks of charged particles are among the most fundamental objects in the reconstruction of *pp* collisions. Tracks can be used to reconstruct the decays of hadrons, photon conversions, and nuclear interactions. In addition, tracks are components in the reconstruction of other objects such as electrons and muons. Reconstructed tracks in the inner tracker are also used to determine the position of the primary interaction vertex in the event and to monitor the position of the beamspot. The beamspot represents the profile of the luminous region where the LHC beams collide at CMS. The beamspot is determined as an average over many events, in contrast to the event-by-event primary vertex which gives the precise position of a single collision.

Immediately around the interaction point the inner tracker (Figure 2.11) serves to

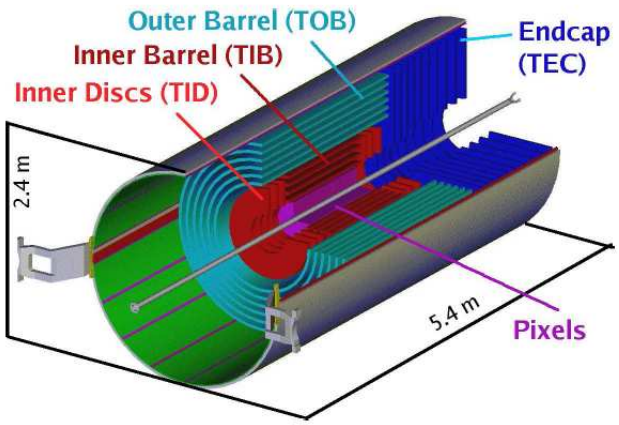


Figure 2.11: Schematic 3D view of the CMS inner tracking system.

identify the tracks of individual particles and match them to the vertices from which they originated. The curvature of charged particle tracks in the high magnetic field, 3.8 T, of the detector allows their charge and momentum to be measured. In addition to the precise momentum measurement, the CMS tracker is designed to have a very high impact parameter resolution. Hence the precise b-jet identification algorithms together with the tools to tag different physics objects like the τ-leptons can be developed. It is also expected to have a high track reconstruction efficiency.

The high charged particle multiplicity resulting from LHC collisions necessitates highly granular sensors to keep the occupancy low. Moreover, the sensors are required to withstand the large radiation fluencies close to the interaction point. The substructures of the CMS tracker are based at the silicon sensor technology. However, depending on the mean distance to the collision and thus the flux of the charged particles, they differ in constituents and arrangement (Figures 2.12). The whole tracking system is kept aligned using the laser beams.

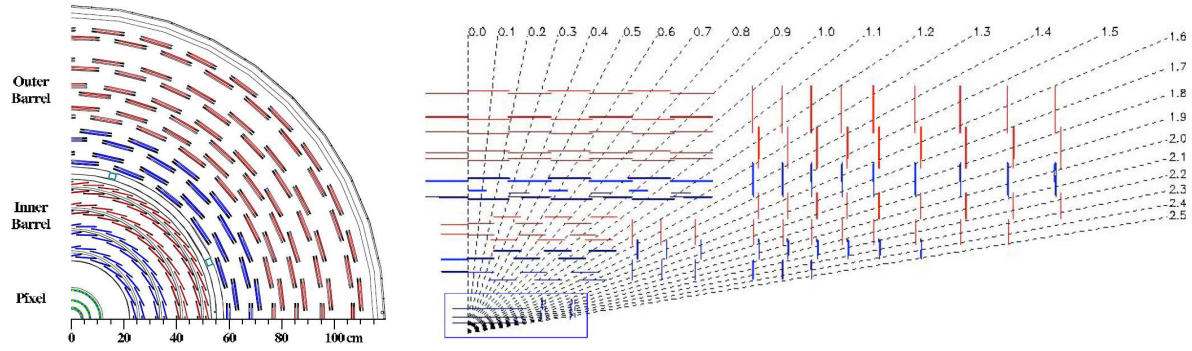


Figure 2.12: Cross section of the barrel tracker (left) and 1/4 of longitudinal view of tracking system, barrel and endcap (right). The blue box contains the pixel layers.

Pixel tracker: The environment for the innermost tracking layers in CMS is characterized by a high density and rate of charged particles ($\sim 10^7\, s^{-1}$ at $r \sim 10\, cm$). With about 66 million $100\times150\,\mu m^2$ pixels arranged at a close distance from the beam line, the Pixel tracker provides 3D space points with fine granularity on a cylindrical barrel and endcap structure. The "almost" square pixel shape design both in r-ϕ and z coordinates together with the small size of the pixels result in an optimal vertex resolution. The spatial resolution is about $10\,\mu m$ ($20\,\mu m$) for the r-ϕ (z) measurement. With the length of 53 cm the barrel part consists of three layers while there are two endcap disks on each side, Figure 2.13. The barrel layers are positioned at the mean radii of 4.4 cm, 7.3 cm and 10.2 cm. The two endcap disks extending from 6 to 15 cm in radius, are located on each side at $|z| = 34.5$ cm. They have a turbine-like geometry with blades rotated by 20 °.

Strip tracker: The CMS silicon strip tracker is the largest micro-strip detector ever built, with an instrumented area of over $200\,m^2$ and 9.6 million micro-strip channels on almost 15400 detector modules. The Silicon Strip detectors are divided in the inner barrel part (TIB), the outer barrel (TOB), the inner disks (TID) and outer endcaps (TEC). The layout of the strip tracker substructures is sketched in Figure 2.14. The TIB is made of four layers and covers up to $|z| < 65$ cm, while with six layers, TOB extends in -110 $< z <$ 110 region. Depending on the distance to the interaction point and the amount of the received radiation, TIB and TOB have silicon sensors with different size and thickness. The r-ϕ resolution for a single point ranges from 23-34 μm in TIB (which has smaller sensors) to and 35-52 μm in TOB. In z direction, the single-point

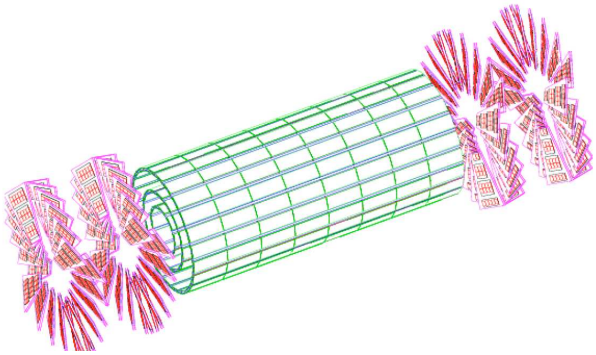


Figure 2.13: The arrangement of layers in the pixel detector, barrel and endcap.

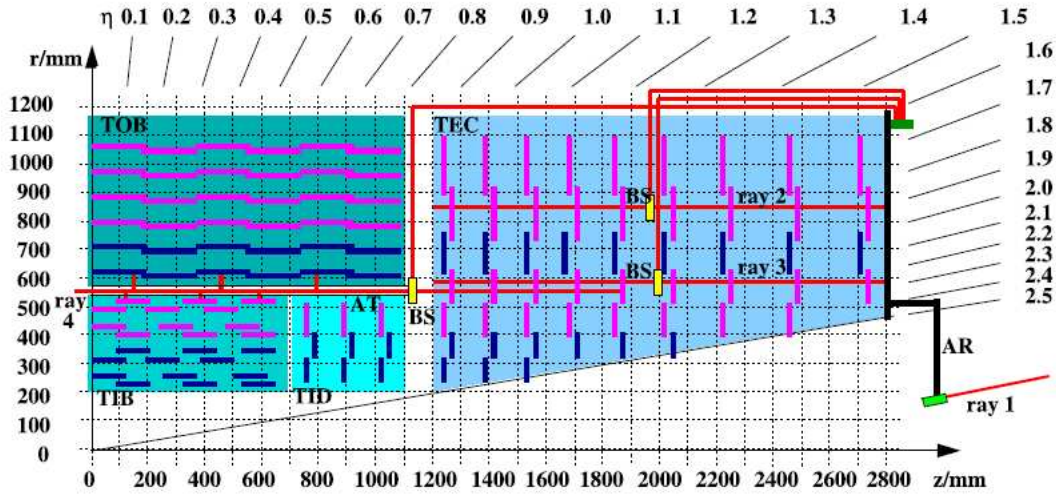


Figure 2.14: A quarter view of the strip tracker with its partitions. The red lines show the laser alignment setup .

resolution is 23 and 52 μm in TIB and TOB, respectively. Each TEC is composed of 9 disks that extend into the region 120 cm $< |z| <$ 280 cm, and each TID consists of three small disks that fill the gap between the TIB and the TEC.
The whole tracking system is positioned in a container with 5.4 m length and 2.4 m diameter and operates at a temperature of about $-20°$C. At low temprature, the radiation damages on the silicon sensors are frozen. Hence the detector quality will not gradually decrease.

Track reconstructions

The default track reconstruction at CMS is performed by the Combinatorial Track Finder (CTF) [84] which starts from the reconstructed hits. The electronic signals of a traveling charged particle in the tracker material are clustered and go through the dedicated reconstruction procedures which results in the reconstucted hits. The hit position and its uncertainty are estimated during the hit reconstruction.
The track reconstruction is decomposed in four logical steps: seed generation, pattern recognition or trajectory building, ambiguity resolution and the final track fit.

Seed generation: Seeds, the sets of reconstructed hits which are supposed to come from a single charged particle track, are generated in the first step using the pixel detector information. The pixel detector is well suited for seeding purposes since it has low occupancy and provides both r-ϕ and z-r measurements. To generate a seed, two hits from two pixel layers in different radii need to fullfil a beam constraint where the beam constraint will be removed during final steps. The combination of pixel layers is supposed to maximize the seed generation efficiency. Computationally, seed creation is much more intensive than just finding relevant hit-pairs since it includes other calculations like the construction of the covariance matrix of the track parameters. The time to generate a seed is in general about 0.3 ms.

Pattern recognition: The trajectory building is based on a combinatorial Kalman filter method [85]. The procedure starts by searching for the detector layers which are basically compatible with a given seed trajectory. The trajectory is extrapolated to these layers on the basis of the equation-of-motion of a charged particle in a constant magnetic field. As the particles travel in the detector material, the multiple Coulomb scattering and the energy loss are considered. Electrons are in addition dealt with to account for the bremsstrahlung effect (see Section 4.1). Hits are compatible with the trajectory if they fulfill a χ^2 criterion. In a given layer, there may be more than one suitable hit and the algorithm creates one trajectory for each. To account for the possibility of a particle leaving no hit in a layer, an extra trajectory candidate is created with no measured hits. This curve crosses the layer at an imaginary point, the so-called *invalid hit*.
The track parameters and the covariance matrix for all trajectories are simultaneously updated by means of the Kalman filter formalism and grown in turn to the next compatible layer. The algorithm stops either by reaching the outermost layer of the tracker or by facing a predefined stopping condition.

Ambiguity resolution: During the pattern recognition, different seeds may end up

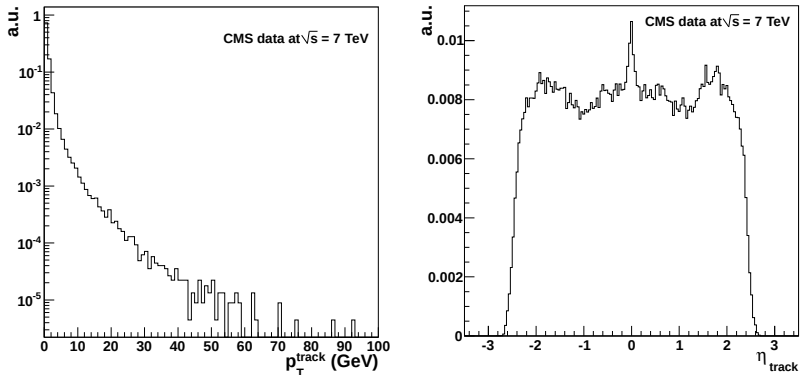
in the same track. It also happens for a seed to participate in more than one trajectory candidate. These ambiguities, are resolved by applying the ambiguity resolution once on all track candidates resulting from a single seed, and then on the complete set of track candidates from all seeds. Hence, the double counting of tracks is avoided. One can define the ambiguity resolution based on the fraction of hits which are shared between two trajectories:

$$f_{\text{shared}} = \frac{N_{\text{shared hits}}}{min\{N^{\text{hits}}_{\text{track 1}}, N^{\text{hits}}_{\text{track 2}}\}} \tag{2.4}$$

This fraction must not exceed a value of 50%, otherwise the track with the least number of hits is discarded, or, if both tracks have the same number of hits, the track with the larger χ^2 value is discarded.
Track fitting and smoothing: The track is built only after finding the last hit. However, the estimation of parameters can be biased due to the applied constraints at seeding level. To resolve the possible bias a combination of a standard Kalman filter and smoother is utilized to refit the track. The Kalman filter is initialized at the location of the innermost hit with an estimate obtained during the seeding step. The corresponding covariance matrix is scaled by a large factor in order to avoid any bias. The position estimate of each valid hit is re-evaluated and it finally leads to more precise measurement especially in the pixel modules.
The complementary smoothing stage is initialized with the result of the previous step except for the covariance matrix, which is scaled by a large factor. Smoothing starts from the outermost hits and runs backwards towards the beam line. To update the parameters at each hit, the smoother benefits from both the values it has obtained at outer hits behind and the information of the innermost hit outwards before the current hit calculated by the first filter.

During reconstruction tracks are separated in categories of expected purity based on a series of cuts on the normalized χ^2, the longitudinal and transverse impact parameters, and their significances [86]. Tracks failing the loosest selection are rejected, while those that pass the tightest selection are labeled as *high purity*.
The design performance of the CMS tracking detector is to provide a transverse momentum, p_T, resolution of about 1-2% for muons of p_T about 100 GeV, an impact parameter resolution of about 10-20 μm for tracks with p_T of 10-20 GeV and the ability to reconstruct tracks in hadronic jets with an efficiency of about 85-90% and a fake rate of less than a few percent. Detailed studies with the LHC collision data at 900 GeV, 2.36 TeV and 7 TeV have proven the good performance of the CMS tracking system [86].
To demonstrate the track properties, *high purity* tracks with $\frac{\Delta p_T}{p_T} < 10\%$ are selected within a subset of 7 TeV LHC collisions data in 2010. The events pass a set of dedicated high level online selections [1] as well as the existence of a good reconstructed interaction vertex criteria. To suppress the beam background, events are asked to have at least 10

[1] Events are selected from the electron primary data set. They are asked to have an electron of 17 GeV eneregy at online selection level. The primary data set and online event selection are defined and detailed in Section 2.3.2 and Section 2.2.4, respectively.



(a) (b)

Figure 2.15: Transverse momentum (a) and pseudo-rapidity (b) of high purity tracks in a subset of data collected in 2010. Events in the subset are further asked to pass a single electron trigger. See Section 2.2.4 for more explanation the trigger selection.

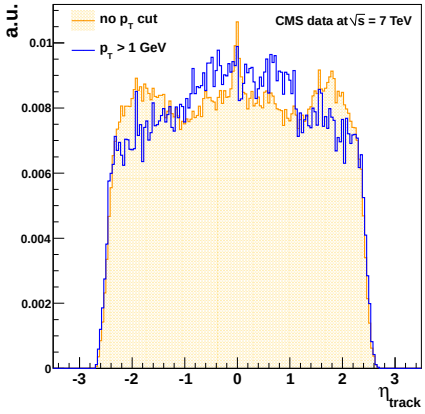
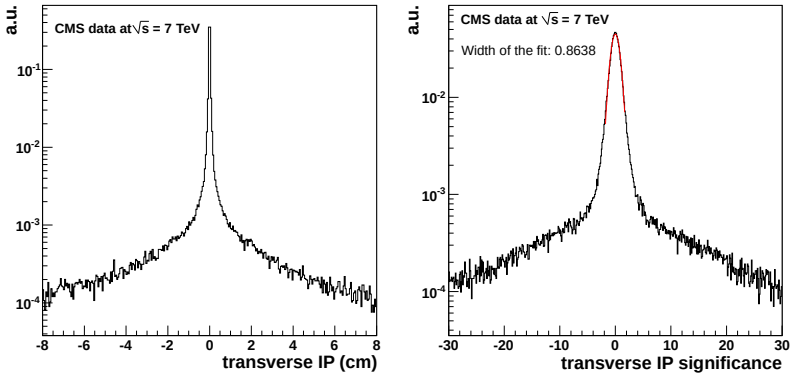


Figure 2.16: The pseudo-rapidity of high purity tracks in a subset of data collected in 2010, plotted for tracks with and without the $p_T > 1\,\text{GeV}$ requirement. Events in the subset are asked to pass a single electron trigger. See Section 2.2.4 for more explanation the trigger selection.

tracks and a *high purity* track fraction greater than 0.25.
This selection leads to the formation of a track sample which contains a lot of low p_T tracks as shown in Figure 2.15 (a). A peak is obsereved in the η distribution at $\eta_{track} \approx 0$ in Figure 2.15 (b) where it is removed after applying the $p_T^{track} > 1\,\text{GeV}$ requirement in Figure 2.16. Such tracks do not contribute in the reconstruction of the physics objects introduced in Chapter 4 since they are rejected by applying a momentum threshold of p_T^{min}.
and 2.17 Within the same track sample, the transverse impact parameter of the track



(a) (b)

Figure 2.17: The transverse impact parameter of the track w.r.t the primary vertex (a) and its significance (a) for high purity tracks in a subset of data collected in 2010. Events in the subset are asked to pass a single electron trigger. See Section 2.2.4 for more explanation the trigger selection.

with respect to the primary vertex is shown in Figure 2.17 (a). For the same tracks, the significance of the transverse impact parameter (IP/ΔIP) is plotted in Figure 2.17 (b) where the distribution is fitted with a Gaussian. The distribution however is not completely described by the Gaussian. For the non-Gaussian part, either the impact parameter or the elements of covariance matrix are badly measured. It should be noted that the tracks in Figure 2.17 are not required for $p_T^{track} > 1\,\text{GeV}$.

Vertex reconstruction

To reconstruct a vertex first the vertex candidates are found and then the estimation of the vertex parameters is optimized via a fitting algorithm, although some procedures utilize the fitting algorithms to find better vertex candidates [81]. On the other hand, there are different approaches to find the vertex candidates depending on the physics case. The vertices correspond to the point of the main interaction, primary vertices, have different properties than vertices created due to decay of particles, i.e. secondary vertices. Vertex reconstruction can also be divided into on-line and off-line reconstruction. While the former is based on the pixel hit triplets and is meant to be used for on-line event selection, the latter uses the information of fully reconstructed track collection.

The primary vertex provides an important tool to select the interesting *hard interaction* from the huge background due to long distance diffractive interactions of proton. Because of its physics importance, the reconstruction of the primary vertex is explained here. The secondary vertex reconstruction is crucial for the b-tagging algorithms and its description is postponed to Section 4.4.

Primary vertex reconstruction: To reconstruct the primary vertex in an efficient way, tracks are filtered based on their distance of closest approach to the beam (transverse impact parameter), number of hits and their normalized χ^2. The application of a cut of the order of 1 GeV on the track transverse momentum depends on the required reconstruction efficiency in different physics scenarios [87]. The selected tracks are clustered in z where the z coordinate of tracks in their point of closest approach to the beam line is taken into account. Clusters are split when there is a gap of > 1 mm and those with at least two tracks are fit with a vertex finding algorithm [81] to estimate vertex parameters as well as an indicator of the success of the fit. Tracks in the cluster which are not compatible with the found vertex are discarded. Vertices are excluded because of their either poor fit indicator or incompatibility with the beam line.
The default fit algorithm in off-line vertex reconstruction in CMS is the *Adaptive Vertex Filter* [88]. Among several vertex fitting algorithms developed and studied in CMS [89, 90], it has proven to give the best estimation of the vertex position and its errors. The algorithm assigns a weight between 0 and 1 to all of the tracks in the cluster. The weight of consistent tracks with the common vertex is close to 1. The out-lier tracks with larger distances to the vertex position are down weighted significantly, which makes the algorithm robust against the outliers. The number of degrees of freedom is defined as

$$\text{ndof} = 2 \times \Sigma_{\text{i=1}}^{\text{nTracks}} w_i - 3, \tag{2.5}$$

where w_i is the weight of i'th track. The variable is strongly correlated to the number of tracks compatible with the primary interaction region. Therefore, the number of degrees of freedom of the vertex can be used to select real proton-proton interactions. The algorithm finally reports the primary vertex position and its coordinates. The resolutions are highly correlated to the impact parameter resolution of the input tracks. They improve with the number of tracks associated to the vertex and with their average p_T.

The resolution of the primary vertex is first studied with the CMS data collected at 900 GeV and 2.36 TeV collisions [86]. It is again studied with 7 TeV data where the reconstruction efficiency is also measured [87]. The primary vertex reconstruction efficiency is estimated to be close to 100% if there are more than two tracks with transverse momenta greater than 0.5 GeV in the vertex. Within the same data set as the one used for the track properties, the primary vertices are selected according to their r-z position and ndof[2]. Vertices flagged as `Fake` are discarded. The number of "good" primary vertices in the event is shown in Figure 2.18. A Poissonian is fitted to the distribution of number of good vertices, resulting in a mean value of $\sim$ 2.3. This means that the signal collision is accompanied by extra interactions. The number of extra interactions can be estimated from the expectation value of the

[2] For details on cut values see Section 5.1.

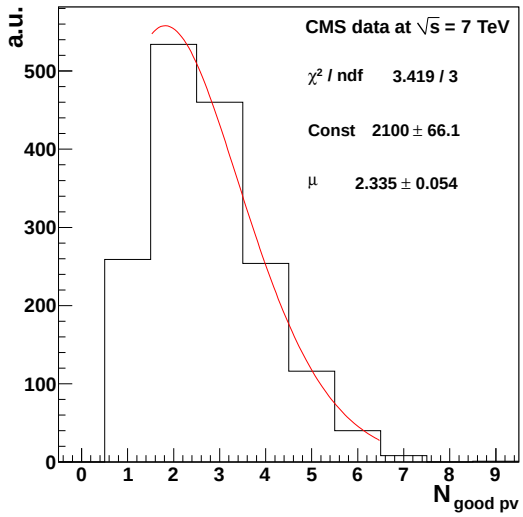


Figure 2.18: Number of good primary vertices for events in a subset of data collected in 2010. Events in the subset are asked to pass a single electron trigger (See Section 2.2.4). They are also required to have at least one good primary vertex where the vertex selection criteria is detailed in Section 5.1.

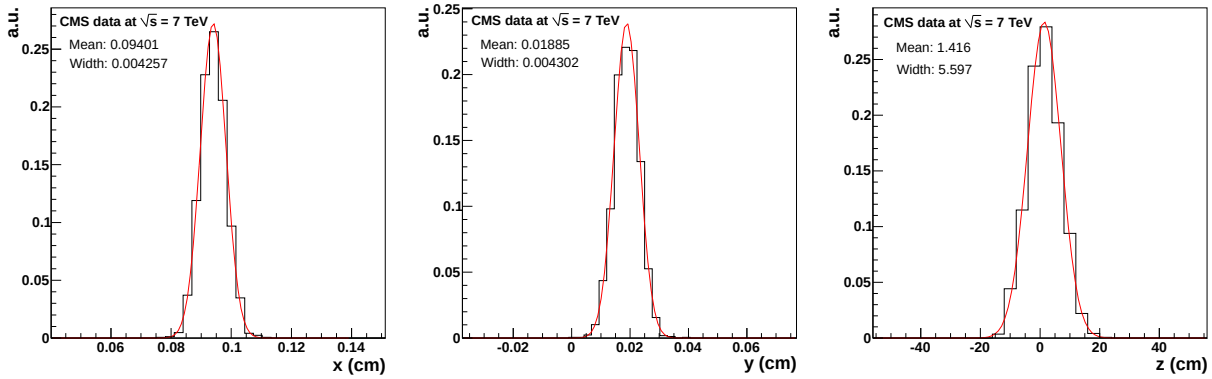


Figure 2.19: The x, y and z positions of the main primary vertex in events from a subset of data collected in 2010. Events in the subset are asked to pass a single electron trigger (See Section 2.2.4). They are also required to have at least one good primary vertex where the vertex selection criteria is detailed in Section 5.1. The mean value of the fitted Gaussian which is reflecting the average estimated position together with the width are indicated for each coordinate.

Poissonian distribution, $< N_{vtx} > -1 = 1.3$ in the current data sample. From the good vertex collection in an event, the vertex with the highest *ndof* is taken as the main primary vertex and becomes a reference point to measure the distances and impact parameters within the event. The x, y and z position of the main primary vertex is plotted in Figure 2.19. The uncertainty on the vertex position in each direction is indicated in the width of the fitted Gaussian. Figure 2.20 shows that the resolution of the primary vertices which is improved in higher track multiplicities since the vertex position becomes more constrained by additional tracks.

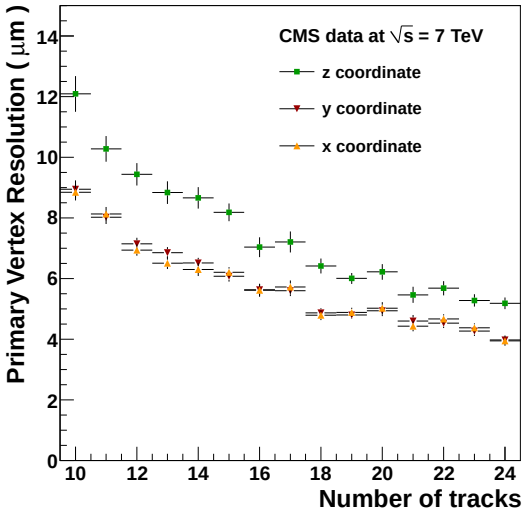


Figure 2.20: The resolution of primary vertex versus the number of associated tracks. The primary vertices are selected from events in a subset of data collected in 2010. Events in the subset are asked to pass a single electron trigger (See Section 2.2.4). They are also required to have at least one good primary vertex where the vertex selection criteria is detailed in Section 5.1.

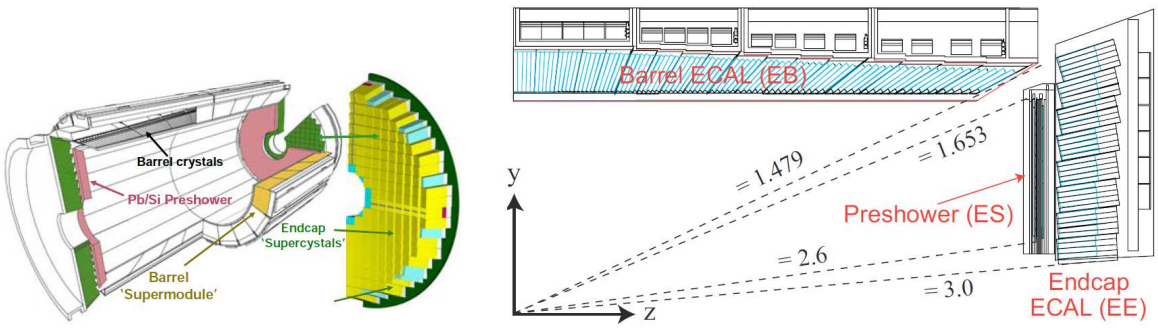
2.2.2 The CMS calorimeter system

Finely grained and with high resolution optimized for the detection of the Higgs boson through its electromagnetic decay, the Electromagnetic Calorimeter (ECAL) plays a crucial role in the search for new physics as well as in precision measurements in the Standard Model. A hermetic and homogeneous structure based on 75848 lead tungstate ($PbWO_4$) scintillating crystals is devoted to the electromagnetic energy measurements in CMS [81]. The advantages of the lead tungstate crystals are their short radiation length ($\chi_0 = 0.89\,\text{cm}$) and Moliere (2.2 cm) radius. The radiation length in the material is both the mean distance over which a high energy electron loses all but 1/e of its energy by bremsstrahlung, and 7/9 of the mean free path for pair production by a high energy photon. The Moliere radius is a characteristic constant of a material giving the scale of the transverse dimension of the fully contained electromagnetic showers initiated by an incident high energy electron or photon. One needs to add the fast

response (80% of the light is emitted within 25 ns) and radiation hard properties to this list. Low light yield, the disadvantage of $PbWO_4$ is compensated by the use of photodetectors with reasonable gain and the ability of operation at high magnetic field. While in the endcap the vacuum photodiods (VPT) are installed, in the barrel the silicon avalanche photodiods (APD) are in use and they require a temperature stability at 0.1°C.
As illustrated in Figures 2.21 in the barrel part (EB) extended in $|\eta|$ <1.479, ECAL consists of 36 supermodules each covering half the barrel length. Every supermodule, has 1700 crystals which are quasi-projective and are covering 0.0174 in η and ϕ. The ECAL endcap (EE), occupies the range of $1.479 < |\eta| < 3.0$ with two Dees structured aluminum plates, from which are cantilevered structural units of 5×5 crystals, known as supercrystals.
The ECAL energy resolution measured in electron test beams is parametrized as [81]

$$\frac{\sigma(E)}{E} = \frac{2.8\%}{\sqrt{E(GeV)}} \oplus \frac{12\%}{E(GeV)} \oplus 0.3\% \tag{2.6}$$

for electrons incident on the center of crystals where the values are obtained from a Gaussian fit to the reconstructed energy distributions. From left to right: The first therm is the stochastic term, the second corresponds to the noise and the last one is the constant. For photons with energy above 100 GeV the energy resolution is dominated by the constant term.
The ECAL performance has been promising during the first LHC collisions in 2010. The percentage of fully working channels in EB and EE is about 99.30% and 98.94%, respectively. The constant term in energy resolution depends a lot on the stability of the system in which the temperature stability of crystals and photodetectors contribute. The temperature stability over two months has been measured to be about 0.0076°C and 0.015°C for EB and EE respectively [91].
In front of most of the fiducial region of each endcap is a preshower device. The prin-



(a) (b)

Figure 2.21: Three dimensional (a) and longitudinal view (b) of ECAL

cipal aim of the CMS preshower detector (ES) is to identify the closely spaced photons from neutral pions in the endcaps within a fiducial region $1.653 < |\eta| < 2.6$ [81], hence

allowing a reduction of the backgrounds to di-photon decay channel of the intermediate mass Higgs boson. It also helps the identification of electrons against minimum ionizing particles, and improves the position determination of electrons and photons with its superior granularity [92]. The ES is a sampling calorimeter comprises 2 silicon strip planes in each endcap. The silicon strips are orthogonally located, one after a lead radiators plate which initiates electromagnetic showers from incoming photons/electrons. The energy deposited and the transverse shower profiles is measured by the silicon strips. The performance of the preshower detector is investigated in 7 TeV collision data where the information of preshower together with the ECAL is used to observe the π^0 mass peak with a very good resolution [91]. The preshower detector has been operational with an efficiency of 99.8%. Designed to measure the energy of quark-gluon made particles, the Hadron Calorimeter at CMS is required to minimize the non-Gaussian tails in the energy resolution and to provide good containment and hermeticity for $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement. In terms of radiation length, the HCAL maximized the material inside the magnet bore while a complementary part of the HCAL, hadron outer (HO), is outside the magnet. The HCAL material is the brass alloy not only for its short interaction length which fulfills the material maximization requirement but also because of its non-magnetic characteristics. Looking for the least possible devoted space, the HCAL read-out system is made up of plastic scintillator tiles with embedded wavelength-shifting light fibers, each tile with a thickness of 3.7 mm sandwiched between the brass layers. The light is detected by the hybrid photodiodes (HPD's) that are capable of providing gain and operating in the high axial magnetic field of the detector. Both thickness and η-ϕ segmentation of the tiles vary in different parts of the HCAL barrel (HB) and endcap (HE)(Figure 2.22). Covering -1.4 $< \eta <$ 1.4 inside

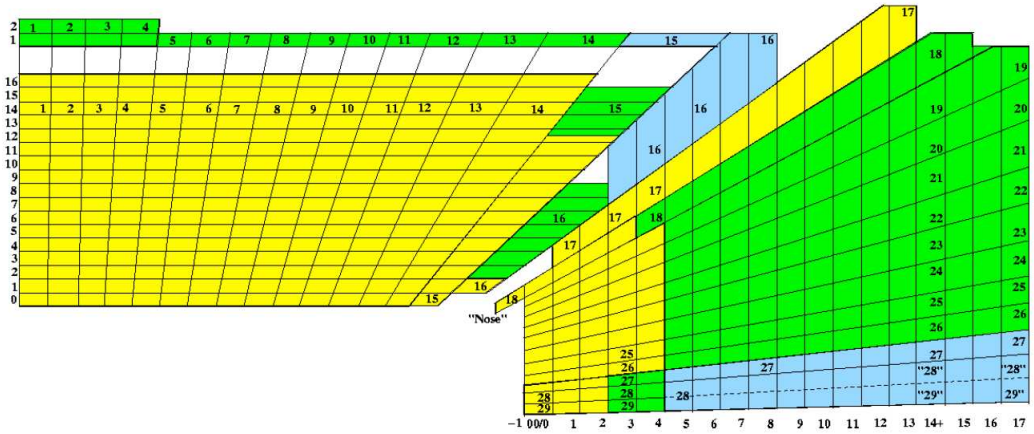


Figure 2.22: Longitudinal view of CMS hadron calorimeter

the magnet, the HCAL barrel consists of 32 towers where each tower includes layers of active materials with $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ segmentation, facing the incoming particles. The HCAL endcap extends in $1.3 < |\eta| < 3.0$ range. It comprises 2304 towers with different segmentations.

Located out of the magnet and inside the muon system, the hadron outer calorimeter has a projective geometry that is influenced by the muon system. It covers the

$|\eta| < 1.26$ range and serves as a "tail-catcher" for hadron showers leaking through the rear of the calorimeters. With scintillators of 10 mm thickness, it increases the effective depth of the HCAL to over 10 interaction lengths. Hence, it reduces the tail of the energy resolution function and improves the resolution of the calorimeter based E_T^{miss}. The geometry of the scintillators in the HO is similar to those in the HB towers.
Different checks on the HCAL operation are done with test beams in which different subsystems were exposed to beams of electrons, pions, protons and muons [93]. The energy resolution of the ECAL+HCAL is parametrized as

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E(GeV)}} \oplus b \tag{2.7}$$

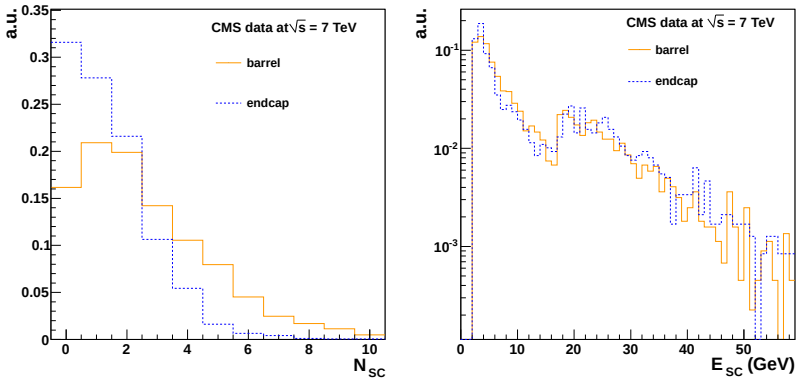
where a is the stochastic term and b is the constant. The reported values for HB are $a = 0.847 \pm 0.016\, GeV^{1/2}$ and $b = 0.074 \pm 0.008$ and the results for HE are similar. The muon test beam results are compared to cosmic ray muons to test the absolute energy scale in the HCAL [94]. In collision data, CMS looked at the jet properties and $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution to investigate the performance of the HCAL [95–98].

Reconstruction of Superclusters in ECAL

The high magnetic field in the experiment influences the electrons which are interacting with the ECAL material and makes their energy spread via bremsstrahlung in the ϕ direction. The ECAL energy deposits are grouped together as clusters and then special algorithms build a *SuperCluster* of clusters extending in ϕ to take all of the electron energy into account. Superclusters are then used in the electron reconstruction algorithm (Section 4.1). The knowledge about the formation of superclusters can illuminate the further discussions presented in Section 4.1.1, regarding the electron shower shape in electron identification.
The Hybrid [99] and "Multi5×5" [100] algorithms are used in CMS to reconstruct the superclusters in the ECAL barrel and endcap respectively. While with the Hybrid algorithm the entire supercluster is made first before it is decomposed to clusters, in the Multi5×5 algorithm the supercluster reconstruction is based on the formation of a basic cluster.

- *Hybrid algorithm:* The Hybrid algorithm starts by looking for a "seed" crystal. The energy deposit at the seed is a local maximum and is greater than some global threshold, $E_{\mathrm{hybrid}}^{\mathrm{seed}}$. Two single crystals are added to each sides of the seed in η and if $E_{seed} > E_{wing}$ (another threshold), another crystal in each side joins this so-called "domino" structure. In the $\pm\phi$ directions, the algorithm adds up other dominoes for N_{steps} and removes dominoes with $E < E_{threshold}$ to account for the detector noises. This complex of ECAL crystals is then clustered in ϕ where each distinct cluster of dominoes is required to have a seed domino with energy greater than E_{seed}. The whole structure which is now broken into clusters is the aimed supercluster. The Hybrid algorithm is originally developed for the reconstruction of relatively high energy electrons in the barrel and has been tuned afterwards to reconstruct the electron showers down to p_T=5 GeV.

- *Multi5×5 algorithm:* The local maxima endcap crystals in terms of energy are the start point of *Multi5×5* if $E_{local\,max} > E_{seed}$. Around the most energetic seed, a 5×5 crystal cluster is made. Clusters do not share crystals except the ones on the edge of a cluster that are still allowed to seed other clusters. To make the supercluster, basic clusters are sorted by energy and the most energetic one with $E > E_{\text{seed}}^{\text{cluster}}$ serves to seed the first supercluster. To be added to the seed, the 5×5 crystal clusters need to be in $\Delta\eta_{multi} \times \Delta\phi_{multi}$ region around the seed. Both $\Delta\eta_{multi}$ and $\Delta\phi_{multi}$ are parameters of the algorithm. The endcap superclusters are later combined with energy deposits in the preshower detectors.



(a) (b)

Figure 2.23: The number of superclusters, (a), and their corrected transverse energy distributions, (b), in the barrel and endcap. The superclusters are taken from a subset of data collected in 2010 where events in the subset are asked to pass a single electron trigger (See Section 2.2.4). The shoulder in energy distribution is due to the online selection criterion.

The energy of the supercluster is the sum of the energy of its contituents and different sources can introduce variations in this clustered energy. Therefore the energy of superclusters are corrected with dedicated algorithms before they are used in the reconstruction of high level objects [81]. The supercluster position is determined by the energy-weighted mean of the clusters position. In each cluster, the position is defined as the energy-weighted mean of constituent crystals position

$$x_{cluster} = \frac{\Sigma x_i . W_i}{\Sigma W_i}. \tag{2.8}$$

In the electron showers the energy density in the ECAL crystals falls exponentially, hence with a simple weighting the position is biased toward the shower cone. To recover for this bias, the crystal weight is defined as a logaritmic function of its energy

$$W_i = W_0 + log\frac{E_i}{\Sigma E_j}. \tag{2.9}$$

Figure 2.23 illuatrates the number of superclusters and the superclusters corrected transverse energy within the same dataset as for the tracks and primary vertices. Categorized into the barrel and endcap superclusters, the candidates are required to have a corrected transverse energy greater than 2 GeV. To exclude the ECAL gap, superclusters with $1.4442 < |\eta_{supercluster}| < 1.566$ are rejected. More superclusters are reconstructed in barrel than in the endcap while there is no significant difference between the energy distributions of the two categories.

Calorimeter towers

Larger in size, each HCAL tower covers several ECAL cells in η-ϕ plane (1:25 in barrel). Hence, jets of hadrons fire more than one ECAL tower per HCAL tower and it leads to the definition of calorimeter energy towers, *CaloTowers* as objects link matching clusters in ECAL and HCAL to produce a projective tower in the calorimetry system. While for calotowers in barrel 5×5 ECAL crystals are associated to one HCAL tower, the different geometry in the endcap region requires a more complex matching between the ECAL and HCAL cells. Calotowers are the main ingredients of the jets, served as input for calorimeter jet reconstruction algorithms(see Section 4.3). Figure 2.24 depicts the η-ϕ distribution of calotowers towers in the mentioned subset of CMS 2010 collision data at 7 TeV.

To reduce the calorimeter noise from the electronic readout system, individual cells

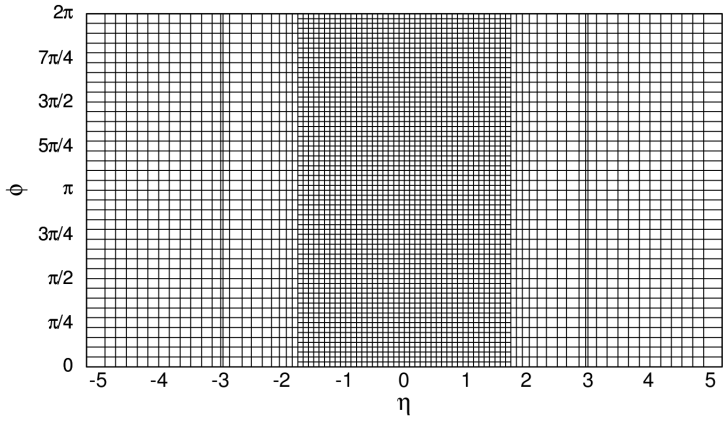


Figure 2.24: The map in η-ϕ of calorimeter towers.

are required to pass an energy threshold according to the scheme in Table 2.1 before bulding the calotowers [97]. In addition, calorimeter towers with $E_T^{calotower}$ <0.3 GeV are rejected to suppress the energy contributions from multiple interactions in the event.

2.2.3 The muon system

Muons provide a very clear signature and they are relatively easy to measure and identify. Hence, physics signals including muons are very important in the physics

Section	Threshold (in GeV)
HB	0.7
HE	0.8
HO	1.1/3.5 (Ring 0/Ring 1,2)
EB	0.07 (per crystal, double sided)
EE	0.3 (per crystal, double sided)
EB Sum	0.2
EB Sum	0.45

Table 2.1: Calorimeter cell thresholds to supress the noise. In ECAL, an overall threshold is applied in addtion on the sum of the energy in crystals associated to a tower.

program of the pp collisions. The main requirements for the design of the CMS muon detector are based on studies to detect and identify processes where a Higgs boson decays to ZZ$\rightarrow 4\mu$. The p_T resolution of the muons in the muon system is expected to be around 9% for $p_T \sim 200\,\text{GeV}$ and from 15 to 40% for $p_T \sim 1\,\text{TeV}$ muons. Considering also the tracker information, the resolution can reach 1% for low p_T muons and 5% at $p_T \sim 1\,\text{TeV}$.

The CMS experiment benefits from three different types of muon detectors all based on the gas ionization in its muon system [81]. While the drift tubes (DT) in the barrel and the cathode strip chambers (CSC) in the endcap provide a good position resolution, the resistive plate chambers (RPC) are specific for the timing in both barrel and endcap. The choice of DT's and CSC's for barrel and endcap respectively is driven by both muon and neutron induced rates which are lower in the barrel together with the magnetic field that is stronger in the endcap. Although the RPC's are not precise in determining the position, they have a fast response and are suited the to identify the correct bunch crossing.

In drift tube chambers, positively charged wires are in the gas volume. They receive the signal once an incoming particle ionizes the gas and free electrons are gathered in the positive wire. In total 250 chambers are arranged in the barrel part ($|\eta|$ <1.2) of the muon system, Muon Barrel (MB), in 4 layers, so-called "stations", which are located from $\sim 4.0\,\text{m}$ to $7.0\,\text{m}$ from the beam axis (see Figure 2.25). To detect the high p_T muons even produced near the boundaries of the sectors, chambers in different stations are staggered by half a cell. In each one of the first three stations (MB1-3), there are 12 layers of drift tubes grouped by 4 with wires in parallel in each group, i.e. in each "superlayer". The superlayers can measure the position in the r-ϕ plane if their wires are along the beam axis. Perpendicular to this direction, the z position is measured. In the first three stations, a z-measuring superlayer is located between two r-ϕ superlayers. The fourth station does not have the z-measuring plane. Each station is expected to give a muon vector in Cartesian coordinates and apart from that, to provide a ϕ measurement. The space resolution is better than $100\,\mu m$ where ϕ is measured with a precision of $\sim 1\,\text{mrad}$. To provide the timing, each DT layer is either sandwiched by two RPC's like in MB1-2 or coupled to only one RPC (MB3-4).

In cathode strip chambers arrays of positively-charged wires crossed with negatively-

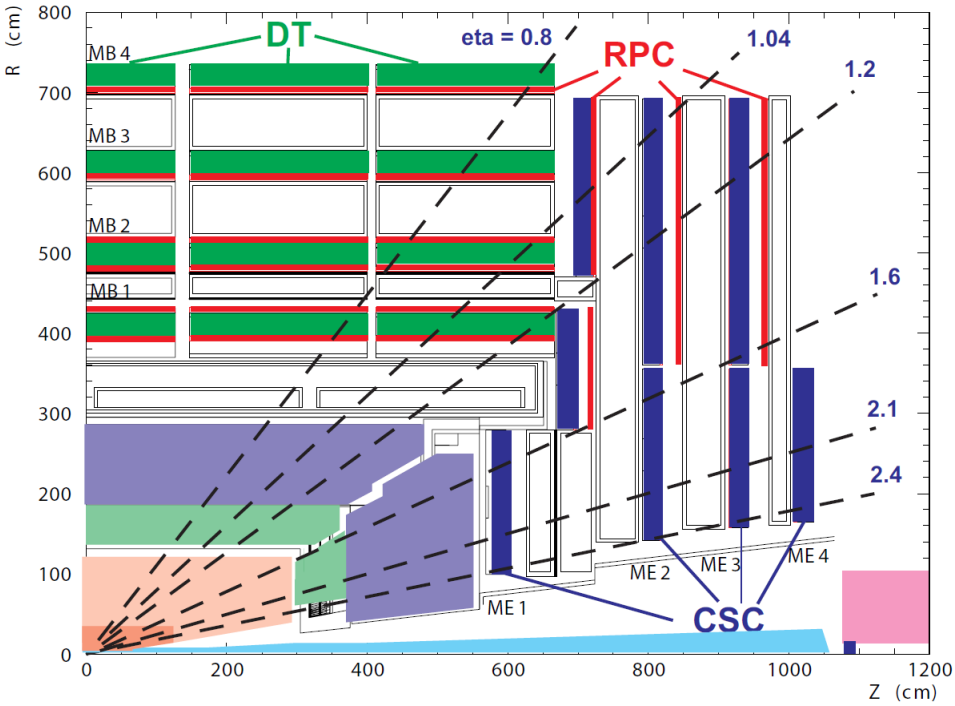


Figure 2.25: One quarter view of CMS muon system

charged copper strips are positioned within a gas volume. The gas ionized by the incoming particle, move toward the strips while the avalanche of the electrons are attracted by wires, so a 2D position measurement is possible. The endcap, extending the coverage of the muon system to $|\eta|$ <2.4, comprises 468 CSC's in total divided into four disk stations (ME) in each side. The CSC's have a trapezoidal shape and consist of 6 gas gaps in which the cathode strips are radial. The CSC's are put together to make the endcap rings in such a way that except for the third ring of the first station, they have ϕ overlap to avoid gaps in muon acceptance. CSC's have a fast response, useful for the online selection. However providing (r,ϕ,z) measurement, their position resolution is coarse. There, the center-of-gravity of the charge distribution induced in the cathode strips leads to a more precise position estimation. The spatial resolution of each CSC is $\sim 200\mu m$ ($100\mu m$ for the first ring in ME1) while the angular resolution is about 10 mrad.

In the endcap, RPC's are mounted up to $|\eta|$=1.6 to help not only in the timing but also in resolving the geometric ambiguities specially in the first station of CSC's where the barrel-endcap overlap exists.

After the test beam results [81] and in the absence of collisions, cosmic muons provided a golden natural source to investigate the muon detector performance. CMS performed detector commissioning in different periods of time with and without the magnetic field [101, 102]. One can point to the average single cell efficiency of 98% for DT's beside the 200-260 μm resolution for a single layer [101]. Combining the muon system information with the tracker system, the resolution reaches 100 μm. Confirming the test beam results, the track segment finding efficiency in CSC's is well above 99%. The CSC resolution in the first ring of ME1 varies from ∼100 to 125 μm depending on

the distance from the beam line. Close to the beam axis, the strips are narrower and hence give a better resolution.

Muon reconstruction

Starting from the muon system, the first step is to reconstruct the points, "hits", in DT's, CSC's and RPC's. Depending on the detector characteristics, different algorithms are used for the hit reconstruction. In the DT's, the distance of the particle to the wire is obtained by converting the drift time to a drift length. A fit on a cluster of strips helps to find the position in the CSC while in the RPC, assuming a uniform probability for the incident point over the plate, the center of gravity is taken as the hit position.
The next step, is to make segments using the hits. Again the approaches differ regarding the local detector properties. In the DT's, segment candidates are made with a set of aligned hits compatible with a track pointing to the nominal interaction point. The best segment among those candidates sharing the hits is selected according to the number of hits and the χ^2 of the segment track fit. Using the segment information, hits are updated and the segment is then refitted. Within the 6 layers of the CSC's, a line connects the two hits in the first and the last layer if the two points have a separation less than 1 cm in r-ϕ plane. Then, other hits are successively added and the linear fit is updated.
In the standard CMS reconstruction for pp collisions [81], beside the tracks reconstructed independently in the silicon tracker (tracker tracks), the "standalone-muon tracks" are made in the muon spectrometer and the muon reconstruction can benefit from both.

Standalone Muon reconstruction: Only the muon subdetectors participate in the standalone reconstruction. Providing direction, position and momentum information, track segments from the innermost muon chambers are used to seed the muon track. The Kalman-Filter technique [85] predicts the muon trajectory and the predicted values are updated according to the next segment. The momentum resolution is improved by means of a beamspot constraint in the fit for the collision data. The resulting tracks are referred to as standalone muons.

Tracker Muon reconstruction: All tracker tracks with $p_T > 0.5\,\text{GeV}$ and with an energy greater than some threshold are extrapolated to the muon system while the energy loss and the uncertainty arises from the multiple scattering are accounted for. Finding one muon segment in the muon system matched to the extrapolated track is enough for the candidate to be considered as a muon track. The efficiency of this reconstruction is good for low energy muons ($< 5\,\text{GeV}$).

Global Muon reconstruction: The trajectory of a "standalone muon" is extrapolated from the innermost muon station to the outermost tracker layer and is matched with a tracker track. The effects of multiple scattering and energy loss in the material plus the influence of the magnetic field are taken into account. The joint track is then fitted

in the region of interest using the Kalman-Filter technique. The region of interest is defined according to the muon trajectory parameters and their uncertainties, assuming the muon originated from the primary interaction point. For muons with $p_T > 200\,\text{GeV}$, the momentum resolution can be improved via the global-muon fit compared to the tracker-only [81, 103]. The performance of the CMS muon reconstruction is studied both with cosmic rays [103] and collision data [104].

2.2.4 Online event selection process

The nominal bunch crossing rate at the LHC, 40 MHz, corresponds to $\sim 10^9$ interaction per seconds. To match to the available data storage and keep the information of only those interactions with a potential physics interest, some online selections must be applied. This rate reduction down to 100 interactions/sec is carried out in two main steps. While the first step, *Level-1 trigger*, is based on custom hardware decisions, the second part,*High Level Trigger*, use more detailed information in more sophisticated algorithms that approach the quality of final reconstruction. The total time specified to the Level-1 trigger selection is $3.2\,\mu s$ of which a latency of more than $\sim 2\mu s$ is allocated to the data transportation from the front-end detector electronics to the trigger boards and vice versa.

The Level-1 trigger

Using only the information of the muon system and the calorimeters, the Level-1 trigger decreases the rate by a factor of $\sim 10^3$. During a time of $1\,\mu s$, the decision is made upon the presence of "trigger primitive" objects like photon, electron, muon and jets with E_T or p_T greater than some threshold. The E_T sum and $E_{\text{T}}^{\text{miss}}$ can be checked as well. The low-resolution and low-granularity data is utilized in the trigger object reconstruction at this level. As an example, in the ECAL barrel the energy deposits in 5×5 ECAL crystals is measured and if it exceeds the threshold, the object is considered as an electron/photon candidate [81]. Based on the physics and technical motivations, the Level-1 trigger can request for qualification criteria like isolation for the electron/photon or ask for more than one qualified trigger object.
The designed rate for the Level-1 trigger is 100 kHz, set by the average time needed to transfer full detector information through the readout system. However, at the startup condition this rate is limited to 50 kHz.

The High-Level Trigger

During the time the Level-1 trigger makes the decision, the high-resolution data is held in pipelined memories from where it is transferred to the temporary memories of the front-end readout for further processing. The data is placed in teh Data Acquisition system (DAQ) that builds the event with the size of about 1.5 MB. These events go under high-level trigger selection in different processors. On the HLT farm, the rate can fall down to 100 Hz. To speed up the procedure even more, the HLT software reconstructs only the relevant objects instead of the full event reconstruction. Moreover,

it applies some virtual trigger levels such that the calorimeter and the muon system information are checked first. Only then the tracker data and eventually the full event information are used. Due to the access to event data with full granularity, simple b- and τ-tagging can be done at this stage.
In the case of electron HLT selection, the event goes through three sublevels: first, based on the Level-1 object the supercluster is formed and its corrected energy is asked to be higher than the threshold. Then, the consistency of relevant pixel hits with the supercluster is checked. The object fails as an electron candidate if the consistency does not exist. At the end, the position and momentum/energy of the supercluster should be matched with that of the tracker track. The requirement of isolation is also implemented to be used when relevant [3].
Various triggers aiming for different physics goals are implemented in the HLT software of the CMS experiment. The thresholds used in these triggers should be optimized in the sense that they need to be neither too low and saturate the trigger nor should they be too high and reduce the physics signal of interest. Another challenge is to fit all triggers within the predefined bandwidth without tightening the cuts too much. Cross-triggers are developed in which combinations of qualified trigger objects are used for triggering both at the Level-1 and within the HLT algorithms (see for example [105]).
The performance of the CMS HLT has been investigated from the timing point of view. The full HLT paths have been run over simulated QCD samples and once more on signal-like simulated events to ensure an unbiased timing estimation. The average time per event accepted by the Level-1 trigger is $42.9 \pm 5.6\, ms$, in agreement with the capabilities of the HLT farm [106].

2.3 Data taking and computing in CMS

In addition to the detector design, a robust and efficient computing infrastructure is required to support the final physics demands so the end users in the chain of data taking, are provided with reliable data to analyze. In the first place, the computing system has to reduce the rate of delivered data to a reasonable value for storing and further processing while the recorded data is expected to be of physics interests. To guarantee the smooth detector operation and the excellence of physics results, the deployment of a monitoring system is also crucial. On the other hand, there are many physicists in different institutes all over the world collaborating with CMS and hence need an access to the CMS recorded data. Therefore, the experiment has to provide a worldwide computing network dedicated to both data storage and analysis purposes. Finally, the data needs to be well-modeled to make all of these requirements attainable.
As explained in Section 2.2.4, the CMS experiment reduces the rate of the delivered data in a basically two-step procedure. Once the Level-1 hardware trigger system accepts data considering the detector data primitives and/or technical and beam conditions, the raw detector data is retrieved by the DAQ system, where it is aggregated to event data and delivered to HLT processors. Having met the criteria of HLT, events are passed to several instances of the Storage Manager (SM) application which is responsi-

[3] The electron reconstruction and the isolation definition are detailed in Section 4.1 and Section 4.1.2, respectively.

ble for storing the data on disks. Along with the storage procedure, the data events go through different reconstruction steps, categorizations and skimming algorithms. They are stored on disks in numerous nodes within the CMS worldwide network in different data tiers and finally become accessible on CMS computing network for the end users. In this section, first the CMS Event Data model and analysis framework together with the data tiers in CMS are explained in Section 2.3.1. The data categorization after the online selection is then reviewed in Section 2.3.2. CMS computing environment is briefly addressed in Section 2.3.3. Last but not least in Section 2.3.4 the CMS Data Quality Monitoring (DQM) system containing the example of TopDQM is reviwed.

2.3.1 The Event Data Format of CMS

The CMS Event Data Model [107] is a software concept that can be defined as a box containing all the information of a *pp* collision. During the processing data goes through different software modules in the format of events, so the information about the collision is accessible via the event format. Events are first formed by the DAQ system using the *RAW* data of the detector. All the information added to the event must be timely coherent in the sense that they have to come from the same collision. After HLT, the high-level trigger objects are added to the event content. Starting from the RAW data, the information is being refined, higher level objects are formed and what is not needed is being dropped. This defines the CMS **data tiers**. The next data tier is *RECO* which refers to fully reconstructed objects like tracks, vertices, jets, etc. It preserves links to the RAW information. Although the event size is dropped from 1.5 MB in RAW to 0.25 MB in RECO, it is still massive to be shipped easily and also part of its content is not interesting for the physics analysis. Analysis Object Data, *AOD*, is derived from RECO information and is dedicated to physics analysis in a convenient, compact format. It further reduces the event size to 0.05 MB. Both RECO and AOD data tiers have the least possible space required while they provide enough flexibility. Specially due to the fact that they contain objects which link to each other. By the way, using the links needs more expertise in terms of programming. Hence the Physics Analysis Toolkit, *PAT* data format is developed to ease the analysis task as much as possible. The PAT event content can be defined by the user so it is not a data tier.
The event content is processed quantitatively in the framework of CMS SoftWare, *CMSSW*. CMSSW is coherently used for the online trigger filtering, monitoring, offline event reconstruction and the physics data analysis. To maximize the flexibility, CMSSW has a modular structure and is fully object oriented based on C++ language. The modules are configured by Python code and are executed in a user-defined sequence. The CMSSW executable, *cmsRun*, is relatively lightweight since only the required modules are loaded at run time. Internally, CMSSW uses the ROOT [108] framework, hence the CMS data formats are ROOT-aware. It means if a framework can load ROOT-friendly CMS shared libraries within a ROOT session, the CMSSW data format will be recognized by ROOT. Based on this idea, the so-called framework-light (FWLite) is developed which is lighter than the full CMSSW framework.

2.3.2 Data categorization for storage and analysis

By definition, individual events are put in files in the Storage Manager in such a way that they are grouped into *stream*s according to the specific HLT paths they have fired. Several streams can be defined based on their offline usage (e.g. express streams, physics streams, etc.) and since the same HLT paths can feed different streams, the individual streams can overlap. The streams are written in a binary data format and referred to as *streamer files*. Within a stream, sets of HLT paths which select similar signatures are regarded for further categorization of data events and the formation of *primary dataset*s (PD's). More specific HLT selections are applied to make the *secondary dataset*s (SD's) that are more specialized for the physics analyses. While both primary and secondary datasets are formed based on the high level trigger information, *central skim*s (CS's) are produced with additional cuts on the reconstructed physics objects. The group skims in which one looks at the data in the region of physics interests, can be promoted to central skims if they contain a small fraction ($< \sim 10\%$) of the primary dataset[109] so they will be centrally produced by CMS which is more convenient.

2.3.3 CMS distributed computing system

The CMS computing system supports the computing requirements of data storage, processing and analysis. It has a multi-tiered architecture [110] based on a distributed infrastructure (Grid) that shares computing resources, CPU and disk space, among a dynamic collection of institutes. Aiming for LHC physics, the Worldwide LHC Computing Grid, WLCG [111], has provided the building blocks for the CMS computing network. The CMS distributed computing system has three tiers of which the Tier-0 is hosted at the CERN computing center. Containing about 20% of all available resources at CMS, the Tier-0 performs the initial processing of the data coming from the detector. The *streamer files* are converted to ROOT-based data format and are split to primary datasets. The primary datasets are then reconstructed and stored on the tapes in Tier-0 [112]. About 40% of the CMS computing resources belong to the Tier-1, where copies of PD's are stored for reprocessing purposes. There in fact, PD's can be re-reconstructed with the updated software or with the new calibration and alignment constants. Moreover, the SD's and CS's are produced at Tier-1's. France, Germany, Italy, Spain, Taiwan, United Kingdom and United states of America host the CMS Tier-1 sites.

Datasets stored at Tier-1 sites are transferred to about 50 Tier-2 sites where the group skims can be performed and the final analyses for physics achievements are carried out. While in 2010 the primary datasets were available on Tier-2 sites, the central skim production will be the only way to access the full PD's afterward so the Tier-2 resources can cope for the increasing luminosity.

For users who have the relevant certification, the data files at Tier-2 sites are accessible via the Cms Remote Analysis Builder (CRAB) [113, 114]. With the CRAB application, users with no knowledge of grid infrastructure have the possibility of creating, submitting and managing job analyses into the grid environment. Once the analysis code is developed interactively, given the required information like the name of the dataset, the analysis parameters, etc. CRAB finds the sites hosting the data sets and

handles the resources availability, job creation and submission, status monitoring and output retrieval.

2.3.4 CMS Data Quality Monitoring

Aiming for a homogeneous monitoring environment across various applications related to the data taking at CMS, the Data Quality Monitoring (DQM) system is arranged [81, 115]. The primary goal is to ensure the quality of physics data collected in general data acquisition. The main requirement for DQM has been the maximum flexibility so it can be used by different groups interactively, e.g. update of histogram code on request. In addition, DQM has to have the least interference with data taking, triggering and data storage. Hence possible problems in DQM can be isolated and CPU-intensive tasks will not slow down the general data taking in DAQ and HLT.

DQM architecture

Following the requirements above, the DQM framework is designed in three main layers. The basic DQM components, monitoring elements (ME's), are produced in *source*s which retrieve the event data information from several storage managers, SM's. Sources are connected to a *collector* in a many-to-one structure. The collectors are responsible for the redistribution and periodic update of monitoring elements on one hand and on the other hand they act like servers for the final users, the *client*s. Therefore, the client is blind to the source, the source remains stable and will not be slowed down. Also, the quick transfer of monitoring information from sources to collectors is facilitated. A client needs to subscribe to the collector, requesting for a subset of monitoring elements which are finally shown to the end user in a histogram format. The Graphical User Interface, GUI, serves the centralized visualization of the DQM histograms.
Although DQM works with event data, it does not give access to individual events since it has a statistical nature. What is seen on histograms is collected over a period of time so punctual problems are not spotted.

DQM operation

Data Quality Monitoring is performed online and offline. While the online DQM is carried out at P5 to support the prompt reaction about the detector status based on a subset of data, the offline monitoring is done with some latency at CERN, DESY and Fermilab and has two main steps which finally end up in data certification for physics analyses. In the first step a subset of data, the express stream, is reconstructed and monitored within about an hour. The goodness of run is examined in terms of the reconstruction software, calibration and alignment constants. After about 48 hours the full dataset is reconstructed with better constants obtained in the previous step. Another offline monitoring sequence is performed when the data is re-processed and re-reconstructed with new software releases.
To certify the goodness of data in both online and offline monitoring, first some test algorithms are run automatically, checking different parts of the detector and reporting possible problems in the DAQ, etc. Then a shifter is in charge to investigate the

goodness of data by looking at the most important distributions and a quality indicator is assigned to each subsystem. The manual certification is registered in the Run Registry that is the central workflow tool for DQM, tracking certifications and quality knowledge. It has an interface with CMS Data Bookkeeping Service [116] in which one can find the results of the automatic certification as well. Once the final certifications are confirmed for different runs and the runs are "signed-off", a list of GOOD runs is prepared. Finally for physics analyses, the certification can even be more specific, resulting in qualified luminosity sections within runs. This information is saved in files with Java Script Object Notation or JSON files and can be used while running the analysis on the relevant dataset. Figure 2.26 shows a simplified picture of the online and offline DQM workflow.

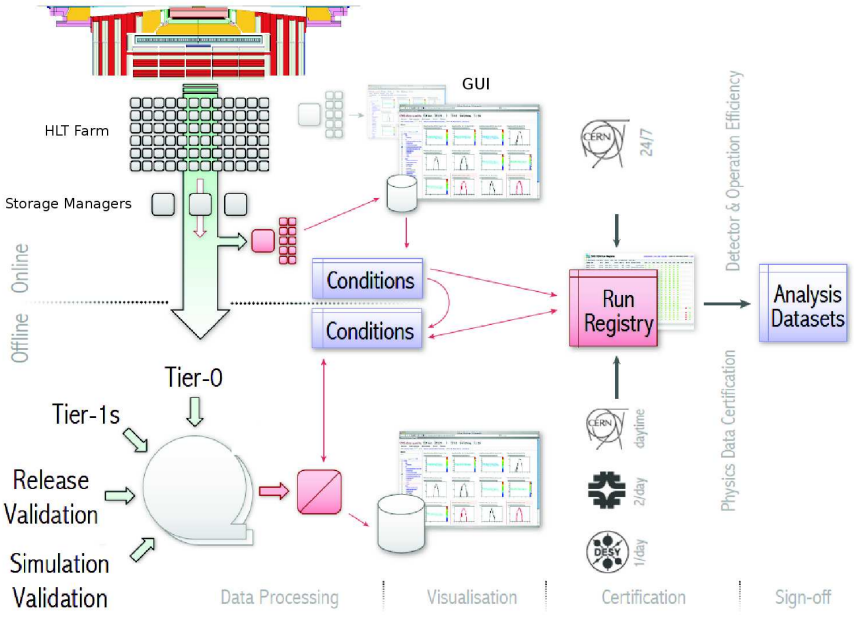


Figure 2.26: A simplified schematic view of the online and offline DQM workflow

Analysis oriented DQM: TopDQM example

In another view, DQM involves different levels from detector to more complex high level quantities. Monitoring the status and behavior of each subsystem up to the local reconstruction is done at the level of DPGs (stands for Detector Performance Group) in both online and offline scenarios. At the level of POGs (stands for Physics Object Group) the reconstructed physics objects like muon, electron, etc. go through qualification tests. Higher level quantities like kinematic distributions with more analysis oriented cuts, are monitored at PAG level (Physics Analysis Group). The POG and PAG monitoring are performed offline.
The motivations behind the DQM at PAG level are to spot possible changes/problems which are not seen at other levels of monitoring and to monitor those part of the

analysis which are sensitive to the changes in the detector conditions, calibration and alignment constants, etc. The quality monitoring in the top quark PAG (TopDQM) [117] is divided into three main categories regarding the $t\bar{t}$ final states, namely the semi-leptonic, di-leptonic and full-hadronic decay channels. The most interesting or sensitive distributions are defined as the reference plots in each category. For the semi-leptonic decay channel, one can point to the jet and lepton multiplicities, the kinematic distributions, the isolation and b-jet identification variables. The histograms are filled after the dedicated selections of the analysis which vary with different luminosity scenarios.

Since TopDQM is based on the CMS DQM framework, it can be integrated in the

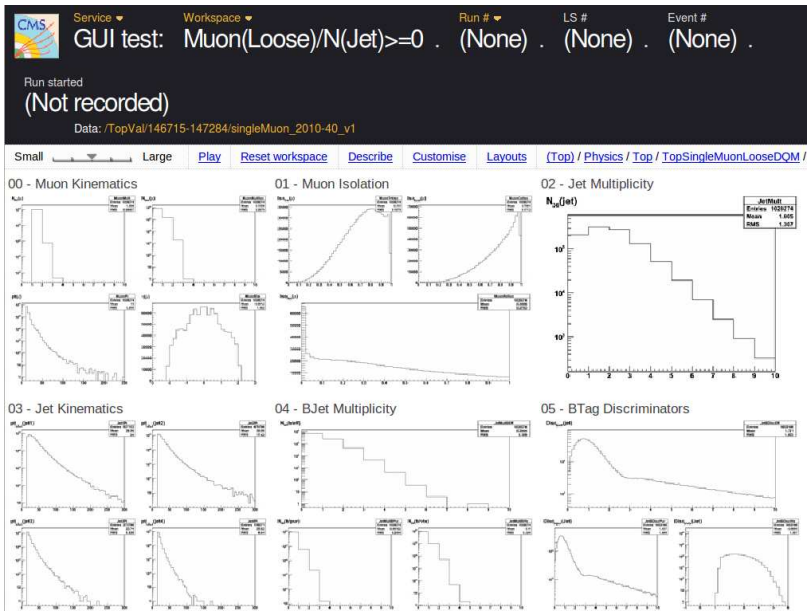


Figure 2.27: Part of CMS data from muon PD taken in October 2010, in TopDQM histograms of semi-muon selection [117]

central data quality monitoring operation. However because of the run-wise nature of central DQM, the central monitoring for TopDQM has been postponed for higher intensity collisions where in each run, more statistic and hence more top-like events are expected. During the 2010 data taking, weekly shifts were carried out so what entered the TopDQM plots was the data taken over the whole week. The files were stored on a private server and the histograms were visible via a local GUI. Figure 2.27 shows the distribution of some variables for the data taken in the 49'th week of 2010.

The same framework in the top-quark analysis group is used to validate the CMSSW releases as well as the new simulated samples or re-processed datasets. The only part of the validation code which is not integrated in the DQM framework since the information is not available in collision data, is the one investigating and confirming the correctness of the particles decay chain within the $t\bar{t}$ event and the branching ratios implemented in simulation. Figure 2.28 shows some of the distributions obtained from a subset of 2010 data [4] using the TopDQM program. The data is re-processed by two

[4] The subset is derived from the electron/photon PD. The events are stored if they fire a single or double lepton trigger.

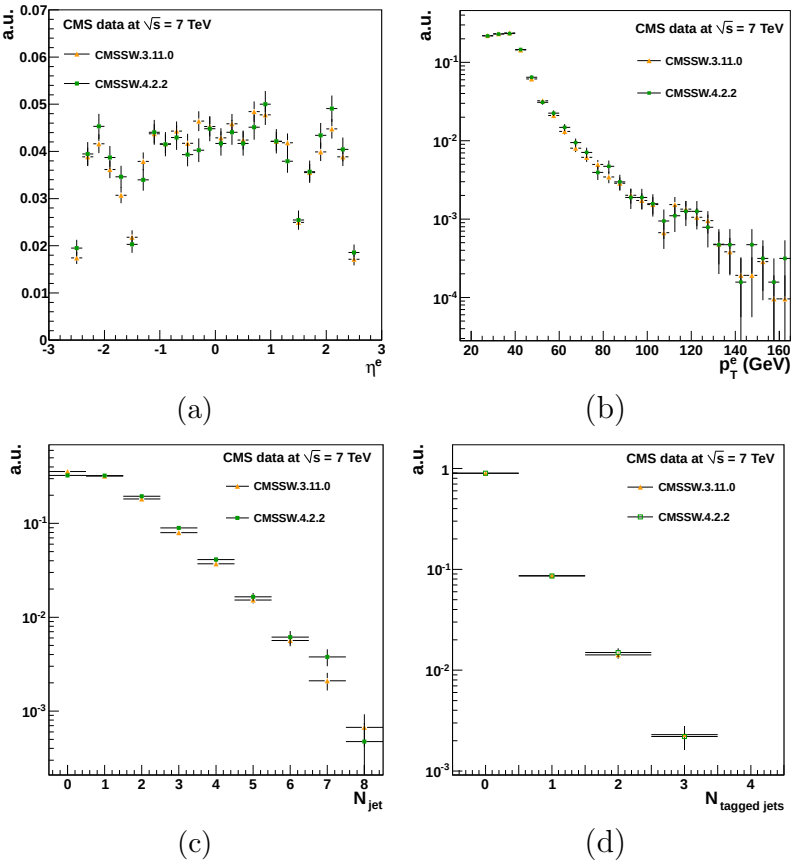


Figure 2.28: Comparison histograms for a subset of 2010 data, re-processed by two CMSSW releases. Distributions are made by events fullfilling the selection criteria for the semi-electron final state of $t\bar{t}$ which are implemented in the TopDQM program.

different CMSSW releases and passes through a selection sequence which is defined to select the top quark event candidates in the semi-electron final state for a medium luminosity scenario. This means that the events are asked to fullfil a single electron trigger requirement and to contain at least one electron candidate with $p_T > 25\,\text{GeV}$ and $|\eta| < 2.5$ (excluding the ECAL barrel-endcap transition region). The kinematics of the electron candidate is shown in histograms (a) and (b). Distribution (c) is the multiplicity of the jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$ within the selected events. The number of jets which are recgonized to be originated from a b-quark by means of a b-jet identification algorithm [5] are illustrated in Figure 2.28 (d). A good agreement is obsereved between the two CMSSW releases.
Such comparisons which give a confidence in new versions of the software can be quickly prepared in the DQM framework. Similar histograms are filled for the data monitoring purposes. The selection criteria are flexible and can be changed along with the data analysis.

[5] The b-jet identification algorithms are detailed in Section 4.4. The algorithm used in the preparation of Figure 2.28 (d) is known as Track Counting High Efficinecy algorithm.

Chapter 3

The simulation of collision events

The proton beams colliding in the Large Hadron Collider, feature interactions which are partialy known. The stable and long-lived unstable particles reaching the CMS detector are registered by the apparatus and provide the information needed for the physics analyses. The simulation of the whole procedure has been the subject of a vast scientific effort both before and after the operation of the machine.
The simulation gives a hint to the regions in the phase space in which the new phenomena can be observed. Therefore, the physics analysis strategies are developed relying on the description of physics processes and the way they can be seen by the detector. Even the discrepancies between the real collisions and the simulation are important since they can lead to a better modeling of physics interactions or probably a discovery. The simulation is also necessary for the design of the detector and triggers.
The current chapter starts with a general discussion about the proton-proton collisions in Section 3.1. Different parts of a generic event including the hard scattering, the non perturbative processes and the underlying interactions together with the way that they are modeled, are described in Section 3.2 where the emphasis is on the production and the decay of the heavy flavor partons.
Section 3.3 is devoted to the event generators used to simulate the top quark pair production in this thesis. In addition, the parameters used in the modeling are varied to study the effect of such variations on different physics observables. The discussion about extra interactions in the bunch crossing, the pile up events, is postponed to Section 3.4 where the simulation of the CMS detector is briefly reviewed. A summary of the simulated samples used for the analysis in this thesis is also given in the final section.

3.1 General features

The $pp \to XY$ process can in general be described in terms of a perturbative parton-parton interaction at high energies and of proton remnants carrying a large fraction of the total energy. This picture results in a nal state X from the hard scattering and an underlying event Y. While the study of the underlying events helps to delve into the

unexplored aspects of multi-parton interactions, many physics processes interesting for the Higgs particle and New Physics searches as well as for reinvestigating the Standard Model lie in the hard scattering.

A generic scattering process of two incoming protons with the four-momenta of P_1

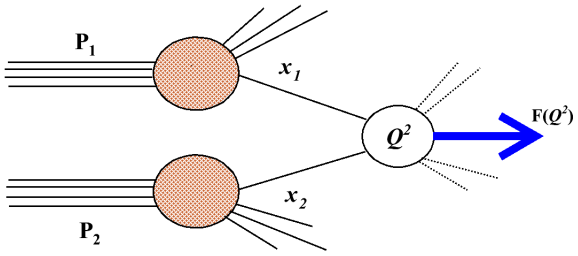


Figure 3.1: The simplified scattering process of two protons with the four momenta of P_1 and P_2 in the parton model.

and P_2 at hadron colliders, referred to as an event, is illustrated in Figure 3.1. In the scattering process the hard interaction happens between two partons, the proton constituents, carrying energy fraction x_i of the $i'th$ proton. The partons are assumed to be free according to the asymptotic freedom [118, 119] expected at the energy scale of the collision, $\sqrt{s} = 7\,\text{TeV}$. The center-of-mass energy of the proton-proton system, $s = (P_1 + P_2)^2$, is reduced to

$$Q^2 \equiv \hat{s} = x_1 x_2 s \tag{3.1}$$

for the energy regimes in which the partons are assumed to be massless. The variable Q^2 is defined as the momentum transfer or the virtuality of the process.
As illustrated in Figure 3.2, the proton-proton collision is more complex than the parton hard interactions:

- The partons can undergo radiation and showering before and after the inelastic scattering. The former is known as Initial State Radiation, ISR, where the latter is referred to as Final State Radiation, FSR.
- The final state radiations can be accompanied by the decay products of the short-lived resonances, like top quarks, produced in the hard interaction.
- At the end of the production chain, where the energy of the colored particles is low enough to break the perturbative QCD, the fragmentation of partons gives rise to the formation of the jets of hadrons.
- The proton remnants which are not contributing in the hard scattering will make the so-called underlying events.

Different event generators have been developed to simulate the complicated process of hadron collisions. While the general purpose event generators like `PYTHIA` [120] are able to simulate the complete process of the proton-proton interaction, the so-called "matrix element" generators such as `MadGraph/MadEvent` [121, 122] are dedicated to the computation of the hard scattering. In the general purpose event generators,

the request for the additional hard partons in the final state is resolved at the level of simulating the parton showers where a more accurate approach is to use the matrix elements, achieved in event generators like MadGraph.
The event generators which are mentioned here, compute the probability of the hard scattering at the lowest order of α_s. Although the additional partons increase the power of α_s in the final scattering probability, the corrections arising from the loop calculations are not included since the production of additional partons is also limited to the tree level calculations. The next-to-leading order approximation is approximated by scaling the LO calculation with the so-called *k-factor* which is the theoretical ratio between the NLO and LO cross sections. The k-factor values may change in the angular phase space, with the choice of the energy scale and the parton density functions which are introduced in Section 3.2.2, [123]. There are however event generators like MC@NLO [124, 125] that provide the next-to-leading order calculations to account for the higher order corrections in the hard scattering.
The event generators are all based on the Monte-Carlo techniques to reflect the stochastic character of the proton-proton collisions.

3.2 The factorization of hadron collisions

The quantum chromodynamic interactions are governed by α_s which can be small at higher energy scales, so the use of perturbative calculations is valid. The energy scale μ_R is the renormalization scale that is used to remove the ultraviolet divergences as discussed in Section 1.2.1. The virtuality of the process, Q^2, can be considered as the hard scale of the interaction for the interactions with a high momentum transfer. For the case of heavy flavor quark production, the hard scale can also be provided by the quark mass or its transverse momentum.

Although the ultraviolet divergences are regulated by the renormalization scale, singularities can still happen due to real gluon emissions. The gluons emitted in the direction of the outgoing parton lead to the so-called collinear divergences while soft divergences take place if a low momentum gluon is emitted. Such emissions, referred to as long-distance, result in terms in the perturbative expansion which are not small anymore and therefore they destroy the validity of the perturbation.
The infrared divergences can be absorbed by imposing a factorization scale, μ_F, on the perturbative expansion in the context of the "factorization theorem" [127]. Using the factorization scale, the short-distance physics that covers the hard process calculable in the perturbative QCD, is separated (factorized) from the non-perurbative long-distance interactions.
According to the factorization theorem and given the partons i and j with energies of $x_i P_1$ and $x_j P_2$, the differential cross section of a $i + j \to f$ process can be written as

$$d\sigma_{pp\to f} = \sum_{i,j} \int_0^1 dx_i \int_0^1 dx_j \, f_1^p(x_i, \mu_F^2) \, f_2^p(x_j, \mu_F^2) \, d\sigma_{i+j\to f}(\alpha_s(\mu_R), Q^2; \mu_F^2). \qquad (3.2)$$

where ps stand for the colliding protons with the four momenta of P_1 and P_2. The equation consists of two main terms, $f_{1(2)}^p(x_i, Q^2)$ and $d\sigma_{i+j\to f}$ which are explained in

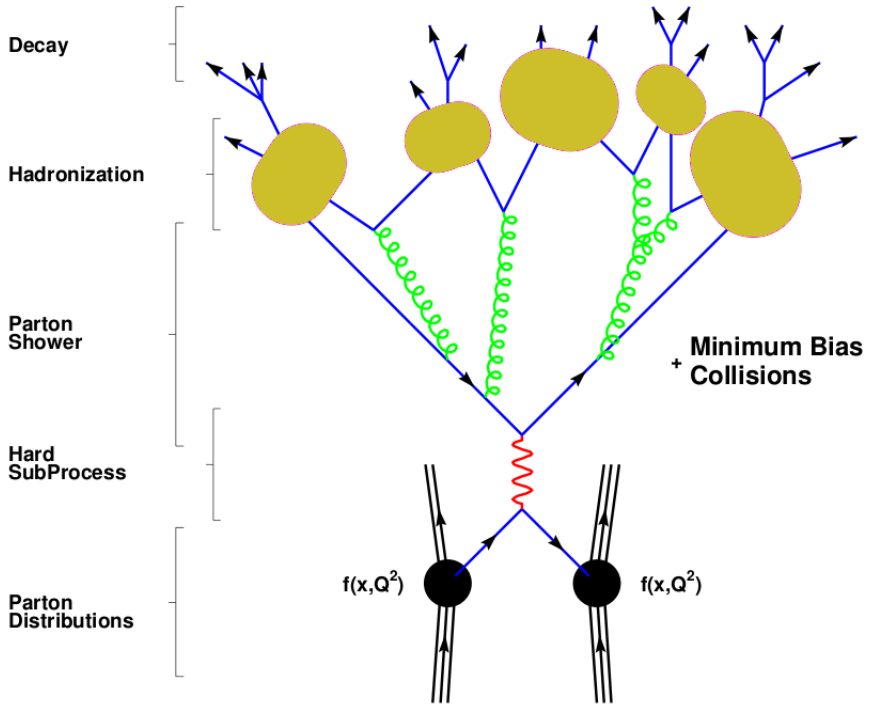


Figure 3.2: The schematic view of the generic structure of a proton-proton scattering including the hard interaction, showering and hadronization and underlying events [126].

the following subsections. The former contains the non-perturbative part of the process where the latter is calculated by perturbative QCD.

3.2.1 The partonic hard scattering

The $d\sigma_{i+j\to f}$ term in Equation 3.2 is the differential cross section of $i+j \to f$ process, indicating the hard interaction in the pp collisions. Figure 3.3 shows and example of the partonic hard scattering resulting in the production of $t\bar{t}$. It can be seen that the partons i and j can either be two gluons or a quark and an anti-quark. The summation over i and j indices in Equation 3.2 gives the overall differential cross section from different production modes where the total cross section is obtained by the calculation of matrix elements and integration over the kinematic phase space. This integration is not indicated in Equation 3.2. Most of the event generators calculate the

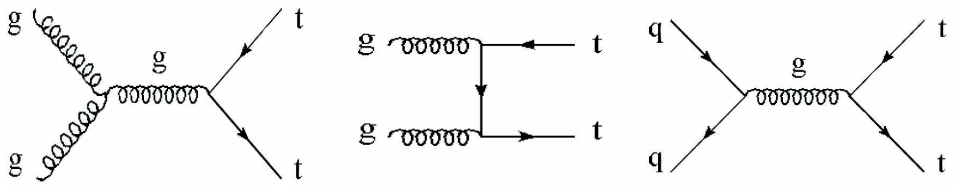


Figure 3.3: The leading order Feynman diagrams for the production of heavy quarks like $t\bar{t}$ in hadron collisions.

hard scatterings at leading order. For the case of heavy flavor productions like $t\bar{t}$, no singularity happens since for the s-channels (diagrams (a) and (c) in Figure 3.3) the energy of the propagator, the gluon, has to be larger than $2m_t$. For the t-channel gluon exchange ((b) in Figure 3.3) it can be shown [128] that the virtuality of the process has to be larger than m_t^2. This sets the scale for α_s and since it is much larger than Λ_{QCD} (introduced in Section 1.2.1), the perturbative QCD is valid for the calculation. Another feature of the top quark production is the suppression of the quark annihilation mode at very high center of mass energies:

$$\sigma(q\bar{q} \to t\bar{t}) \to \frac{1}{\hat{s}} \tag{3.3}$$

and

$$\sigma(gg \to t\bar{t}) \to \frac{1}{\hat{s}}\left(\frac{1}{\beta} log\left(\frac{1+\beta}{1-\beta}\right) - 2\right), \tag{3.4}$$

where $\beta \equiv \sqrt{1 - \frac{4m_t^2}{\hat{s}}}$ is the relativistic velocity of the top quark. The $t\bar{t}$ production at the LHC is dominated by gluon fusion since at high energies the quark annihilation is suppressed more quickly.

3.2.2 The parton density functions

The partonic hard interaction can happen between any two partons allowed by QCD where the desired partons can be extracted from the proton with a certain probability. Therefore, the hard scattering term is corrected by the Parton Distribution Functions, $f_i^p(x_i, Q^2)$, for each of the contributing partons. The parton density function as stated in Equation 3.2 is the creation probability of parton i with fraction x_i of a proton total energy, computed at the virtuality of $Q^2 = \mu_F^2$.

The parton distribution functions are obtained by global fits on the experimental

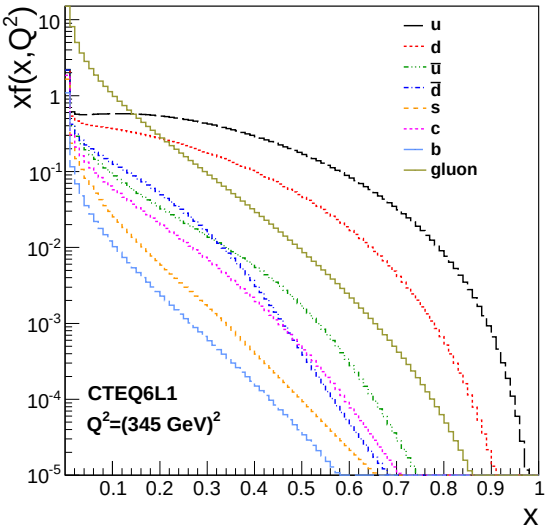


Figure 3.4: Different parton density functions at $Q^2 = (2m_t)^2 = (345\,\text{GeV})^2$, according to the CTEQ6L1 release [129].

results at different virtualities. Different groups like CTEQ [129] perform the global fit evaluation for the parton distribution functions with new data and the theoretical predictions. The parton distribution functions from the CTEQ6L1 [130] measurement are plotted for different partons in Figure 3.4. The same distributions are used to generate the simulated samples used in this thesis.

The evolution of PDFs

A parton distribution function which is evaluated at scale Q^2, can be rescaled to any other scale Q'^2 as far as the α_s remains small at this new scale so the perturbation is retained. Hence the PDFs evaluated at lower energies can be rescaled to meet the energy of the LHC. The evolution of the parton distribution functions to other energies is governed by the DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi) equations [131–133],

$$Q^2\frac{\partial}{\partial Q^2}f_q = \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dy}{y}\left[f_q(y,Q^2)\mathcal{P}_{qq}\left(\frac{x}{y}\right) + f_g(y,Q^2)\mathcal{P}_{qg}\left(\frac{x}{y}\right)\right] \tag{3.5}$$

and

$$Q^2\frac{\partial}{\partial Q^2}f_g = \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dy}{y}\left[\sum_{i=1}^{2n_f} f_{q^i}(y,Q^2)\mathcal{P}_{gq}\left(\frac{x}{y}\right) + f_g(y,Q^2)\mathcal{P}_{gg}\left(\frac{x}{y}\right)\right]. \tag{3.6}$$

The parton creation probability at other energies can changes due to different processes. The number of quark q with a given flavor can change if a gluon splits into $q\bar{q}$. Quarks can also be found at lower momenta by emitting gluons. This is stated in Equation 3.5 by $\mathcal{P}_{qq}$ and $\mathcal{P}_{qg}$ which are known as splitting functions.
The number of gluons is enhanced if a gluon is created by radiation from quarks, $\mathcal{P}_{gq}$, or a gluon splits to two gluons, $\mathcal{P}_{gg}$. As stated in Equation 3.6 a summation over quark and anti-quarks from all flavors, is needed for the quark radiations.

3.2.3 The parton showers

Successive splitting processes occur before and after the hard scattering and result in showers of partons. The showering continues until the energy of the partons reaches values below Λ_{QCD} for which the perturbative showering approach is not valid.
Although the accurate description of parton showering at leading order can be provided by the matrix element event generators, the singularities arising from the soft and collinear gluons are hardly regulated. Hence the perturbative approach based on the DGLAP equations is implemented in the parton showering programs. While the final showers (FSR) are started from an upper scale Q^2_{max} which typically is the scale of the interaction, the initial showers (ISR) are simulated in reverse. They are start from the scale of interaction and backpropagated to the scale at which the initial parton was extracted from the proton.
The `PYTHIA` parton shower algorithm is used to describe the showering content

of the events simulated for the analysis in this thesis. A set of parameters control the showering procedure and therefore the amount of ISR/FSR in PYTHIA. The simulated samples useful to study the systematic effects (see Section 3.3.1) of radiations on the physics observables are provided by different tunes for these parameters.

Heavy flavor production in parton showers

The hard scattering process explained in Section 3.2.1 is generic for the heavy quark production. The charm and bottom quark can be produced in addition via processes called gluon splitting and flavor excitation, shown in Figure 3.5. This gives a next-to-leading order contribution to the inclusive heavy quark $Q\bar{Q}$ production. In the splitting

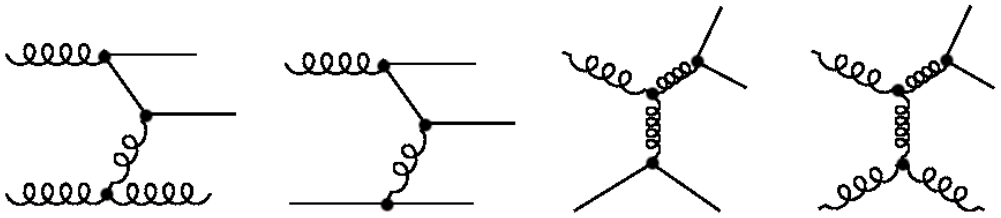


Figure 3.5: The next-to-leading order diagrams for heavy quark production. From left to right, the first two diagrams relates to flavor excitation process while the last two represent gluon splitting.

process, which happens in the final state showering, a radiated gluon with a virtuality of $Q^2 > 4m_Q^2$ is split into the $Q\bar{Q}$ pair. Heavy flavor quarks produced in gluon splitting can carry a large transverse momentum. They therefor are very close and often end up in the same jet. The flavor excitation corresponds to the splitting of an initial state gluon to a $Q\bar{Q}$ pair of which a quark undergoes hard scattering. The other one which is part of the proton remnants is often outside the acceptance region. Figure 3.6 shows different contributions to the total cross section of b-quark production as a function of the center-of-mass energy in proton-proton collisions.

The matching between parton shower and matrix element

The parton showers produced by PYTHIA are added to the final state of the hard scattering computed by the matrix element generators to give a more complete picture of the proton-proton collision. For the simulated samples used in this thesis the MadGraph/MadEvent program is interfaced with the PYTHIA parton showering. A double counting issue arises since the partons produced in showering steps can also be obtained in the matrix element calculation. This can be resolved by matching the partons in the final states of two generators following different schemes like MLM [135, 136]. In the MLM approach which is used for the samples in this thesis, the matrix element generated partons within the pseudo-rapidity acceptance are asked to have a p_T greater than some threshold. They are also required to have a certain separation in the η-ϕ plane.

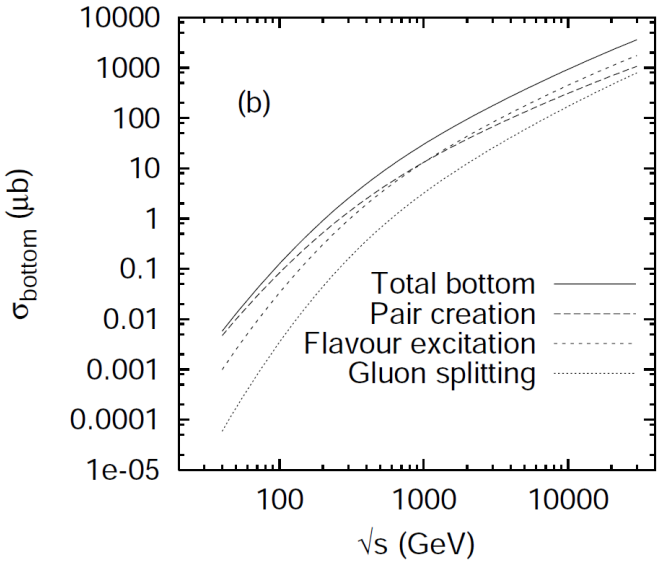


Figure 3.6: Total cross section of b quark production together with the contributions from pair production, flavor excitation and gluon splitting. The cross sections increase as a function of the center-of-mass energy, $\sqrt{s}$ [134].

The selected partons undergo showering via the PYTHIA showering algorithm and are clustered in the jets afterward. Jets with a transverse energy greater than E_T^{min} are taken to match with the generated partons. A jet-parton pair is considered as "matched" if the separation between the two is not larger than some threshold which normally is the typical size of the jets. Removing the pair from the list, the matching is performed successively until no parton remains. The event is rejected if a parton remains with no associated jet. To produce the inclusive samples like $t\bar{t}$+jets, the MLM scheme is applied in parton multiplicities where extra jets are allowed at the highest generated parton multiplicity.

For the next-to-leading order event generators, the situation is more complicated. Extra partons are generated at the higher order of perturbation. Therefore, the association of the jet of showered parton to the generated parton becomes ambiguous. Such ambiguities are avoided by the approach implemented in the MC@NLO event generator.

For a given n-parton final state, computed at NLO, the "real" parton corrections result in an (n+1)-parton event. The virtual corrections (loop calculations) do not increase the number of partons. The (n+1) parton multiplicity has in fact contributions from the NLO correction of the n-parton bin and from the parton showering of n-parton events. This is where the double counting can happen. The showering of the n-parton event on the other hand is investigated analytically to estimate how the showered final state of an n-parton event would populate the phase space of the (n+1)-parton topology. Hence, the contribution of the showering n-parton events can be subtracted from the (n+1)-parton final states.

3.2.4 The hadronization in the final state

Sequentially radiating, the partons enter the energy regime for which the perturbative QCD formalism is not reliable. This happens due to the extremely large coupling constant, α_s, at low energies which enforces the partons to be confined in color-singlet combinations. The fragmentation of colored partons into colorless hadrons is therefore simulated using phenomenological models such as the Lund string model [137].
By means of such phenomenological models, the long-distance physics of the hadronization process is incorporated in the "fragmentation functions" which are independent from the hard scattering. It means that the fragmentation functions are universal and the models tuned for e^+e^- and ep collision data are applicable on the LHC data.
Qualitatively speaking, the Lund model assumes a color flux tube with an increasing potential versus distance that is formed between two partons in a $q\bar{q}$ pair once they tend to move apart. The color-singlets of $q\bar{q}'$ and $\bar{q}q'$ are created if the color tube (the so-called string) breaks. The formation of the new pairs is explained by the concept of quantum tunneling. Considering the string stretched in the longitudinal direction and having no transverse excitation, the transverse momentum is divided between quarks. The probability of tunneling is controlled by the mass and the transverse momentum of the generated quarks. Qualitatively speaking, the creation of quarks with higher masses or momenta is less probable,

$$\mathcal{P} \propto \exp\left(-\frac{\pi m^2}{\kappa}\right) \exp\left(-\frac{\pi p_T^2}{\kappa}\right). \tag{3.7}$$

The break down of strings and the formation of color singlets continue until the invariant mass of the $q\bar{q}$ system is not high enough to support further fragmentations.
The fragmentation function, $f(z)$, proposed by the Lund model based on the tunneling assumption is

$$f(z) \propto \frac{(1-z)^{a_l}}{z} \exp\left(-\frac{b_l(m_h^2 + p_{T,h}^2)}{z}\right) \tag{3.8}$$

which is valid for the u, d and s quarks. The variable z is the fractional momentum of hadron $q\bar{q}'$ that is split of from the string and leaves the $(1-z)$ momentum fraction for the remainder of the string. The quantities m_h and $p_{T,h}$ are respectively the mass and the transverse momentum of the created hadron. The parameters a_l and b_l are tuned by fitting to the experimental observations.
For the hadronization of heavy quarks, the experimental data are very well described by the Peterson [138] fragmentation function,

$$f(z) \propto \frac{1}{z}\left(1 - \frac{1}{z} - \frac{\epsilon_q}{1-z}\right)^{-2}. \tag{3.9}$$

The parameter ϵ_q is measured in the experiment and expected to behave as $\epsilon_q \sim 1/m_q$. For the simulated sample used in this thesis, the hadronization process implemented in PYTHIA is used. The default values of $a_l = 0.3$ and $b_l = 0.58\,(\mathrm{c/GeV})^2$ are taken for non-heavy quarks while $\epsilon_c = -0.05$ and $\epsilon_b = -0.005$ are set for the c and b quarks, respectively.

The decay of heavy flavored hadrons

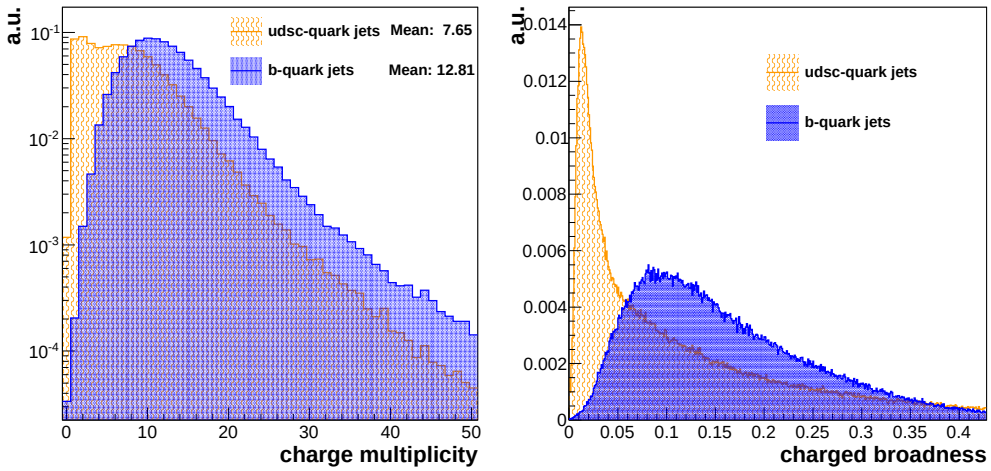
The jets are made along with the hadronization and they contain the decay products of unstable hadrons. The interaction between the decay products of long-lived unstable hadrons, such as B mesons and the detector material provides special signatures by which the flavor of the jets can be identified. The b-flavor identification of the jets is of particular physics interest and is the basis for the analysis developed in this thesis. Most of the b-hadrons are composed of a b-quark and a light quark, namely u,d,s. Explained by the spectator model [139], the b-hadrons decay via the weak interaction of b-quark, $b \to c(u)W^*$, while the other quark has the role of spectator. The decay is highly dominated by $b \to cW^*$ due to the "relatively" large CKM [1] element, $(|V_{ub}|/|V_{cb}|)^2 < 0.01$ [3]. The W^* decays to a lepton and a neutrino in about 10% of the time per lepton flavor where for each flavor the leptonic final state can be enhanced by $W^* \to cX \to l\nu_l X$ with an extra 10%.
The small value of $|V_{cb}| = 0.0412 \pm 0.0011$ introduces a relatively long life time of $\tau \sim 10^{-12}$ s[3] for b-hadrons which corresponds to an average decay lenght of $c\tau \approx 450\,\mu m$. Considering the boost factor,

$$\text{flight path} = \beta\gamma c\tau = \frac{p_B}{m_B} c\tau, \tag{3.10}$$

the flight path in the LHC rest frame is about 3-5 mm. The quantities p_B and m_B are the momentum and the mass of the b-hadron.
This can be observed as a displaced vertex in the CMS detector together with tracks with large impact parameter. Because of the b-quark mass, the decay products carry a large transverse momentum with respect to the jet direction and hence can be identified. The difference between the b- and non-b quark jets can be expressed in terms of the physics observables related to the jet[2] properties. Regarding the interaction of charged particles with the tracker material, the number of tracks associated to a jet reflects the jet charge multiplicity. Figure 3.7 (a) illustrates the charge multiplicity in b-flavored jets compared to other quark jets. The number of tracks is on average 5 more in the jets originated from b-quarks. Another interesting quantity is the so-called "charged broadness" of the jet. It is the radius of a cone around the jet axis which contains 75% of the total charged energy in the jet. The charged energy is the jet energy fraction that belongs to the charged particles and the jet axis specifies the jet direction. Charged particles with higher transverse momentum result in a larger jet charged broadness. Figure 3.7 (b) compares the charged broadness of the b-quark jets to the jets from other quarks. It can be seen that for non-b-quark jets, the charged particles are mostly collimated around the jet axis. The jets are selected from the final state of the $t\bar{t} \to q'qb\bar{b}e\nu_e$ process, asked $p_T > 30$ GeV and $|\eta| < 2.4$. The algorithms to identify the b-flavored jets are explained in Chapter 4.
The event generators need to be equipped with programs to describe the decay of hadrons as well as the properties of stable particles produced in the hard interaction, e.g. the leptons from the decay of W boson. The PYTHIA event generator gives a

[1] See Section 1.1.3 for definition.

[2] Jets are reconstructed with Anti-κ_T algorithm with the recombination parameter of 0.5, as explained in Section 4.3



(a) (b)

Figure 3.7: The charge multiplicity, (a), and the jet "charged broadness", (b), of the jets in the semi-electron final state of $t\bar{t}$.

proper description for the decay of particles with $c\tau < 10\,\text{mm}$. It is interfaced with the TAUOLA package to include the spin polarizations in the decay of τ-leptons.

3.2.5 Underlying events

The event generation explained so far, concerns the hard scattering, the X part in the generic interaction of $pp \to XY$. The simulation of underlying events, part Y, has to be taken into account as well to give a realistic insight into the proton-proton collisions. Largely independent of the hard interaction, underlying events consist of the colored remnants of the protons undergone the hard interaction. These colored particles are eventually hadronized and increase the particle multiplicity. Underlying events also include the possible hard or semi-hard interactions between those partons from each proton that have not contributed in the main interaction. These secondary interactions have in general lower transverse momentum and populate mostly the forward regions of the detector. They therefore lead to correlations between the energy flow in the central and forward regions.
To simulate the first category of underlying events, known as "beam remnant", a primordial transverse momentum $\kappa_\perp$ is ascribed by PYTHIA to the partons of the main interaction and its recoil is carried by the remnants of proton. The distribution of $\kappa_\perp$ is assumed to be Gaussian with width of σ_κ that can be tuned with data. The D6T [140] tunes are obtained from the Tevatron data and are implemented to produce part of the samples used in this thesis. The tunes have been updated with the LHC data at $\sqrt{s} = 900\,\text{GeV}$ and $\sqrt{s} = 7\,\text{TeV}$. The "CMS UE Tune Z1" [141] has been introduced to describe the underlying events observed by the CMS experiment. Changing the PDF parameters from CTEQ5L to CTEQ6L, the so-called "Z2 Tunes" [142] are obtained which are used in the recent production of simulated samples in CMS. Table 3.1 sum-

marizes the values of $\kappa_\perp$ and $\sigma_{\kappa_\perp}$ together with the upper bound value of $\kappa_\perp^{max}$ for D6T and Z2 tunes.
For the second category of underlying events, the multiple interactions, a conceptual definition is considered in PYTHIA for the size of the overlap between the two colliding hadrons. This has been quantified by approximating the density of the hadronic matter with a double Gaussian distribution,

$$\rho(r) \propto \frac{1-\beta}{a_1^3}\exp\left\{-\frac{r^2}{a_1^2}\right\} + \frac{\beta}{a_2^3}\exp\left\{-\frac{r^2}{a_2^2}\right\}, \tag{3.11}$$

introducing parameters a_2/a_1 and β for tuning. The core region, containing the β fraction of total hadron matter is assumed to be centered in an sphere with radius a_1 and it is surrounded by the rest, up to the radius a_2. The transverse momentum lower threshold for the interaction to happen, $p_\perp^{min}$, together with the cut-off $p_\perp^0$ to regularize the divergences at $p_\perp \to 0$ are tuned at the reference energy $\sqrt{s}^{ref}$. The energy scaling is considered as $s^{-\frac{x_{sc}}{s}}$. The summary of multi-interaction parameters for D6T and Z2 tunes can be found in Table 3.1.

	$\kappa_\perp^{max}$	$\sigma_{\kappa_\perp}$	β	a_2/a_1	$p_\perp^{min}$	$p_\perp^0$	$\sqrt{s}^{ref}$	x_{sc}
D6T	15 *	2.1 *	0.5	0.4	1.9 *	1.8387 *	1960 *	0.16
Z2	10 *	1 *	0.356	0.651	1.9 *	1.832 *	1800 *	0.275

Table 3.1: The parameters to model the underlying events tuned for the Tevatron (D6T) and the CMS experiment at the LHC (Z2). The notation * stands for the GeV energy/momentum unit.

3.3 Top quark production with different generators

Although all event generators are aiming to predict the same physics expected from the theory or observed in the experiment, they may differ in the kinematics of the event due to dissimilar algorithms implemented to generate the events and different matching schemes. The simulated samples used in this thesis are generated with MadGraph (see Section 3.4) for which the matrix elements are calculated at the leading order. Therefore, it is interesting to see the effect of the next-to-leading order corrections, implemented in the MC@NLO event generator. The next-to-leading order corrections are categorized into the real emissions, described in Section 3.2.3, and the virtual emissions corresponding to the loop calculations. A useful discussion about the effect of the next-to-leading order corrections can be found in [128] while one can qualitatively point to the changes expected either in the $\hat{s}$ dependence of the cross section or in the kinematic distributions, due to new processes that appear at next-to-leading order.
To compare these two generators, $t\bar{t}$ events with the semi-electron final state, $t\bar{t} \to q'qb\bar{b}e\nu_e$, are considered. Figure 3.8 illustrates some kinematic variables obtained from the $t\bar{t}$ samples produced by MadGraph ($m_t = 172.5$) and MC@NLO ($m_t = 170.9$). The distributions show differences reflecting the inequal m_t input together with the different

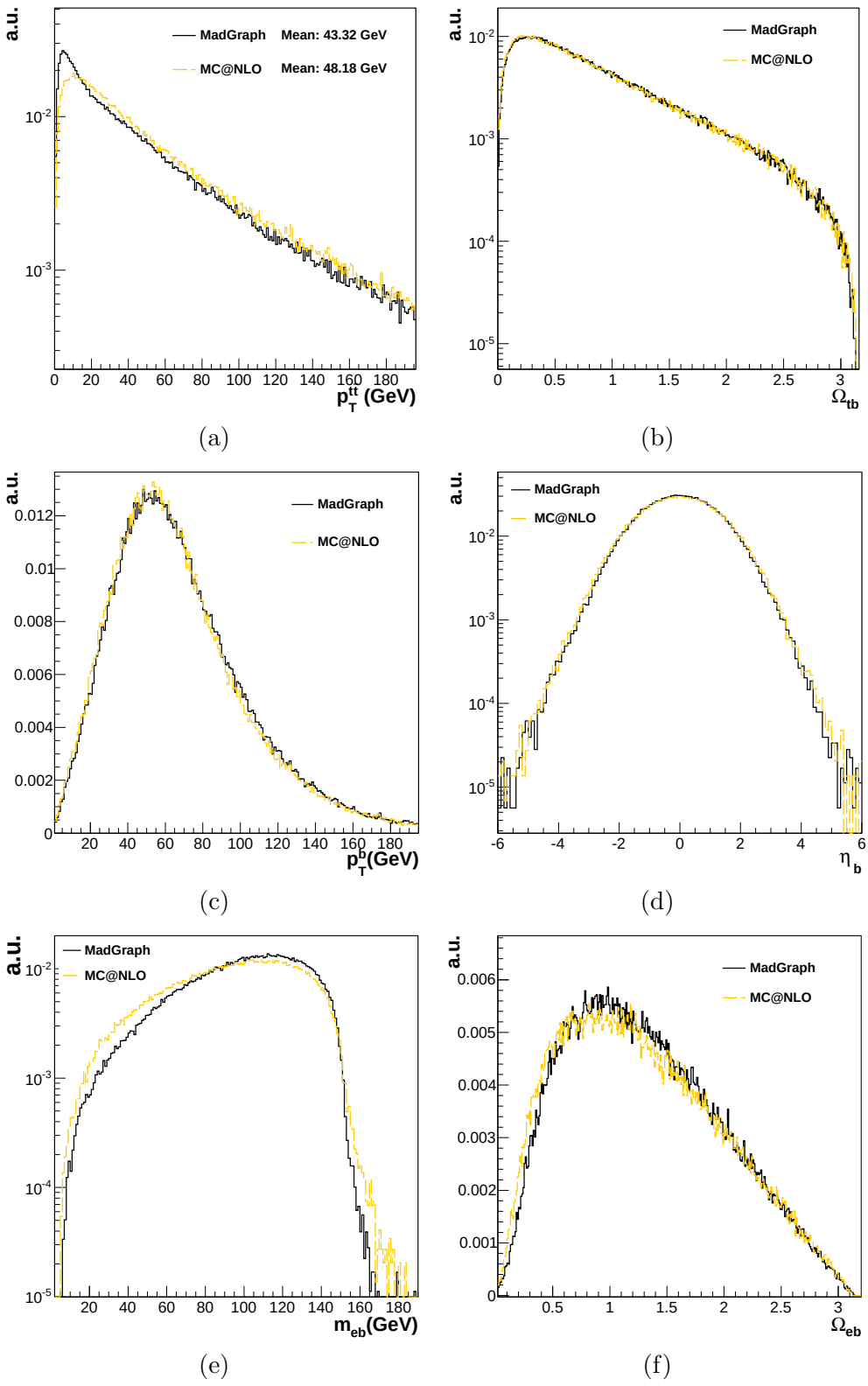


Figure 3.8: Some kinematic distributions in the $t\bar{t} \to q'qb\bar{b}e\nu_e$ process for two different event generators. The p_T of the $t\bar{t}$ system (a), the space angle between the top and the b-quark on the leptonic side (b), the transverse momentum (c) and pseudorapidity (d) of the b-quark, the invariant mass of the electron and the b-quark (e), together with their space angle (f) are illustrated.

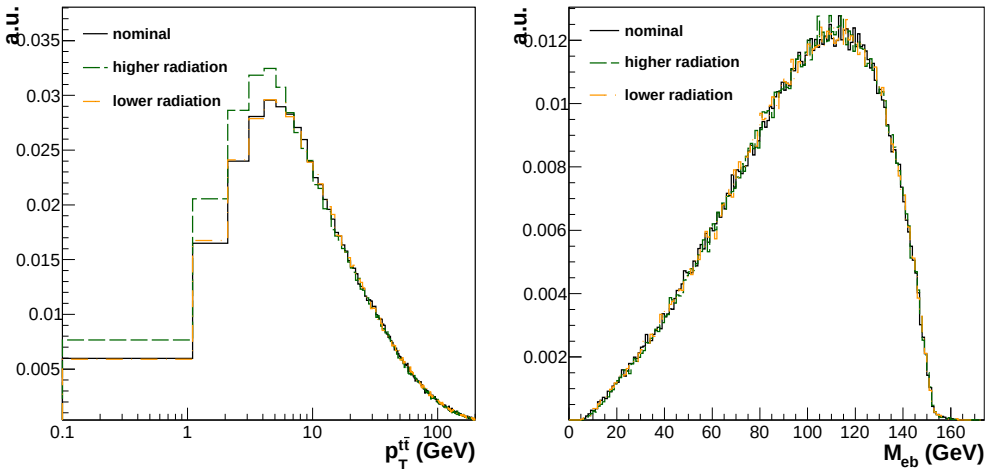
orders of precision in calculations. The p_T of the $t\bar{t}$ system simulated by MC@NLO is on average $\sim 5\,\text{GeV}$ higher than the one of MadGraph. The space angle which is calculated in the $t \to bW \to be\nu_e$ process between the top and the b quark is in good agreement between the two generators. The shapes for the η and the p_T of the b-quark are quiet similar except the fact that in MC@NLO, the p_T distribution is slightly tending to lower values which is expected due to the lower top quark mass.
The kinematic correlations between the b-quark and the electron are interesting to study since they provide the main ingredients for the method developed in this thesis (see Appendix B). The space angle between the electron and the b-quark is a bit smaller for the MC@NLO generator. This can be deduced from the larger boost in the top quark produced by MC@NLO which makes the decay products more collimated.
The key variable of the analysis, the invariant mass of the electron and the b-quark, is broader in MC@NLO. The systematic uncertainty arising for such differences is studied in Section 5.5.5.

3.3.1 Parameter variation for systematic uncertainties

The MadGraph event generator is interfaced with PYTHIA for the proper description of the parton showers. As explained in Section 3.2.3, the showering in PYTHIA follows the DGLAP equation and is governed by parameter tuning. The variation of tunes introduces the systematic uncertainty on the physics estimators. Therefore, samples with different tunes are simulated to study these systematic effects.

Radiation

The scale Q^2 in the DGLAP equations is meaningful when it is compared to a reference scale, i.e. Λ_{QCD}. Parameter PARP(61) governs the amount of ISR while tuning of parameter PARP(72) influences the FSR content of the events. Parameter PARP(81) is more general, controlling the amount of final state showering in the decay of resonances. The values of these parameters changed from 0.25 GeV (default value) to 0.35 GeV in order to increase the amount of initial and final state radiations via the DGLAP equations. Moreover for the ISR evolution, parameter PARP(64) is multiplied to the scale at which α_s is calculated. This parameter is increased from 0.2 in the nominal sample to 1 in the sample with increased radiation content.
For the samples with less radiation, another approach is used. While by default the maximum virtuality of emissions is set to the center-of-mass energy so the shower is allowed to populate the full phase space, for the less radiation it is cut at μ_F. The change of the MSTP(68) parameter from 3 (default) to 1 is dedicated to decrease the radiation content.
It can be seen in Figure 3.9 (a) that the $p_T^{t\bar{t}}$ in the semi-electron final state has slightly decreased for larger amount of emissions since the energy of the system is partly carried by the radiated partons. For the smaller amount of emission, no significant difference is seen. This can be the consequence of the parametrization for lower radiation scenario. As illustrated in Figure 3.9 (b), the electron-b-quark invariant mass, m_{eb} is quiet similar for the three different definitions of the radiation content.



(a) (b)

Figure 3.9: The transverse momentum of the $t\bar{t}$ system, (a), and the invariant mass of the b-quark and the electron, (b), in the semi-electron final state of $t\bar{t}$ (radiation scenarios).

Scale

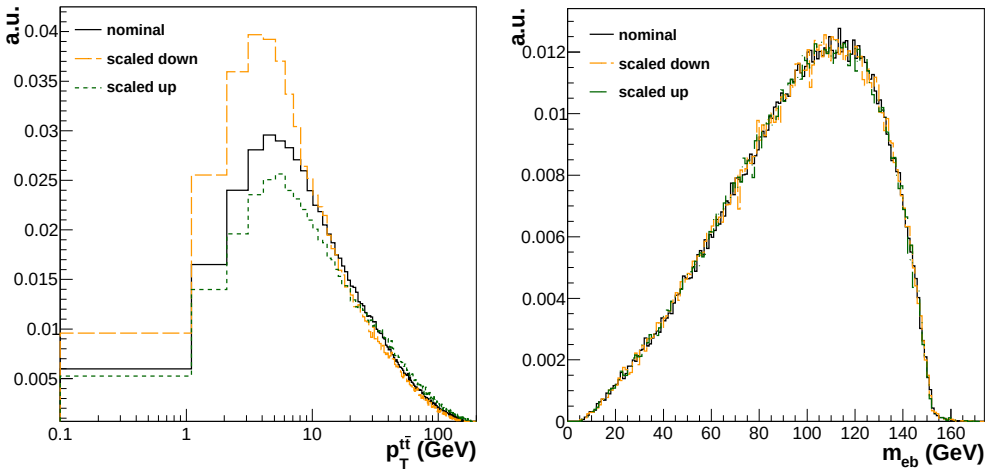
The factorization scale used in the matrix element calculation in `MadGraph` changes event-by-event and is defined as

$$Q^2 = m_t^2 + \sum p_T^2(jets), \tag{3.12}$$

where the summation runs over all generated partons in the hard scattering. Since the topology of the final event depends on the factorization scale, dedicated samples with scaled Q^2 by half and twice are simulated to study the related systematic effects. Figure 3.10 shows the $p_T^{t\bar{t}}$ and the m_{eb} quantities for the three scaling scenarios. While the m_{eb} distribution remains unchanged, in the $p_T^{t\bar{t}}$ lower Q^2 results in smaller values for the transverse momentum of the $t\bar{t}$ system.

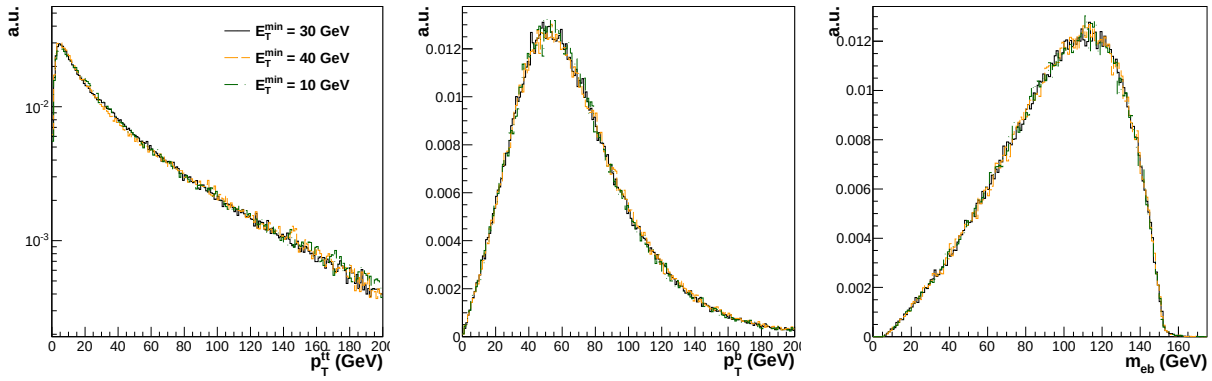
Matching

Another source of systematic effect is induced by the matrix element and parton shower matching threshold, E_T^{min}, introduce in Section 3.2.3. The threshold is taken 30 GeV for the nominal production while it is varied to 10 GeV (matching down) and 40 GeV (matching up) for the systematic studies. As illustrated in Figure 3.11, the matching threshold does not have a significant influence on $p_T^{t\bar{t}}$ and the b-quark transverse momentum. The m_{eb} distributions are also quiet similar. The systematic uncertainties due to the amount of ISR/FSR, Q^2 scaling and ME-PS matching are investigated under the subject of model dependent uncertainties in Section 5.5.4.



(a) (b)

Figure 3.10: The transverse momentum of the $t\bar{t}$ system, (a), and the invariant mass of the b-quark and the electron, (b), in the semi-electron final state of $t\bar{t}$ (scaling scenarios).



(a) (b) (c)

Figure 3.11: The transverse momentum of the $t\bar{t}$ system, (a), the p_T of the b-quark, (b), and the invariant mass of the b-quark and the electron, (c), in the semi-electron final state of $t\bar{t}$ (matching scenarios).

3.3.2 Cross section of $t\bar{t}$ production

The theoretical cross section of $t\bar{t}$ production can be calculated analytically up to next-to-leading order precision [143]. The soft gluon corrections are then added to $q\bar{q}$ and gg processes to obtain the next-to-next-to-leading order approximation.
For pp collisions at the LHC at 7 TeV center of mass energy, the $t\bar{t}$ cross section at next-to-leading order is evaluated by MCFM program [144] using the CTEQ PDF set [145],

$$\sigma^{NLO}_{pp\to t\bar{t}}(7\,\mathrm{TeV}, m_t = 172.5\,\mathrm{GeV}, \mathrm{CTEQ}) = 157.5^{+14.7+18}_{-14.7-19.5}\mathrm{pb} = 157.5^{+23.2}_{-24.4}\,\mathrm{pb}. \quad (3.13)$$

The first term in the uncertainty arises from the uncertainty on PDF and α_s, determined by following the procedures from other PDF sets, namely MSTW2008 [146] and NNPDF2.0 [147]. The second term is due to the variation of the normalization and factorization scales. While a common choice for μ_F and μ_R is the mass of top quark, m_t, a variation in the $0.5 \leq \mu/m_t \leq 2$ range is normally performed to account for the scale uncertainty. For the special value quoted in Equation 3.13, two extremes, $\mu/m_t = 0.5$ and $\mu/m_t = 2$ are checked.
An extensive effort has been made to approximate the NNLO $t\bar{t}$ cross section at $\sqrt{s} = 7$ TeV, recently has lead to e.g. [148]

$$\sigma^{NNLOapprox}_{pp\to t\bar{t}}(7\,\mathrm{TeV}, m_t = 173\,\mathrm{GeV}, \mathrm{MSTW2008}) = 163^{+7+9}_{-5-9}\mathrm{pb} = 163^{+11}_{-10}\,\mathrm{pb}, \quad (3.14)$$

or with a different approach to[149]

$$\sigma^{NNLOapprox}_{pp\to t\bar{t}}(7\,\mathrm{TeV}, m_t = 173.1\,\mathrm{GeV}, \mathrm{MSTW2008}) = 149^{+7+8}_{-7-8}\mathrm{pb} = 149 \pm 10\,\mathrm{pb}, \quad (3.15)$$

for which the first error results from scale variations and the second reects PDF uncertainties [3]. Although the NNLO calculation has a significant contribution to the NLO $t\bar{t}$ cross section, it considerably reduces the scale dependence.

3.4 The simulation of the CMS detector

The event generators provide a complete picture of the proton-proton collisions and the stable particles in the final state. In reality however, the products of the scattering process cannot be studied unless they interact with the detector material. Such interactions result in electronic signals in the detector readout which are further digitized to be used in the subsequent reconstruction algorithms [4].
Therefore a precise simulation of the detector response is necessary to describe the observation of the physics objects in real data. The full detector simulation with the desirable precision is based on the GEANT4 [150] toolkit. This involves simulating the geometry of the detector and the description of the material used to detect the traversing particles as well as the details about the inactive components such as supports, cooling system, etc.

[3] A detailed discussion about the difference between the two approximations can be found in [148].
[4] The real data taking procedure has been discussed in Section 2.3 where the focus in Chapter 4 is on the reconstruction of the physics objects.

The incoming particles can be influenced by the magnetic field. They in addition, undergo processes like multiple scattering within the active material of the detector. Hence, dedicated models for such interactions together with the map of the magnetic field are needed. Similar to the real data taking, the modeled interactions create simulated electronic signals which are digitized before going through the reconstruction steps.

3.4.1 Pile up simulation

The $pp \rightarrow XY$ process which fires the triggers, so-called the signal collision, is not the only scattering happening during the beam crossing. There are extra collisions between other protons in the colliding beams which "pile up" on top of the signal collision. The number of pile up interactions per signal collision is therefore increased by luminosity. The pileup can be a general term including also the diffractive processes in which a proton emerges intact from the interaction with a few percent of energy loss. The interaction of particles produced by pile up collisions with the detector material contaminates the signal collision. Such interactions influence the physics objects and observables, hence introduce a systematic effect on the physics analyses.
Regarding the CPU-time needed for the event simulation ($\sim$ min/event), the simulation of pile up collisions is performed separately. To reflect the randomness of pile up processes, the number of pile up collisions per signal collision, N_{pu}, in PYTHIA is described by a Poisson distribution, where the mean value $< N_{pu} >$ changes with the luminosity. Events with soft partonic interactions, minimum bias events, are generated for pile up simulation where each event contains a vertex of interaction.
A number of N_{pu} pile up collisions are randomly taken from the minimum bias sample. The list of non-decayed particles, produced in the minimum bias interaction and to be later decayed and propagated through the detector material, is associated to each pile up vertex. The pile up vertices and their decay products are then mixed with the signal collision. For the 2010 pile up simulation the choice of $< N_{pu} >= 1$ is made.
The addition of pile up events to the signal collision leads to an increase in the multiplicity of charged and neutral particles. Figure 3.12 illustrates the charge multiplicity introduced in Section 3.2.4 for the b-flavored jet with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ in the semi-electron final state of $t\bar{t}$ events. The b-quark jets charge multiplicity has been increased but only slightly due to the small number of additional pile up vertices.
The pile up category which has been explained is known as "in-time" where the "out-of-time" pile up is related to events coming from bunch crossings before and after the triggered event and depends on the time response of different subdetectors.

Overview on the simulated samples

The method developed in this thesis is based on the semi-electron final state of top quark pair events, $t\bar{t} \rightarrow W^+W^-b\bar{b} \rightarrow qq'e\nu_e b\bar{b}$. Hence, this decay mode is considered as signal where the other final states of $t\bar{t}$ with some additional physics processes form the background event samples. The generation of $t\bar{t}$ events in all channels has been centrally performed by the CMS collaboration using MadGraph where up to four extra

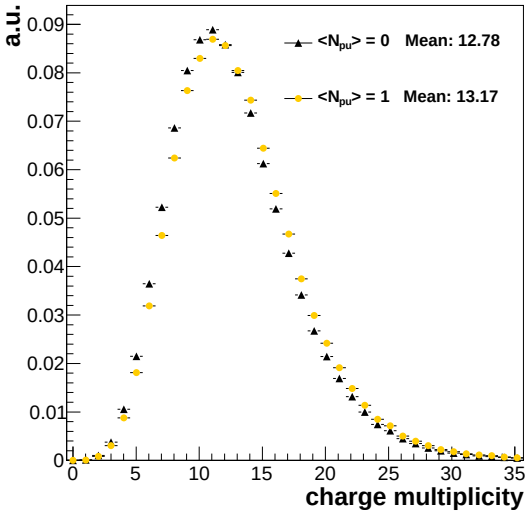


Figure 3.12: The charge multiplicity of b-quark jets in the semi-electron final state of *tbart* events, illustrated for events with no pile up and events with $< N_{pu} >= 1$.

partons are allowed in matrix element calculation.
With a similar signature to the signal events, the W and Z bosons in association with extra jets are considered as background if the vector bosons decay leptonically. Additional samples have been provided containing the production of vector bosons with two additional heavy quarks (Vqq+jets) and the W boson with an extra c quark (Wc+jets). These samples have possible overlaps with the inclusive Z/W+jets samples (see Section 5.1). The electro-weak production of single top quarks is also taken as background. These samples are all generated by `MadGraph` and interfaced with `PYTHIA` for showering where the MLM method with $E_T^{min} = 30\,GeV$ has been applied for matching.
The `PYTHIA` event generator has been used to produce the QCD multi-jets events. For a realistic estimation of QCD backgrounds, samples with huge statistics are needed. Hence, the QCD multi-jets events are generated in three $\hat{p_T}$ bins, 20-30 GeV, 30-80 GeV, and 80-170 GeV, where the variable $\hat{p_T}$ is the transverse momentum of the hard interaction in its rest frame. This has been complemented by a filtering at generator level to enhance the statistics of the events that are likely to pass the electron selection requirements.
Two sets of QCD multi-jets are generated by means of two filters which are explicitly orthogonal. Events containing an electron within the tracker acceptance and with at least an energy of 10 GeV are filtered with the `BCtoE` filter. As the name indicates, the electrons are required to be produced via the decay of b- and c-hadrons where the multi step decays are also considered. The other filter, `em-enriched` looks for stable particles in the final state which can be reconstructed as an electron candidate. The energies of generated $K^{\pm}, \pi^{\pm}$, photon and electrons with $E_T \geq 1$ GeV are clustered and asked to be greater than 20 GeV. Then, individual charged particles of the mentioned set with $E_T \geq 20$ GeV are looked for. The particles have to be isolated and to fall in the tracker acceptance.
Finally, the photon+jets events generated by `MadGraph` in three different $\hat{p_T}$ bins (40-100 GeV, 100-200 GeV, and $>$ 200 GeV) are added to the background samples. The

full detector simulation based on GEANT4 has been applied on all samples.
Table 3.2 gives the summary of simulated samples used in the thesis. The samples with the `Spring10` indicator are used to develop the method on simulation where the `Fall10` label is assigned to the simulated samples used in data analysis, for data-simulation comparisons. The tunes to simulate the underlying events are also quoted. To study the systematic uncertainties, another set of the $t\bar{t}$+jets samples has been

process	generator	σ_{eff} (pb)	#events	tune `Spring10` vs. `Fall10`
$t\bar{t}$+jets	`MadGraph`	$157^{+23.2}_{-24.4}$	1.5 M	D6T/D6T
single top ($t \to bl\nu$)	`MadGraph`			
t channel		$20.91^{+1.10}_{-1.04}$	529 k	D6T/Z2
tW channel		10.6 ± 0.8	466 k	D6T/Z2
s channel		1.36 ± 0.08	495 k	D6T/Z2
W+jets	`MadGraph`	31314 ± 1558	10 M	D6T/D6T
$Z/\gamma^*(\to l^+l^-)$+jets	`MadGraph`	3048 ± 132	1.1 M	D6T/Z2
m_{ll}> 50 GeV				
QCD `BCtoE`*	`PYTHIA`			D6T/Z2
$\hat{p_T}$: 20-30		108330	2.8 M	
$\hat{p_T}$: 30-80		138762	2.5 M	
$\hat{p_T}$: 80-170		9422.4	1.2 M	
QCD `em-enriched`*	`PYTHIA`			D6T/Z2
$\hat{p_T}$: 20-30		1719150	34 M	
$\hat{p_T}$: 30-80		3498700	42 M	
$\hat{p_T}$: 80-170		134088	5.5 M	
γ + jets*	`MadGraph`			D6T/D6T
$\hat{p_T}$: 40-100		23620	2.1 M	
$\hat{p_T}$: 100-200		3476	1.1 M	
$\hat{p_T}$: >200		485	1.0 M	

Table 3.2: The summary of the simulated samples used for the method developed in this thesis (*Spring10*) together with those used for the data-simulation comparisons (*Fall10*). The NLO cross sections for all but the QCD and γ+jets samples for which the LO was available, are taken from [145]. The notation σ_{eff} is to account for the branching ratios where only the leptonic decays are considered, for the cut on m_{ll} in the Z/γ^*+jets process and for the QCD and γ+jets filtering efficiencies. It should be noted that the experimental values of BR($W \to l\nu$) = 0.1080±0.0009 [3] is taken instead of 1/9. The uncertainties include the scale and PDF fluctuations. The notation * means the $\hat{p_T}$ are in GeV unit. the The top mass is taken $m_t = 172.5$ GeV.

centrally generated by `MadGraph` with different tunes for the relevant parameters as described in Section 3.3.1. They are all produced with the input top mass value of $m = 172.5$ GeV and with the D6T tunes for underlying events. The statistics of the

samples ranges between 0.8 M and 1.5 M events.

Chapter 4

Reconstruction and identification of the physics objects

After the hard scatterings in pp collisions at the LHC, the non-stable particles decay subsequently to particles that need to be long-lived enough to leave their "footprints" as hits or energy deposits in the detector. At the detector level these footprints are actually the electronic signals that are collected from the relevant subdetectors to built up the primitive objects. These primitive objects are then combined by means of dedicated algorithms to form the familiar physics objects like electrons and jets.
The quality of a physics analysis relies on the goodness of the reconstruction performance of the physics objects. Therefore the reconstructed physics objects are asked to meet some identification criteria depending on their expected signature in the detector. While some of the objects may not be qualified enough to be used in the analysis, other physics objects with similar signatures could be wrongly identified by the identification algorithms. Therefore extra methods with good performances are needed to estimate the efficiency and mis-identification rate of such algorithms. To select the $t\bar{t} \rightarrow W^+bW^-\bar{b} \rightarrow e\nu_e qq'b\bar{b}$ events with one electron and multiple jets in the final states, the experimental signature which is focused in this thesis, one of the essential requirements is the presence of a well-reconstructed electron i.e. an electron candidate passing the identification criteria. In Section 4.1 the reconstruction of the electron in the CMS experiment is explained together its isolation and identification variables. A data-driven method to measure the efficiency of the electron identification and isolation algorithm is developed in Section 4.2. It is described how to use these efficiencies in the measurement of the top quark cross section.
The reconstruction of jets, the other components of the semi-electron $t\bar{t}$ final state, together with the jet identification variables is addressed in Section 4.3. Different algorithms to identify the flavor of the jets are used in the CMS for the analyses involving jets originated from b-quarks. Section 4.4 is devoted to the description of these algorithms in addition to the methods developed for the measurement of their efficiencies.

4.1 Electron reconstruction

Starting from the interaction point, the electron leaves hits in the tracker layers and releases its energy finally in the ECAL. Therefore, the electron object has two major components: a supercluster (Section 2.2.2) in the ECAL matched with a track segment (Section 2.2.1) in the inner tracking system. However, the reconstruction of the electron is more challenging than this simple matching and needs special treatments both in th reconstruction of the track and to select the supercluster.
Within the tracker material distributed in front of the ECAL, the electron radiates bremsstrahlung photons while it is bent in the presence of the strong magnetic field. Thus depending on the electron transverse momentum, $p_{\mathrm{T}}^{\mathrm{elec}}$, the energy which reaches the ECAL is spread in ϕ. The tracker material varies with η and the electron can emit a considerable amount of photons if it traverses for example through the edge of the barrel ($|\eta| \approx 1.5$) where the tracker material budget is about $2.0X_0$ [1]. The emission pattern can vary from event to event in a "non-Gaussian" way where the amplitude of such a fluctuation increases by the amount of the tracker material. These non-Gaussian effects should be well taken care of in the energy measurement in the ECAL and the momentum estimation in the tracker as well as in the electron identification algorithm.
The electron reconstruction [151] can be started by looking either for a suitable supercluster in the ECAL (*ECAL driven* seeding) or for a track candidate in the high purity track collection (*Tracker driven* seeding). While the former performs efficiently for isolated electrons with $p_T \gtrsim 10\,\mathrm{GeV}$, the latter helps to recover the low p_T electrons and the electrons inside jets. With the large transverse momentum, the isolated electrons in $t\bar{t}$ events can be well reconstructed with the ECAL driven approach. In the analysis presented in this thesis, the electrons are asked to have $p_T > 30\,\mathrm{GeV}$.

ECAL driven electron seeding

The electron supercluster is reconstructed by a hybrid algorithm in the barrel while in the endcap, the Multi5×5 method is utilized (see Section 2.2.2 for the algorithms description). In both cases, the clusters are grouped if their position lies within a ϕ road of width 0.3 rad.
The superclusters with $E_T > 4\,\mathrm{GeV}$ are further asked to pass a hadronic veto. Since the electrons are supposed to not leak in the hadron calorimeter, the amount of energy found behind the supercluster within a cone of $\Delta R = 0.15$ [2] in the HCAL needs to be relatively small. It introduces the H/E quantity where H (E) is the energy deposited in the HCAL (ECAL). For supercluster candidates H/E must be less than 0.15.
Assuming the supercluster candidates are on the helix trajectory of an electron, the path is backpropagated towards the interaction point with both charge hypotheses. The tracker hit pairs or triplets are searched in a ϕ-z window around the extrapolated trajectory in the tracker system. In the case of triplets it is enough to have two hits out of three inside the window. To gain in efficiency in the forward regions where the pixel detector is limited, one can take advantage of both the pixel and the TEC detector to

[1] Since the Physics TDR [81], the description of the tracker material has become more realistic leading to an overall budget peaking at $2.0X_0$ instead of $1.5X_0$ for a pseudorapidity $|\eta| \approx 1.5$.
[2] $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$.

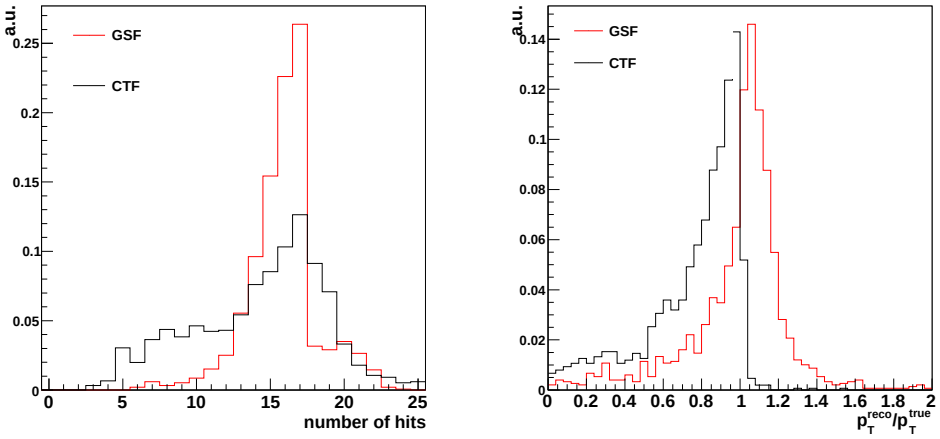
make the hit pairs/triplets.
The ϕ-z window is wider at the beginning because of the uncertainty on longitudinal position of the beamspot and to account for residual effects. Once the first hit is found, the track parameters are re-evaluated, the uncertainties decrease and the next hit is expected to be in a smaller window around the trajectory. A further refinement on the first window size is performed to reduce the probabilty for a jet of particles to be mistaken as an electron. In general the window is narrower for high energy superclusters. The same strategy is used at the HLT for the electron trigger.

Tracker driven electron seeding

For an efficient reconstruction of low momentum electrons, the ϕ road was first extended to 0.3 rad [81], the size which is still in use for superclustering. The supercluster seed threshold was lowered at the same time to 1 GeV instead of 4 GeV. However for electrons with $p_T \lesssim 10\,\text{GeV}$, the 0.3 rad ϕ extension in the supercluster finding is still insufficient to collect all of their energy. In fact, since the soft electron is bent more in tracker, the bremsstrahlung photons can make clusters in the ECAL which are well separated from the electron cluster. Widening the ϕ band or lowering the seed threshold is not the best way to recover these electrons since there is a good chance for the noise contributions. On the other hand for the electron in jets, the energy deposited by neutral particles in the jet can contribute in the supercluster energy hence biases the electron track finding procedure.
The tracker driven electron seeding, developed in the context of the particle-flow reconstruction [152], has proved to be suitable for the low-p_T electron reconstruction. It has increased the electron reconstruction efficiency at $p_T^{\text{elec}} = 5\,\text{GeV}$ by 12.5% while for high p_T^{elec} electrons, the gain is 1-2% [151]. More details about tracker driven electron seeding can be found elsewhere [153].

Electron track reconstruction

Similar to the general procedure of the track reconstruction the track finding is the first step in the electron tracking. A dedicated combinatorial Kalman Filter in which the energy loss is modeled with the Bethe-Heitler [154] function is used to find the preliminary electron track candidates. The default Kalman filter is a linear least-squares estimator approximating the electron energy loss by a single Gaussian, leading to crude results. Hence, it has been generalized to the non-linear Gaussian-Sum Filter (GSF) [155, 156] in which the Bethe-Heitler distribution is estimated by a mixture of Gaussians with different weights. More details about the implementation of this algorithm for the CMS electron tracking can be found in [157].
In electron tracking, the maximum number of compatible hits in each layer is limited to 5 to control the combinatorics. To cope with the curvature changes due to the bremsstrahlung effect, a very loose χ^2 requirement (< 2000) is applied in the fit. Since the hit finding algorithm tolerates one layer without a hit on the track trajectory, this could contaminate the track collection with electrons from conversion. A stronger χ^2 requirement of the $\chi^2 < 90$ is imposed in these cases. Figure 4.1 (a) illustrates the



(a) (b)

Figure 4.1: Number of electron track hits (a) and the track momentum magnitude divided by electron true momentum (b) for CTF and GSF algorithms in $t\bar{t}$ events.

number of CTF [3] and GSF track hits for electrons in a $t\bar{t}$ sample. The GSF algorithm seems more successful in dealing with the bremsstrahlung curvature changes hence recovering more track hits. The GSF algorithm is used to refine the track parameters for the products of the finding stage. There, the distributions of all state vectors are allowed to be weighted sums of Gaussians, instead of single Gaussians as in the default Kalman filter. The propagation to the next layer is done independently for each component and does not affect the weights. Despite of the full set of information, in practice the state vector is approximated by the component with the highest weight ("mode") which has shown to result in a better precision compared to the weighted mean of components for the tracks with low radiation [151]. Comparing the GSF tracking algorithm using the "mode" estimation with the CTF algorithm, the p_T reconstruction shows a less biased measurement for the tracks subjected to bremsstrahlung emission (Figure 4.1 b).

Track-supercluster matching

For those ECAL driven electrons which succeeded to pass the $E_T > 4\,\text{GeV}$ cut and the calorimeter veto $H/E < 0.15$, the geometrical constraints are imposed to match the electron track with the supercluster. The closest approach of the supercluster to the electron track extrapolated from the innermost tracker layers, introduces the $(\eta_{\text{in}}^{\text{extrap}}, \phi_{\text{in}}^{\text{extrap}})$ coordinate on the track curve as the electron position before the ECAL. The difference in η (ϕ) between the supercluster position (see Section 2.2.2) and the track position before the ECAL needs to be $|\Delta\eta_{\text{in}}| < 0.02$ ($|\Delta\phi_{\text{in}}| < 0.15$) in both barrel and endcap.

Ambiguity removal: The described matching could involve ambiguities in the sense that the emitted bremsstrahlung photons may undergo a so-called conversion and pro-

[3] The general track reconstruction using the Combinatorial Track Finder is detailed in Section 2.2.1.

duce an e^+e^- pair. The tracks of such secondary electrons tend to end up in the same supercluster as the primary electron. Hence, different electron candidates are sharing the same supercluster. In particular when photons take more than 50% of the electron energy, the predicted position in the next layer would be closer to the photon (that converts to secondary electrons) than the primary electron after emission.
To resolve this ambiguity, the electrons which have the supercluster in common are classified according to their innermost track hit position. Between two candidates with the innermost track hit in different layers, the one with its hit at more inner layer is taken. If the innermost track hits are in the same layer for two ECAL driven electrons, candidates are judged by their E_{sc}/p_{trk} value. For the cases where a track is in common between two superclusters, the ambiguity is also resolved based on the value of E_{sc}/p_{trk}.

Electron momentum determination

To measure the momentum of electrons, CMS takes advantage of both the inner tracking system and the ECAL. While for low p_T electrons and the electrons in the ECAL crack regions the tracker momentum estimation is more precise, for energetic electrons the ECAL energy measurement has a better resolution. Hence the electrons are classified regarding to their p_T and some other properties. For each class a dedicated momentum determination is performed. The integrated amount of energy an electron loses along its trajectory due to the bremsstrahlung effect, f_{brem}, plays a key role in this classification. This value is estimated as the normalized difference between the momentum magnitude at the outermost and innermost track position. The electron classes are

golden with low radiation and good track-supercluster matching: Supercluster contains one cluster (no bremsstrahlung subcluster); $E_{sc}/p_{trk} > 0.9$ and $f_{brem} < 0.5$.

big brem but no evidence of energy loss effects: Supercluster contains one cluster; $E_{sc}/p_{trk} > 0.9$ and $f_{brem} > 0.5$.

showering with the energy pattern quiet influenced by bremsstrahlung losses: Supercluster contains one cluster but the E_{sc}/p_{trk} and f_{brem} are such that the electron does not fit within the other classes; or the supercluster constitutes several clusters.

One could add the *crack* electrons to this list to account for the electrons for which the η value of the starting crystal of their supercluster is either close to the boundary of the ECAL barrel modules or near the ECAL barrel-endcap boundaries. Before making the combined ECAL-tracker momentum estimation, it is checked if the supercluster energy needs extra corrections (see Section 2.2.2). Although an offset of $\simeq 0.3\%$ is found for showering electrons and a residual $\Delta\eta$ trend is observed for *golden* electrons, the effects are small enough so no extra correction is applied [151].
The combined momentum measurement is carried out according to the E_{sc}/p_{trk} variable which contains the information from both subdetectors and is sensitive to the amount of bremsstrahlung. Figures 4.2 (a) and 4.2 (b) illustrate the sensitivity of the electron

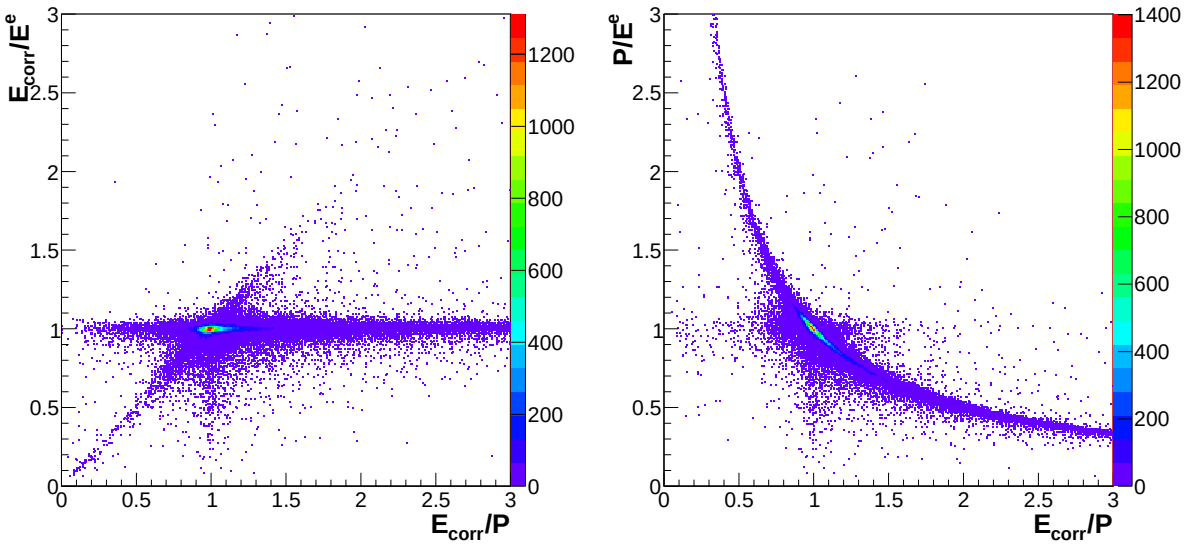


Figure 4.2: The P/E^e and E_{corr}/E^e as a function of E_{sc}/p_{trk} for barrel electrons in $t\bar{t}$ events. E^e is true energy of the electron while P and E_{corr} denote the electron track momentum and corrected supercluster energy. In $E_{sc}/p_{trk} > 1$ the tracker momentum is always underestimated. The $E_{sc}/p_{trk} < 1$ region is dominated by showering electrons where either the ECAL energy or the tracker momentum is not correct.

track momentum and supercluster energy estimation to E_{sc}/p_{trk} for barrel electrons in $t\bar{t}$ events. Another important feature considered in the combination is the opposite behavior of $\sigma(E_{sc})/E_{sc}$ and $\sigma(p_{trk})/p_{trk}$ which is shown in Figure 4.3. The weighted mean of the tracker and supercluster measurement is taken as the electron momentum if $|E_{sc}/p_{trk} - 1| < 2.5\,\sigma(E_{sc}/p_{trk})$. The weights are the normalized inverse of the variance of each measurement, hence the more precise measurement contributes more. The supercluster energy is used for all other cases except for the *golden* electrons with $E < 13\,\text{GeV}$ and $E_{sc}/p_{trk} < 1.15$ in the endcap, and for the three cases in the barrel where only the tracker measurement is taken into account:

- *golden* electrons with $E < 15\,\text{GeV}$ and $E_{sc}/p_{trk} < 1.15$.
- *showering* electrons with $E < 18\,\text{GeV}$ and $E_{sc}/p_{trk} < 1 - 2.5\,\sigma(E_{sc}/p_{trk})$.
- *crack* electrons with $E < 60\,\text{GeV}$ and $E_{sc}/p_{trk} < 1 - 2.5\,\sigma(E_{sc}/p_{trk})$.

The common feature in each category is either the low p_T of the electron or the imperfect ECAL measurement. With the combined measurement, the precision is in particular improved for electrons in the $p_T \lesssim$ 20-30 GeV range. In addition, a resolution of 1% is achieved for *golden* electrons [151]. The resolution of the combined measurement for electrons in $t\bar{t}$ events is shown in Figure 4.3 as a function of the electron energy. In each bin of the electron energy, a Gaussian is fitted to the relative difference between the energy or momentum of the reconstructed electron candidate with that of the generated electron. The width of the fitted function is taken as the energy or momentum

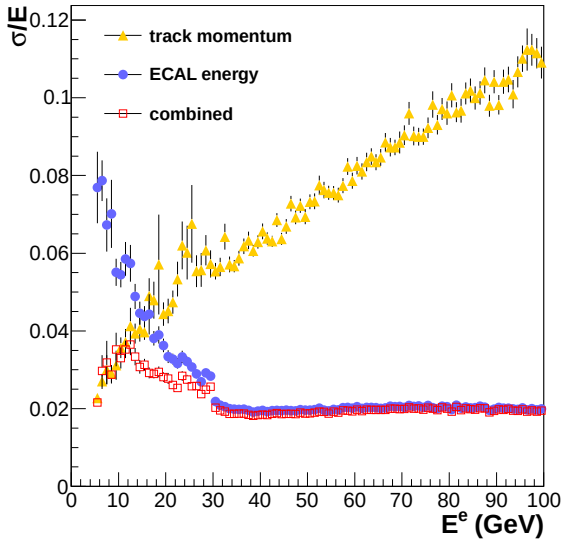


Figure 4.3: The resolution of the energy (momentum) measurement in the ECAL (tracker) as well as the resolution of the combined estimation for electrons in $t\bar{t}$ events.

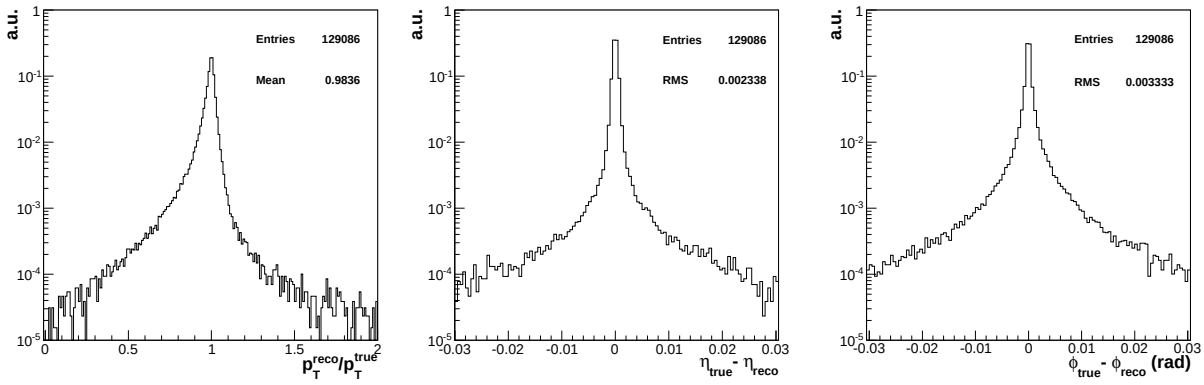


Figure 4.4: From left to right, the momentum and the geometrical properties of reconstructed electrons are compared to those of the true electrons in $t\bar{t}$ events.

resolution. It can be seen how the combined momentum determination has improved the resolution.
Within the same sample, the momentum and the geometrical properties are compared between the reconstructed and true electron in Figure 4.4. The relative momentum is close to one where the small bias is due the electrons subjected to the bremsstrahlung emissions. The determination of η seams to be more precise than the estimatio of ϕ coordinate regarding the RMS value which is slightly less in the $\Delta\eta$ distribution. The energy spread in ϕ direction as a consequence of the bremsstrahlung emission leads to a wider $\Delta\phi$ distribution. The electron reconstruction performance has been studied using the early 7 TeV CMS data in 2010 where the simulation has well predicted the data [158].

4.1.1 Electron Identification

The fact that an electron object consists of a set of information form the tracker and calorimeter, provides various choices of variables for the electron identification [159]. These variables ranging from the quality of track-supercluster matching to the amount of tracker and calorimeter activities around the electron, the energy deposit pattern of superclusters, etc. are used to discriminate the prompt electrons (e.g. from the W/Z-boson decay) from the hadron showers and/or the electron in jets. The most discriminating quantities are given to neural network or likelihood algorithms along with the cut based methods to determine the quality of the electron [160]. While ultimately the most performant selection could be obtained using the multivariate techniques, for early data taking the cut based selections can provide a useful tool to understand the data and to make a robust and efficient selection.
For the cut based selection either a fixed threshold is applied on all different type of electrons or with a further refinement, the cut values change for different categories where the categories are defined according to f_{brem} and E_{sc}/p_{trk} [160]. Both of the approaches were investigated prior to the 2010 data taking. The cut values were optimized to reject the backgrounds as much as possible while keeping the prompt electrons [158]. In $t\bar{t}$ analyses, the simple cut based selection has been taken for the sake of simplicity and transparency. The top quark analyses also take advantage from the possibility of decomposition in the current cut based identification tool, hence the isolation and conversion rejection are applied separately.
Apart from the conversion and isolation criteria, the simple cut based identification is based on H/E, $\Delta\eta_{\text{in}}$ and $\Delta\phi_{\text{in}}$ together with $\sigma_{i\eta i\eta}$ which is a shower shape variable taken from the covariance matrix using logarithmic energy weights for crystals in a cluster [4]. The conceptual idea behind $\sigma_{i\eta i\eta}$ is to check if the electron shower is narrow in η as expected. Figure 4.5 illustrates the shape difference of the mentioned electron identification variables between the semi-electron final state of $t\bar{t}$ and the QCD multi-jet events. All electron are ECAL-driven, required for a $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$

[4] The following definition is used: $\sigma^2_{i\eta i\eta} = \frac{\sum_{\text{i}}^{5\times5} w_i.(n_i\times0.0175+\eta_{seed}-\bar{\eta}_{5\times5})^2}{\sum_i^{5\times5} w_i}$, where n_i is the number of the crystal in the η direction from the seed and 0.0175 is the average η size of the ECAL crystals; $\bar{\eta}_{5\times5}$ is the energy weighted η mean of the cluster and w_i is the logarithmic energy dependent weight function; The summation runs over all crystals in a 5×5 cluster.

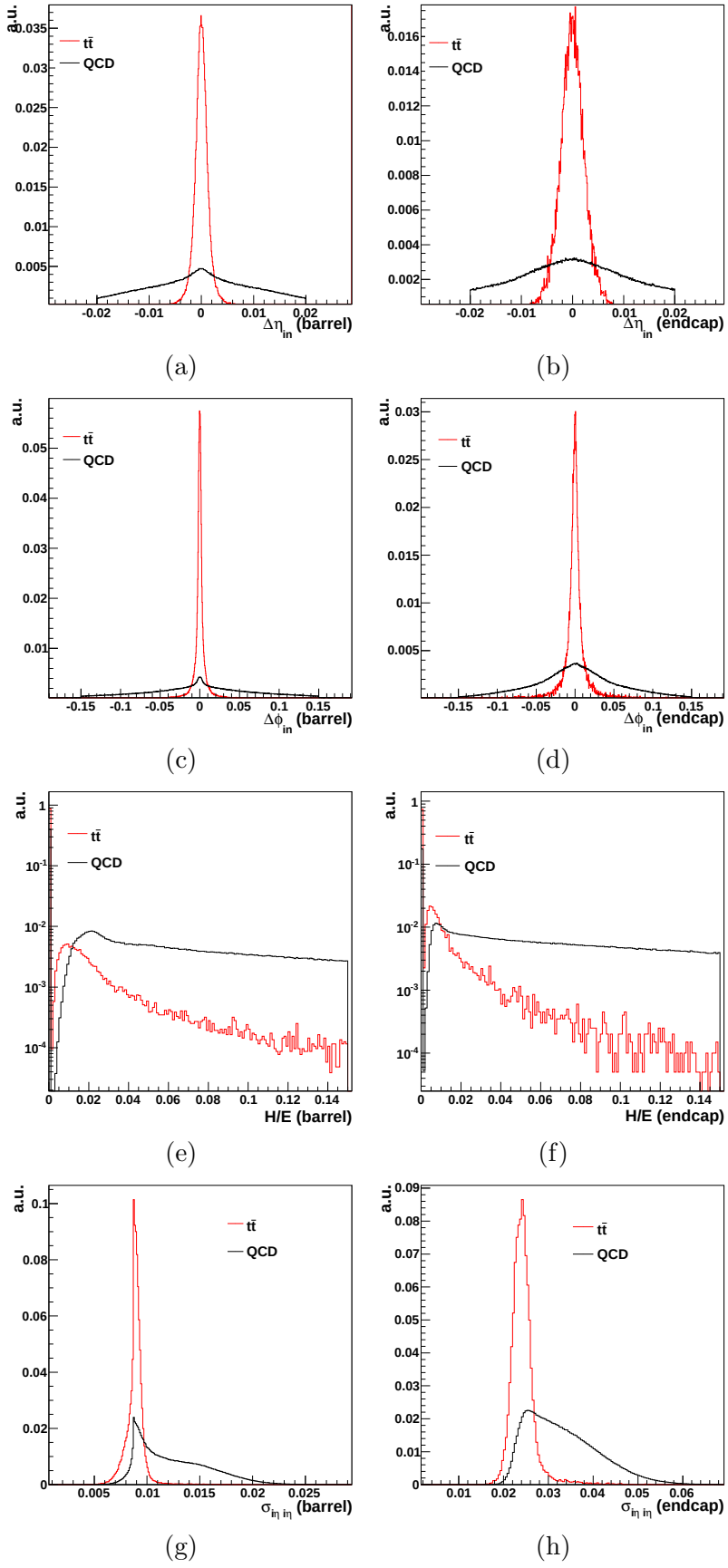


Figure 4.5: Comparison of the electron identification variables in the semi-electron final state of $t\bar{t}$ and QCD multi-jet events for the electron candidates with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$ (ECAL gap excluded). The electron candidates in the semi-electron final state of $t\bar{t}$ events are checked to be matched with an electron from the W boson decay.

where electrons having the supercluster in the ECAL barrel-endcap transition region are rejected. In $t\bar{t}$ events, the electron candidates are further matched with generated electrons in the W boson decay. Both in the barrel and endcap, the variables are discriminating between the prompt electron candidates in $t\bar{t}$ and poor or fake electron candidates in QCD where the distributions are in general wider in the barrel. Except for $\sigma_{i\eta i\eta}$, the effect of the cut at the reconstruction level is visible on the distributions. The cuts on this set of variables referred from now on as "electron ID variables", are tuned for a maximum background rejection with a given signal efficiency. Hence, different working points (WPs) are defined. Table 4.1 presents the upper limits on the electron ID variables for WP70 and WP95 that are corresponding to a signal efficiency of 70% and 95% respectively. The signal here is defined as electrons in $W \to e\nu_e$ events while electrons from processes that are considered to be background to the $W \to e\nu_e$ process, are rejected. An E_T cut of 25 GeV is applied on the electron supercluster along with the requirement that no second electron with supercluster $E_T > 20$ GeV is in the event. The electron supercluster is asked to be out of the barrel-endcap transition in the ECAL.

Variable	Cut values for WP70		Cut values for WP95	
	barrel	endcap	barrel	endcap
H/E	0.025	0.025	0.15	0.07
$\Delta\eta_{\text{in}}$	0.004	0.005	0.007	0.01
$\Delta\phi_{\text{in}}$	0.03	0.02	0.8	0.7
$\sigma_{i\eta i\eta}$	0.01	0.03	0.01	0.03

Table 4.1: The upper limits for the electron ID variables for 70% and 90% signal efficiency in the barrel and endcap.

4.1.2 Electron isolation

To distinguish between the electrons produced in high p_T processes like $t\bar{t}$ and those produced within the jets in QCD multi-jets backgrounds, isolation requirements are imposed on electrons. The idea is to measure the activities from other particles around the electron in different subdetectors. Within a cone around the electron ($R = 0.3$ for the $t\bar{t}$ analysis), the transverse energy (momentum) is summed up in the calorimeters (tracker) while the E_T (p_T) associated to the particle itself is excluded. In general a smaller internal cone around the particle is excluded and because of the bremsstrahlung and conversion it sometimes gets a more complicated shape than a simple cone.

Isolation in tracker: The electron footprint in the tracker is influenced by conversions in such a way that the secondary electrons make a strip-like shape along ϕ in the η-ϕ plane. A superposition of a conical and strip veto is therefore used to remove the electron hits. While in the endcap the η-width of the strip is

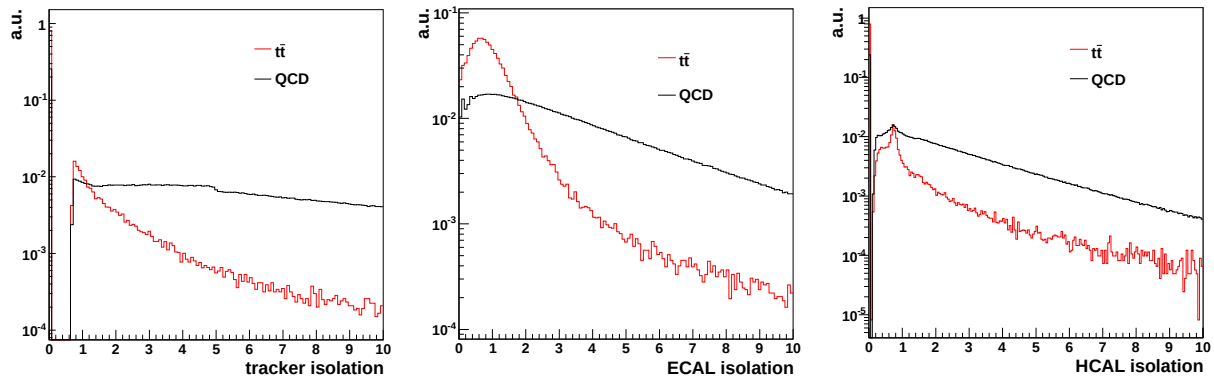


Figure 4.6: The comparison of isolation variables in different subdetectors for the electrons in the semi-electron final state of $t\bar{t}$ and the QCD multi-jet events. From left to right, isolation in the tracker, ECAL and HCAL, respectively.

optimized to 0.005, in the barrel the removal is done only conical. The internal cone size is optimized to 0.015 for barrel and endcap [161].

ECAL isolation: In the ECAL the electron footprint is again a composition of a bulge around the electron and a strip in the ϕ direction. While the former is considered as the electron energy deposit, the latter exists due to the radiation and electron conversion. The innercone radius of three ECAL crystals gives the best performance in the barrel and endcap. The η-width of the strip veto is optimized to one crystal in the barrel and 1.5 crystals in the endcap [161]. An additional energy threshold of 80 MeV (0.2 GeV) is applied on the ECAL hits in the barrel (endcap) to avoid the noise contribution.

Isolation in HCAL: The electrons are expected not to leak in the HCAL. Nevertheless for possible energy deposits, a conical veto is applied in the HCAL as well. The best performance is achieved with a veto cone of $R = 0.05$ and using the HCAL towers with energy greater than 0.5 GeV. However the recommended cone size is 0.15 to make the isolation independent of the H/E cut which is applied at reconstruction and identification [161].
The isolations in different subdetectors are combined to give the optimal performance while the p_T of the electron is also considered,

$$\text{relIso} \equiv \text{relative combined isolation} = \frac{\text{Iso}_{\text{tracker}} + \text{Iso}_{\text{ECAL}} + \text{Iso}_{\text{HCAL}}}{\text{p}_{\text{T}}^{\text{elec}}}. \quad (4.1)$$

Different isolation pieces contributing in Equation 4.1 are illustrated in Figure 4.6 for the semi-electron final state of $t\bar{t}$ and the QCD multi-jet events where a better discrimination is seen in the ECAL isolation variable. The electrons are ECAL driven with $p_T > 30$ GeV and $|\eta| < 2.5$. In $t\bar{t}$ events, electron candidates which are not coming

from the W boson decay are rejected. For the same set of electrons, the isolation variables are combined into *relIso*, following Equation 4.1, shown in Figure 4.7. It can be seen that a cut of $relIso < 0.1$ rejects most of the background electron candidates while keeping a fare amout of the well-defined ones.

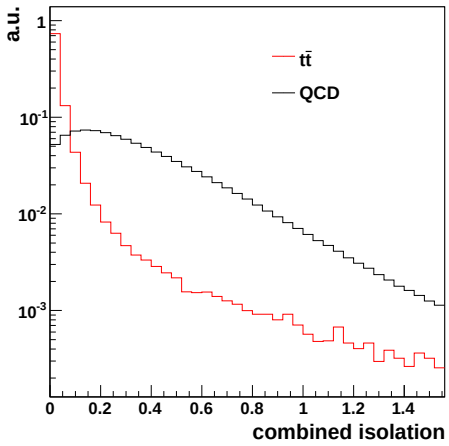


Figure 4.7: Combined isolation as in Equation 4.1. Electrons in $t\bar{t}$ event are compared to the electrons in QCD multi-jets.

4.1.3 Conversion rejection

A non-negligible background to prompt electrons is the electrons from photon conversion, known as conversion electrons. Depending on the cuts applied, the amount of the conversion electrons in the QCD multi-jet events is roughly approximated to be ∼15-30% [160]. One can distinguish between the prompt electrons and those from photon conversion by looking at the transverse impact parameter (d_0) or the hit pattern of the electron track as well as by searching for the track of the other particle initiated from the same photon.

Impact parameter: Since the photon conversion happens within the tracker material, the electrons from conversions have on average a larger distance to the beamspot, i.e. a greater d_0. The tight $d_0(< 200\mu m)$ requirement can reject the conversion electrons while keeping the prompt candidates.

Hit pattern: For the same reason as the impact parameter, the electrons from conversions may not necessarily have hits in the innermost tracker layers while for prompt electrons coming from the primary vertex the track has hits almost in all tracker layers. Therefore the number of missing tracker layer can discriminates the electrons from conversions.

Partner track search: A signature for a conversion is the tracks of an electron-positron pair which are parallel at the point of decay, and remain so in the $r - z$ plane. Based on this feature, all oppositely charged CTF tracks within a cone of $R = 0.3$ around the electron are taken [160]. One variable which is checked is the

$x - y$ distance between each CTF track in the cone and the electron GSF track where they become parallel after extrapolation. Another quantity is defined as the difference in the cotangent of the CTF and the GSF tracks polar angles,

$$\Delta\cot(\Theta) = \cot(\Theta_{\text{CTF track}}) - \cot(\Theta_{\text{GSF track}}). \tag{4.2}$$

Electrons which have a partner track with $|\Delta\cot(\Theta)| < 0.02$ and $|\text{Dist}| < 0.02\,\text{cm}$ are considered as coming from conversion and so are discarded.

The variables used for conversion rejection are shown in Figure 4.8 for electrons in the semi-electron final state of $t\bar{t}$ and the QCD multi-jet events. Electrons are ECAL driven and have already passed the $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$ requirements. Electron candidates in QCD multi-jets events have larger impact parameter and more missing hits than the electrons in $t\bar{t}$. The two dimensional distribution of $(\text{Dist}, \Delta\cot\Theta)$ is illustrated for the QCD multi-jets and $t\bar{t}$ events where the efficiency of the partner track veto is also indicated. About 10% of the electrons in $t\bar{t}$ events are rejected by the partner track veto where the rejection for the electron candidates in QCD multi-jets events is about 16%. The cuts for the partner track veto are optimized using single electron and single photon simulated samples and the effect of other activities in the event is therefore not considered [162].

4.2 Electron isolation and identification efficiency

The estimation of the isolation and identification efficiency is of great importance for analyses based on the prompt electron selection. These kind of efficiencies can be easily calculated in simulation, however a more consistent measurement is achieved if the efficiencies are derived from data itself. In data one needs to be confident that the candidate on which the efficiency is measured, is with a very high probability an electron. For this reason a method called *Tag&Probe* is developed to be applied on resonances that have electrons in the final state. Many analyses including $t\bar{t}$ rely on the efficiency results from the Tag&Probe method that is applied on $Z \to ee$.
The idea of the method is to look for the events with two electrons in the final state, ask one of the electrons to be of high quality (where the definition of quality depends on what is to be measured) and require the second electron to give with the first one, an invariant mass of about the Z-boson mass. The high quality electron is called *tag* while the second one is *probe* on which the efficiency will be estimated. The *probe* candidate is considered an electron because of the Z-boson mass criteria and the presence of the *tagged* electron.
Under the assumption of independence between the isolation and identification, one can factorize the efficiency as

$$\epsilon = \epsilon_{iso} \times \epsilon_{id}. \tag{4.3}$$

According to the Tag&Probe method, electron pairs are made and events can be divided in three categories. In the first category, both electrons are isolated and pass the identification criteria

$$N_{\text{TT}} = \epsilon^2_{\text{iso}} \cdot \epsilon^2_{\text{id}} \cdot N_{\text{total}}. \tag{4.4}$$

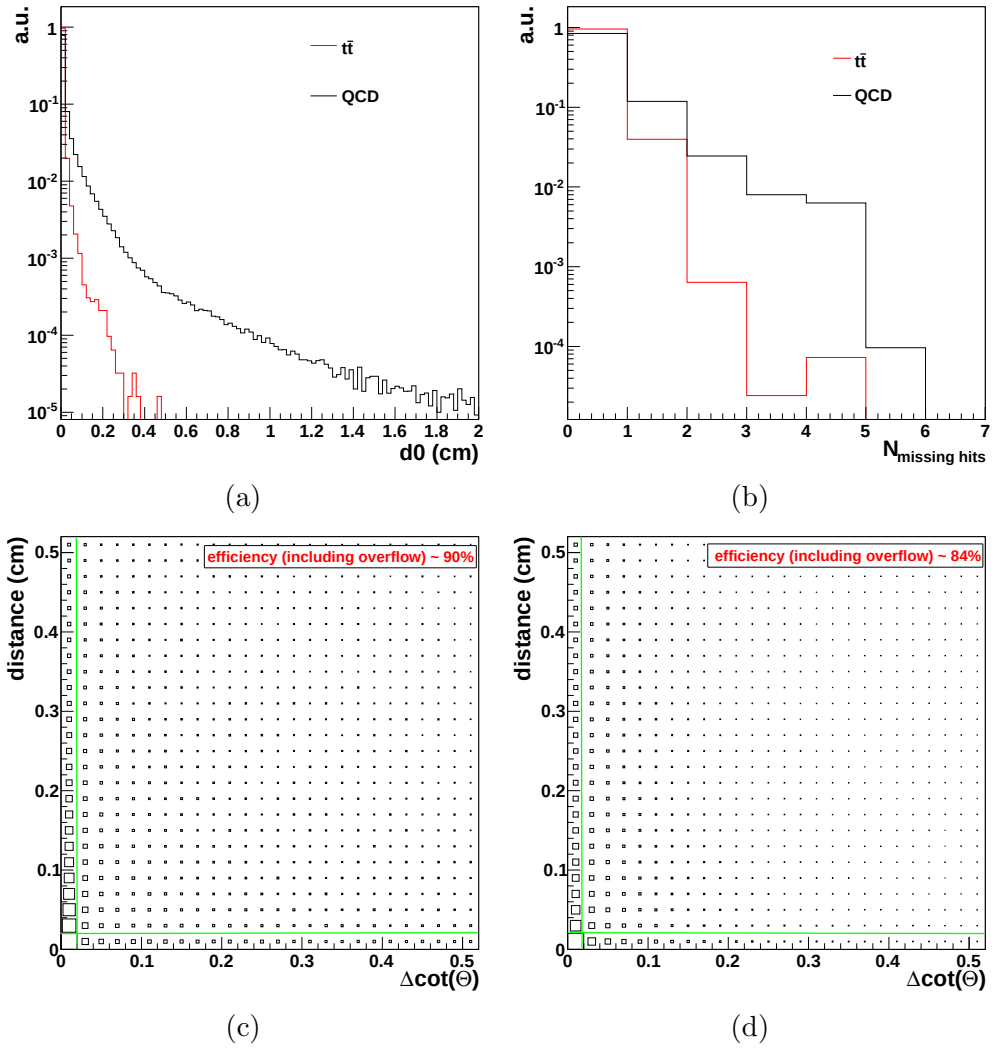


Figure 4.8: The quantities sensitive to conversion, the transverse impact parameter of the electron w.r.t the beam spot (a) and the number of missing tracker layers (b) together with the two dimensional (Dist, $\Delta\cot\Theta$) distributions for the electron candidates in $t\bar{t}$ (c) and QCD multi-jet (d) events. The electron candidates in $t\bar{t}$ are checked to be coming from the decay of the W boson. The lines in the two dimensional plots indicate the cut values on the Dist and $\Delta\cot\Theta$ quantities.

The second category consists of pairs for which the *probe* passes some quality cuts, for example isolation, but not the identification criterion

$$N_{\mathrm{TP}} = 2\,\epsilon_{\mathrm{iso}}^{2} \cdot \epsilon_{\mathrm{id}} \cdot (1 - \epsilon_{\mathrm{id}}) \cdot N_{\mathrm{total}}. \tag{4.5}$$

The factor 2 is to account for the permutation between the two electrons in the pair. The last group of pairs contains the *tag* together with the *probe* that fails the isolation in this example

$$N_{\mathrm{TF}} = 2\,\epsilon_{\mathrm{iso}} \cdot \epsilon_{\mathrm{id}} \cdot (1 - \epsilon_{\mathrm{iso}}) \cdot N_{\mathrm{total}}. \tag{4.6}$$

Combining the equations from three categories, the isolation efficiency can be derived as follow

$$\epsilon_{\mathrm{iso}} = \frac{2N_{\mathrm{TT}} + N_{\mathrm{TP}}}{2N_{\mathrm{TT}} + N_{\mathrm{TP}} + N_{\mathrm{TF}}}. \tag{4.7}$$

The identification efficiency can be calculated using a similar argument. This data-driven method can be applied both on the simulated samples containing Z-boson decay final states and on the collision data. Hence instead of the absolute values of efficiencies, the ratio between the efficiencies in simulation and collision data is calculated

$$SF = \frac{\epsilon(data)}{\epsilon(simulation)}. \tag{4.8}$$

This Scale Factor (SF) relies more on the difference between the data and the simulation rather than the event properties. Therefore it is more general and can be applied on other types of physics processes like $t\bar{t}$. However as it will be detailed in Section 4.2.1, the difference between the physics processes introduces a systematic uncertainty which needs to be taken into account.

4.2.1 Electron efficiency in $t\bar{t}$ events

To measure the $t\bar{t}$ cross section in the electron plus jets final state within the 2010 data [25, 163], only events firing an electron trigger are considered. The events are further asked to have a reconstructed electron with $p_T > 30\,\mathrm{GeV}$ and $|\eta| < 2.5$ where the supercluster position of the electron should be out of the ECAL gap. The z-coordinate of the electron candidate has to be close to the primary vertex, $\Delta z < 1\,\mathrm{cm}$. Electron candidates are kept if they pass the WP70 identification requirements and have a combined relative isolation less than 0.1. The electrons from conversions are rejected. All these requirements are explained in Sections 4.1.1- 4.1.3.
For the final $t\bar{t}$ cross section, Equation 2.1 can be written as

$$\sigma_{t\bar{t}} = \frac{N_{selected\,events}}{\mathcal{L} \cdot \epsilon_{\mathrm{selection}}^{\mathrm{electron}} \cdot \epsilon_{\mathrm{selections}}^{\mathrm{other}}}, \tag{4.9}$$

where $\epsilon_{\mathrm{selections}}^{\mathrm{other}}$ denotes the efficiency of extra selection criteria including the jets and possible b-tagging requirements. Here the main focus is on the electron part and the full detailed event selection can be found in Section 5.6.1.
As can be seen in Equation 4.9, the efficiencies are important ingredients for the cross

section measurements and need to be calculated elsewhere if the desire is to extract them from data. For the 2010 top-quark analyses, the information on the electron sector has been provided via a central efficiency estimation using the Tag&Probe method [164] where the invariant mass of the *tag* and *probe* candidates is requested to be in the 76 GeV< $M_{tag,probe}$ <106 GeV range. The jet selection efficiency has been estimated to be equal to one and in the analyses using b-jet identification, an in-situ b-tagging efficiency measurement has been performed [25, 163].
The electron efficiency, $\epsilon^{\text{electron}}_{\text{selection}}$, can be factorized in four pieces

$$\epsilon^{\text{electron}}_{\text{selection}} = \epsilon_{reco} \cdot \epsilon_{trigger} \cdot \epsilon_{iso} \cdot \epsilon_{id}. \tag{4.10}$$

For each piece the scale factor (Equation 4.8) is calculated and used to correct the efficiency estimated in the $t\bar{t}$ simulated events, $\epsilon^{\text{t}\bar{\text{t}}}_{\text{MC}}$. The corrected value is assumed to be the real efficiency in the $t\bar{t}$ process, $\epsilon^{\text{t}\bar{\text{t}}}_{\text{data}}$, and finally enters the Equation 4.9. This correction is coming from the data-driven efficiency measurement in $Z \rightarrow ee$ processes and is applied on $t\bar{t}$ events under the assumption of

$$\frac{\epsilon^{\text{Z}}_{\text{data}}}{\epsilon^{\text{Z}}_{\text{MC}}} = \frac{\epsilon^{\text{t}\bar{\text{t}}}_{\text{data}}}{\epsilon^{\text{t}\bar{\text{t}}}_{\text{MC}}}. \tag{4.11}$$

The idea behind this assumption is based on the knowledge about the detector behavior which is plugged into GEANT to simulate the physics processes in the detector. This information provided to GEANT may however be not perfect, which means one needs to estimate the difference between the GEANT input and the real detector functioning in the presence of the collision data. Hence the difference between the data and simulation is investigated by looking at the $\frac{\epsilon^{\text{Z}}_{\text{data}}}{\epsilon^{\text{Z}}_{\text{MC}}}$ that would be equal to one in the ideal world. A complementary assumption then is, that the imperfection in the knowledge on the detector is the same, no matter which process is investigated and this leads to Equation 4.11.
For the reconstruction efficiency in Equation 4.10, ϵ_{reco}, the scale factor has been calculated in [165] with 2.88 pb^{-1} of pp collision data at 7 TeV. The SF_{reco} values for the barrel and endcap have been used as input for the work in [164] where they needed additional corrections for the E_T acceptance. The E_T-acceptance scale factor is defined as the data/simulation ratio for an electron candidate with E^{sc}_{T} >20 GeV, $|\eta^{sc}| \leq 1.4442$ or $|\eta^{sc}| \geq 1.566$ and $|\eta| < 2.5$ to have $E_T > 30$ GeV. This transition scale factor is needed to account for the E_T requirement difference in [165] and [164].
In the case of the isolation and identification scale factors, different approaches have been cross checked to deal with backgrounds and to extract the number of events under the Z-boson mass peak. The calculated scale factors for 36 pb^{-1} CMS 2010 data are listed in Table 4.2.

The electron trigger definition in the 2010 data taking evolved so that seven different trigger paths have been introduced for $t\bar{t}$ analyses in the electron plus jets final state [164]. These trigger paths have not been available in the simulated event samples and so no trigger requirement is applied in the selection of the simulated events. Therefore instead of a scale factor, a trigger efficiency from data is used to correct the simulated event yield to match with the data. The average trigger efficiency is estimated to be $\epsilon_{trigger} = 0.982 \pm 0.001$ where the fraction of those electrons matched with the electron trigger object is considered.

	Scale factor
Reconstruction	1.001 ± 0.013 (EB) 0.999 ± 0.016 (EE)
E_T-acceptance	0.99 ± 0.01
Identification	0.96 ± 0.01
Isolation	1.000 ± 0.006

Table 4.2: The scale factors for the electron reconstruction [165], E_T-acceptance, isolation and identification [164].

4.2.2 A cross check for electron isolation and identification scale factors

To measure the electron efficiency at the level of isolation and identification for the $t\bar{t}$ analysis, one needs to make sure that all the electron selection requirements in the $t\bar{t}$ analysis are consistent with those applied in the Tag&Probe method on the $Z \rightarrow ee$ events. According to the top-quark analysis group recommendation for the event selection presented in Section 5.6.1, the electron passes the following set of cuts before the identification and isolation:

- $p_T > 30\,\text{GeV}$ and $|\eta|$ <2.5 while the pseudorapidity of the electron supercluster position is out of the EB-EE transition region, $|\eta^{sc}| \leq 1.4442$ or $|\eta^{sc}| \geq 1.566$;
- Small z-distance between the primary vertex and the electron position in the inner tracker, $|z_e - z_{pv}| < 1\,\text{cm}$;
- Reasonable transverse impact parameter with respect to the average beam spot, $d_0(b.s.) < 200\,\mu m$.

Hence in the following cross check, both the *tag* and the *probe* candidates are required to pass the same cuts. The *tag* candidate is further requested to meet both the isolation and the identification criteria: identified as an electron by identification requirements at the working point with 70% efficiency [5] together with the $tag_{relIso} < 0.1$. The *tag-probe* pair needs to have an invariant mass close to the mass of Z-boson, 76 GeV< $M_{tag,probe}$ <106 GeV. Events with more than two pairs are rejected. In addition, if a single *tag* electron makes pairs with two different *probe* candidates, the whole event is discarded since rejecting one pair may bias the efficiency.
The efficiencies are computed according to Equation 4.7 in bins of the *probe* electron kinematic variables. Having the values both in data and simulation, the scale factors are calculated at the end.
The study on the simulated samples, presented in [164], has shown that the above definition of the *tag* candidate reduces the background contamination under the Z-mass

[5] The working point efficiencies are derived from the cut optimization in the $W \rightarrow e\nu$ process. Hence, both the event environment in general and the electron selection in particular are different from $t\bar{t}$ and one would not expect to get the same efficiency of 70% in the $t\bar{t}$ analysis.

peak by a large factor. However due to the uncertainty arising from the difficulties in the modeling of the QCD multi-jet events as well as the limited size of the simulated QCD samples, further background subtraction can lead to more robust results. In the worst case scenario where the uncertainty on the QCD multi-jets contamination is taken 100%, the change in the efficiency would be a bit more than ∼1% while applying a side band subtraction method to reject backgrounds reduces the fluctuation down to ≤0.1%.
The application of the side band subtraction method needs more care since the desired mass window around the Z-mass peak is not wide enough to cover the side band regions. Therefore, the analysis is performed twice: once with the invariant mass interval extended in both sides, 50 GeV< $M_{tag,probe}$ <130 GeV, and once more with the desired invariant mass requirement. In the first round, (50 GeV,76 GeV) and (106 GeV,130 GeV)

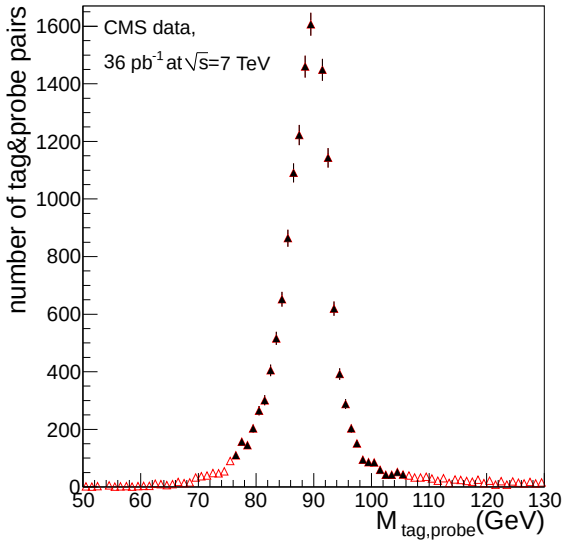
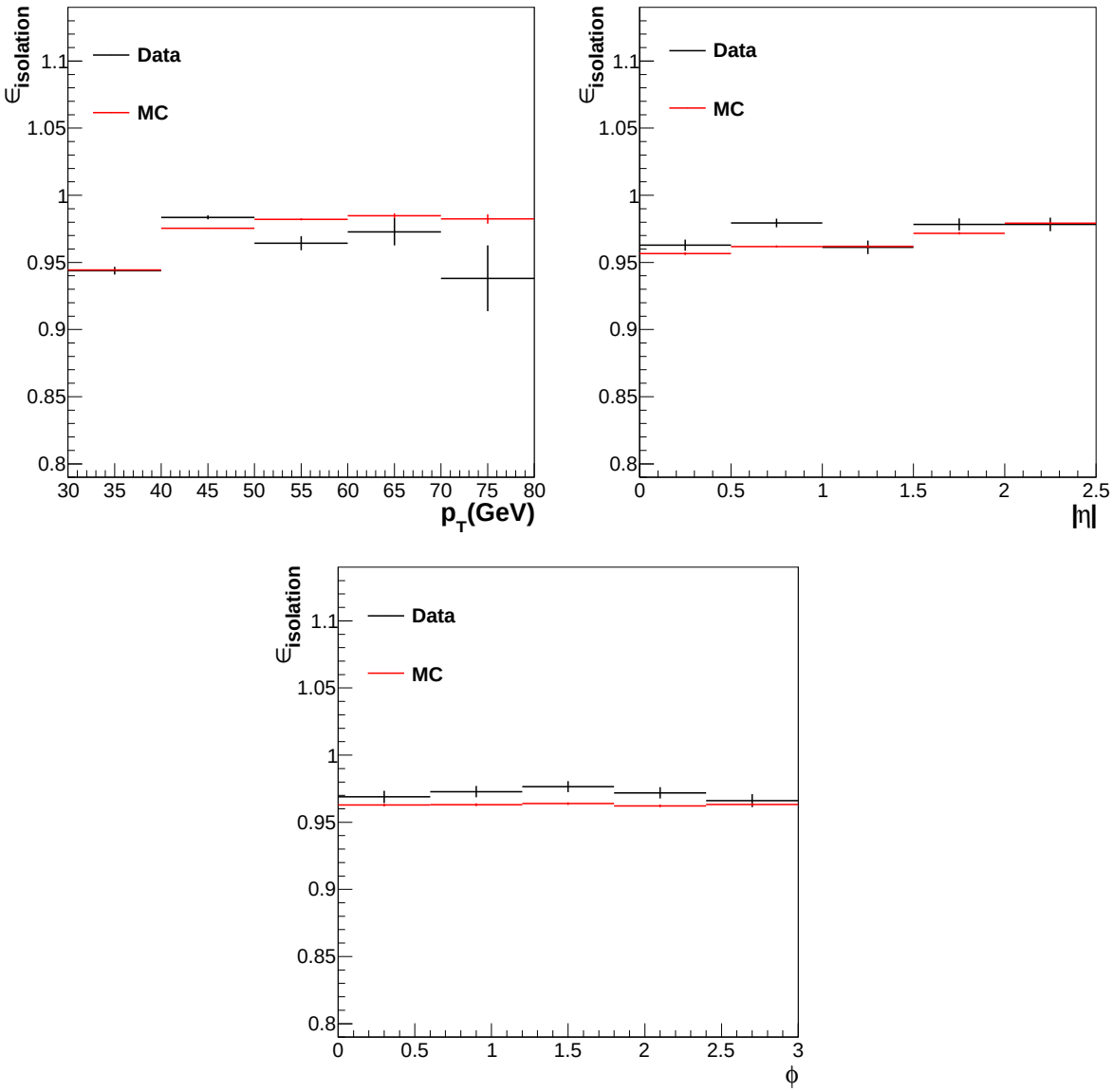


Figure 4.9: The *tag* and *probe* invariant mass spectrum. Side band regions are shown in red.

ranges have been taken as the side band regions in which the number of background electrons is estimated. Figure 4.9 illustrates the side band intervals on each side of the signal region.
The estimated number of background entries in each side band is assigned to the mean mass value of that area, hence one point in each band is provided. The shape of the background entries is estimated by a line connecting this two points. The integral of this line over the signal range is subtracted from the number of entries under the Z-mass peak. The resulting numbers are given to Equation 4.7 for the efficiency calculation.
The background subtraction is performed in bins of the *probe* electron kinematic variables to finally give a differential efficiency and scale factor. The Z-events have limited statistics in high jet bins so the values are calculated inclusively in terms of the number of jets. The method is applied on the full set of 2010 collision data collected by the CMS experiment. On the simulation side, a Drell-Yan sample with additional jets (so-called Z+jets sample) generated by `MadGraph` is used (see Section 3.4 for more explanation). For the coherence of the method the side band subtraction is applied on

the simulated sample as well however, it has no significant effect as expected.

Figures 4.10 and 4.11 show respectively the isolation and identification efficiencies in

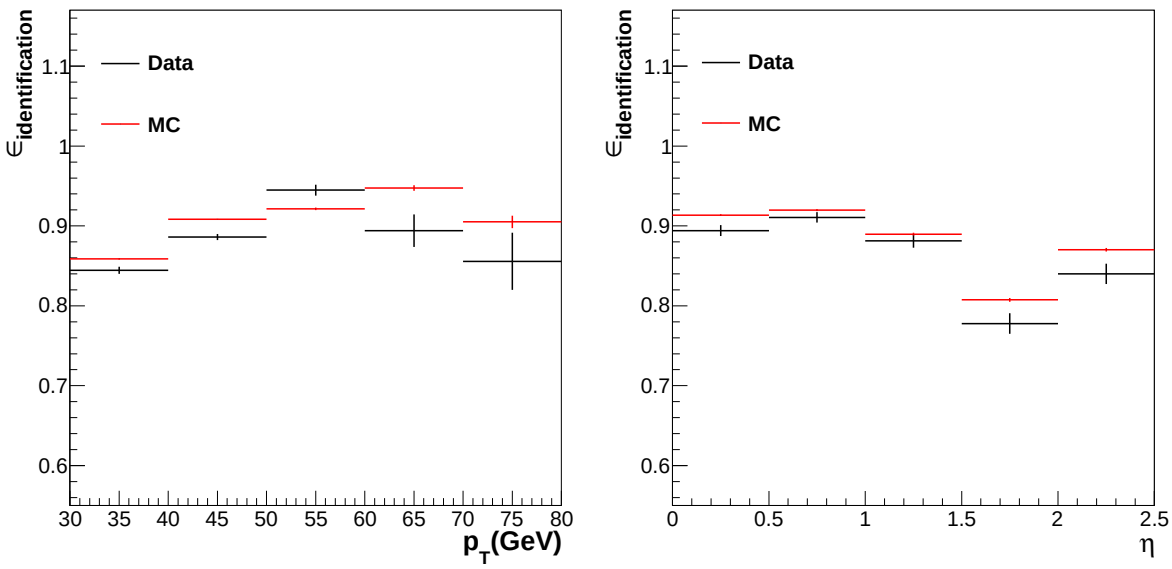


(c)

Figure 4.10: The differential isolation efficiency for the electrons passing the same electron requirements as in the $t\bar{t}$ event selection. The *Tag&Probe* method is applied on the 36 pb^{-1} CMS data in 2010.

different bins of η, ϕ and p_T of the *probe* electron. To calculate the scale factor according to Equation 4.8, efficiency distributions in ϕ from data and simulation are divided and fitted with a straight line. The scale factors versus ϕ are plotted in Figure 4.12.

Possible sources of systematic uncertainties to this method are the shape of the background and the width of the side band region. Trying different shapes for backgrounds,



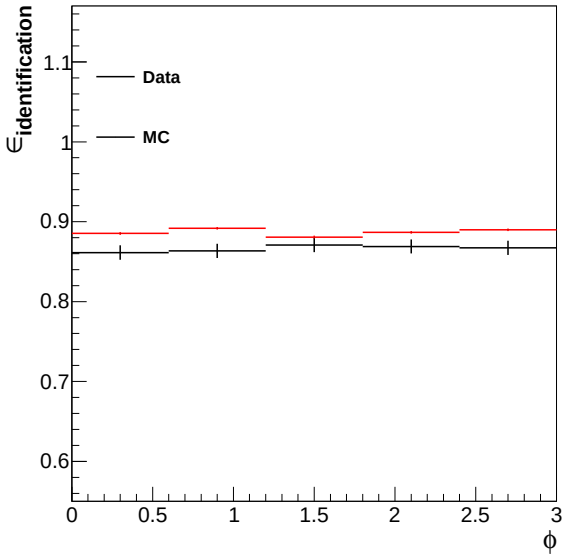


Figure 4.11: The differential identification efficiency for the electrons passing the same electron requirements as in the $t\bar{t}$ event selection. The *Tag&Probe* method is applied on the $36\,pb^{-1}$ CMS data in 2010.

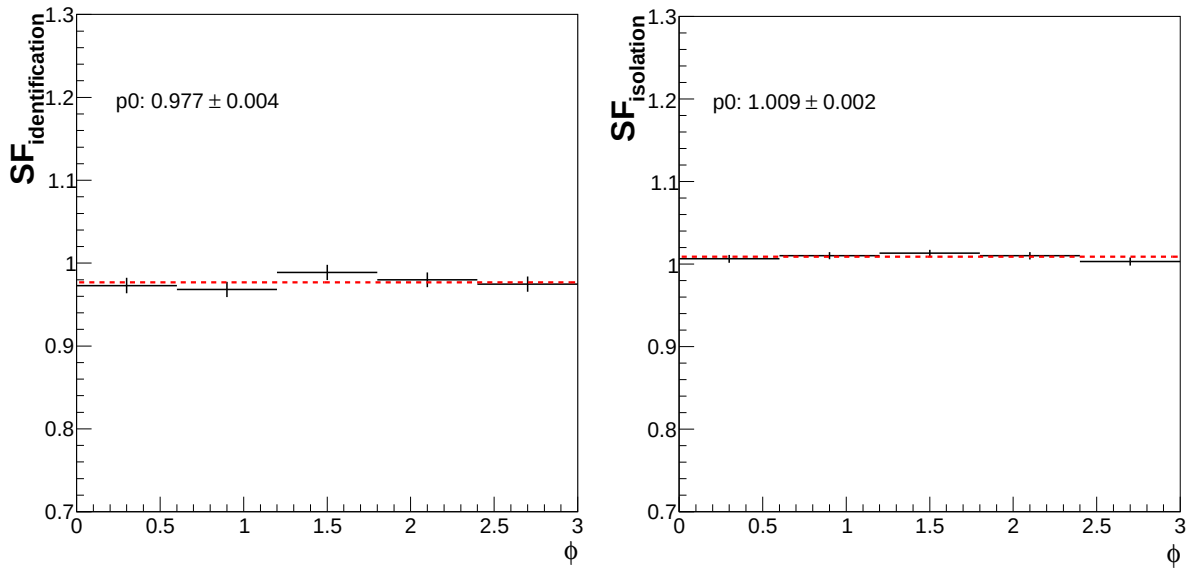


Figure 4.12: The data over simulation scale factor for the electron identification (left) and isolation (right) efficiencies. The electrons pass the same electron requirements as in the $t\bar{t}$ event selection. The *Tag&Probe* method is applied on the $36\,pb^{-1}$ CMS data in 2010 as well as the Z-boson plus jets simulated sample.

	scale factor
identification	0.98 ± 0.02
isolation	1.009 ± 0.007

Table 4.3: The scale factors for the electron isolation and identification extracted using a side band subtraction method to suppress the background electrons under the Z-mass peak.

the uncertainty on the scale factor is estimated to be ∼0.003 (0.004) for SF_{iso} (SF_{id}). An uncertainty of $\Delta SF \approx 0.006$ (0.02) for isolation (identification) arises from the uncertainty on the width of the side band range. This uncertainty has been treated more carefully since in the simulated sample a cut of 50 GeV has been applied on the invariant mass at the generator level. Table 4.3 summarizes the scale factors and their uncertainties for the electron isolation and identification. The values are in good agreement with the results in [164] where another method is used to suppress the background pairs.

4.2.3 The additional source of systematic uncertainty

In the assumption leading to Equation 4.11, the electrons are considered as individual objects interacting with the detector material. Taking into account the whole event activities, i.e. all of the particles interacting with the detector material at the same time, that is much more in the $t\bar{t} \to e\nu b\bar{b}qq'$ events than the $Z \to ee$ processes, one would conclude that Equation 4.11 is only true at the first orders.
On the other hand it is not straight forward to find the exact relation between the $t\bar{t}$ and $Z \to ee$ scale factors since with the relatively small $t\bar{t}$ cross section a direct measurement of the efficiency for top leptons can hardly be performed in the current dataset. Hence a systematic uncertainty is assessed to cover the reasonable difference between $\frac{\epsilon^{Z}_{data}}{\epsilon^{Z}_{MC}}$ and $\frac{\epsilon^{t\bar{t}}_{data}}{\epsilon^{t\bar{t}}_{MC}}$. This systematic uncertainty is uncorrelated to those coming from the background suppression method so needs to be finally added to the total uncertainty in quadrature.
The difference in isolation and identification efficiencies between $Z \to ee$ and $t\bar{t}$ events is investigated using the simulated event samples (see Section 3.4). While on the "Z-side" the *Tag&Probe* method is applied[6], on the side of top quark events the electron candidates are matched to the true electrons from the W-boson decay at generator level. The $t\bar{t}$ simulated events are already filtered to the semi-electron final state using the generator level information.
The electron candidate in the $t\bar{t}$ event passes the same criteria as the *probe* candidate on "Z-side": $p_T > 30\,\text{GeV}$, $|\eta| < 2.5$ (EE-EB transition region excluded), $|z_e - z_{pv}| < 1\,\text{cm}$ and $d_0(b.s.) < 200\,\mu m$. As before, the *tag* electron fulfills in addition the isolation and identification conditions.
In Figure 4.13, the p_T and η distributions of the electron in $t\bar{t}$ event are compared to those of the *probe* candidate in the Z event. Electrons seem more boosted in the $t\bar{t}$ event while the *probe* candidates in Z-events are more central.
Concerning the jet activity, Z-events have lower jet multiplicities and the jets are well separated from the *probe* candidate. The *tag* and *probe* electrons are also produced back to back (Figure 4.14). To avoid those electrons which are mistaken as jets, as it is discussed in Section 4.3, jets closer than $\Delta R = 0.3$ to an electron candidate in $t\bar{t}$ and in the *tag* or *probe* candidates in the Z-event are removed from the jet collection.
The inclusive estimation of the electron identification efficiency, ϵ_{id}, is about 88.6% in top-quark events. In $Z \to ee$ events the *Tag&Probe* method results in an identification efficiency of $\epsilon_{id} \approx 89.5\%$. In Figure 4.15, $\epsilon^{t\bar{t}}_{\text{id}}$ and ϵ^{Z}_{id} are plotted versus the kinematic

[6] It has been checked that the true electrons under the Z-boson mass peak give the same efficiency.

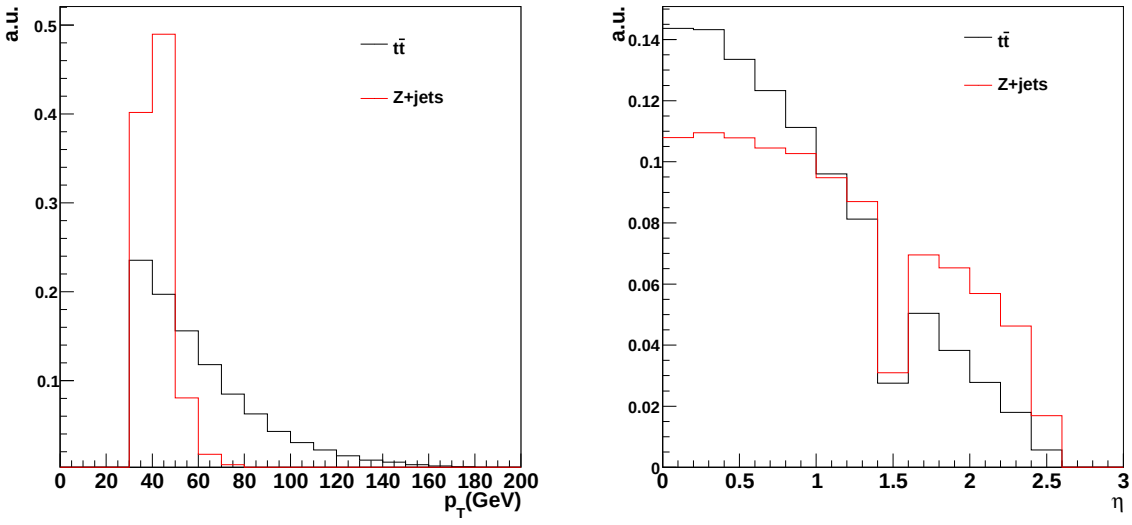
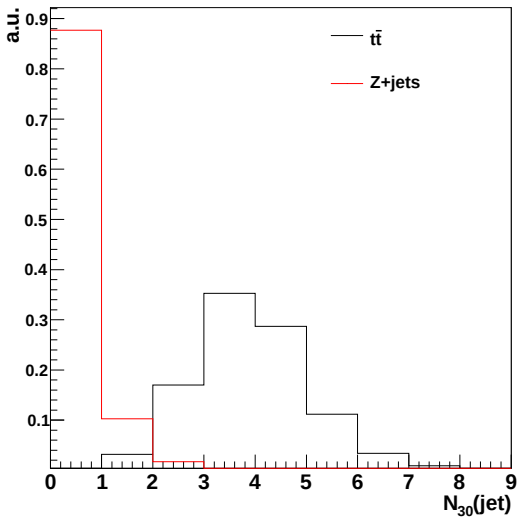


Figure 4.13: The p_T (left) and $|\eta|$ (right) distributions for the electrons in $Z \rightarrow ee$ and semi-electron final state of the $t\bar{t}$ events.

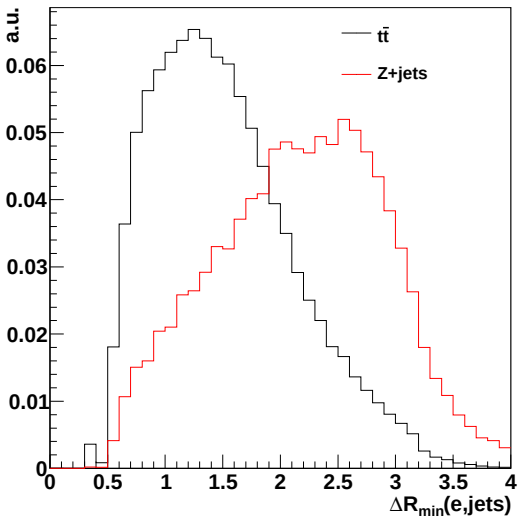
variables of the electron. To have a closer look, the electron ID variables in Z- and top-quark events are compared in Figure 4.16. Except the H/E quantity which looks slightly different, the electron ID variables are quiet similar in the both kinds of events. For the isolation efficiency, ϵ_{iso}, a difference of $\sim$ 6% is observed ($\epsilon_{\mathrm{id}}^{\mathrm{t\bar{t}}} \approx 89.9\%$ and $\epsilon_{\mathrm{id}}^{\mathrm{Z}} \approx 96.0\%$).

Figure 4.17 illustrates the $\epsilon_{\mathrm{iso}}^{\mathrm{Z(t\bar{t})}}$ versus the kinematic variables of the electron. It can be seen that the difference between the efficiencies is larger in the barrel with respect to the endcap. It also decreases at higher p_T's. The p_T dependence of ϵ_{iso} is expected from the definition of the isolation variable in Equation 4.1. Unlike the situation on the Z-side, the inclusive isolation efficiency in $t\bar{t}$ events is driven by the events with multiple jets in the final state. Despite of the limited statistics for the $Z \rightarrow ee$ process at higher jet bins, a decreasing trend as a function of the jet multiplicity can be recognized in Figure 4.18 a. In addition to the number of jets in general, what can influence the isolation more specifically is the electron-jet separation. This variable is quiet different in the $t\bar{t}$ and Z-events (Figure 4.14). The isolation efficiency is plotted with respected to the $\Delta R_{min}(e, jets)$ in Figure 4.18 b. The low statistics in small ΔR_{min}'s makes it difficult to judge the value of $\epsilon_{\mathrm{iso}}^{\mathrm{Z}}$. However for the same reason, the higher isolation efficiency in $Z \rightarrow ee$ is understandable. The isolation efficiency becomes flat at large ΔR_{min}'s which is expected. One can also look at the electron isolation quantities in different subdetectors for the top-quark and Z-events as in Figure 4.19. These variables can give the information about those energy deposits around the electron which were not clustered by the jet algorithms but were high enough to be included in the isolation quantities. The isolation quantities for the electron in $t\bar{t}$ have longer tails in all subdetectors.

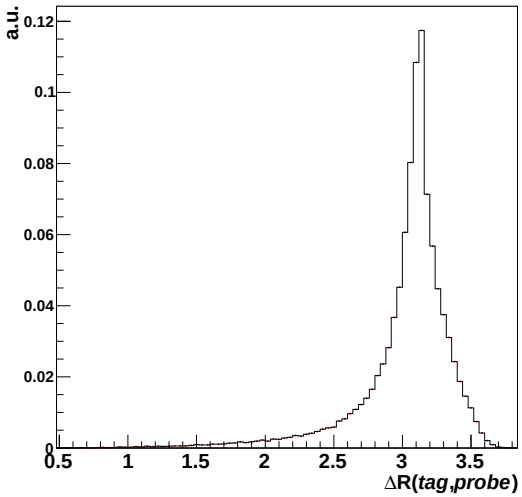
The knowledge about $\Delta\epsilon(Z, t\bar{t})$ needs to be included in the systematic uncertainty. The



(a)



(b)



(c)

Figure 4.14: The jet multiplicity (a) and the minimum $\Delta R(electron, jets)$ (b) are compared in $Z \rightarrow ee$ and semi-electron final state of the $t\bar{t}$ events. The *tag* and *probe* separation in Z-events is illustrated in (c).

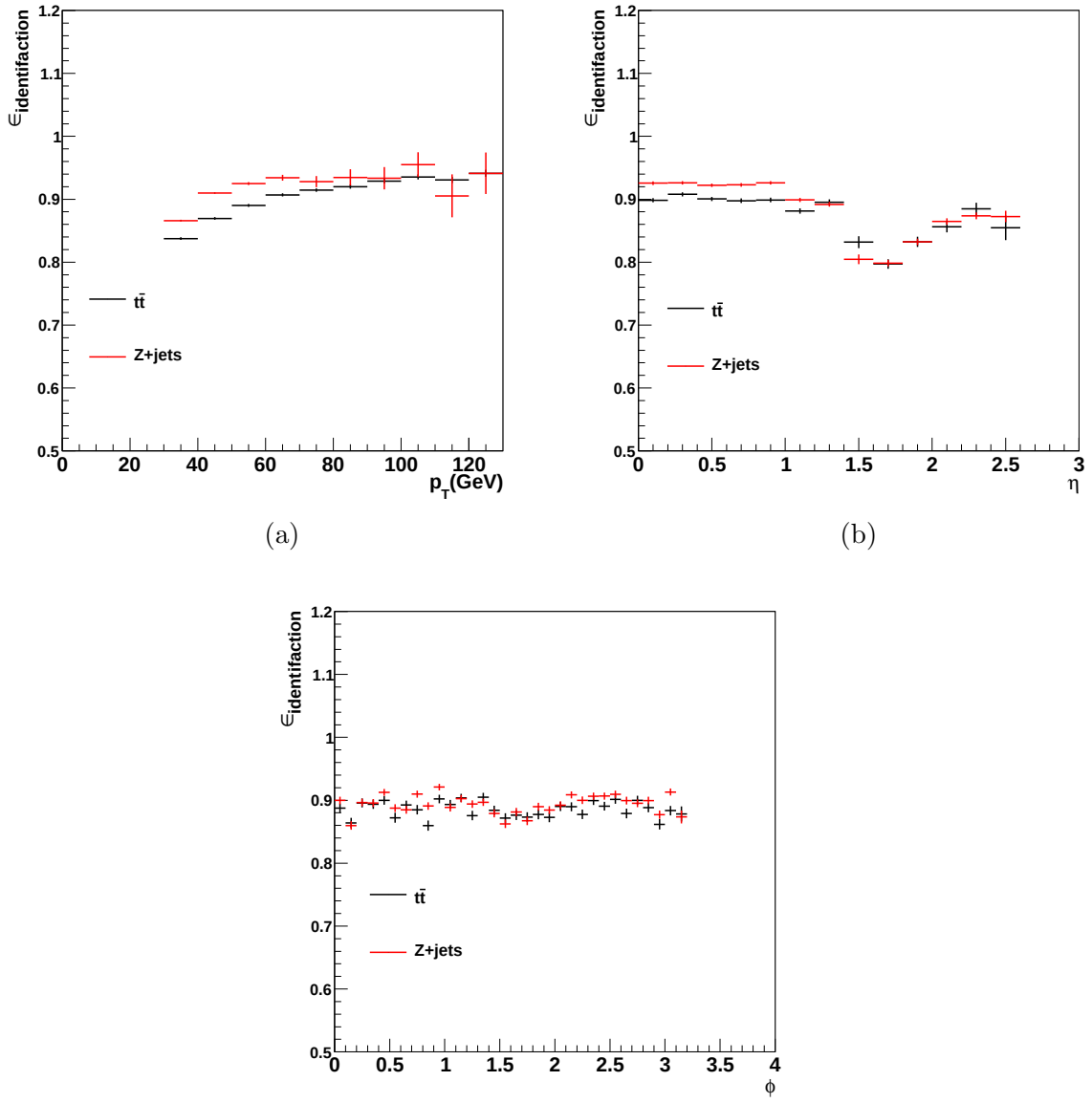


Figure 4.15: The differential electron identification efficiency in $Z \rightarrow ee$ and semi-electron final state of the $t\bar{t}$ events.

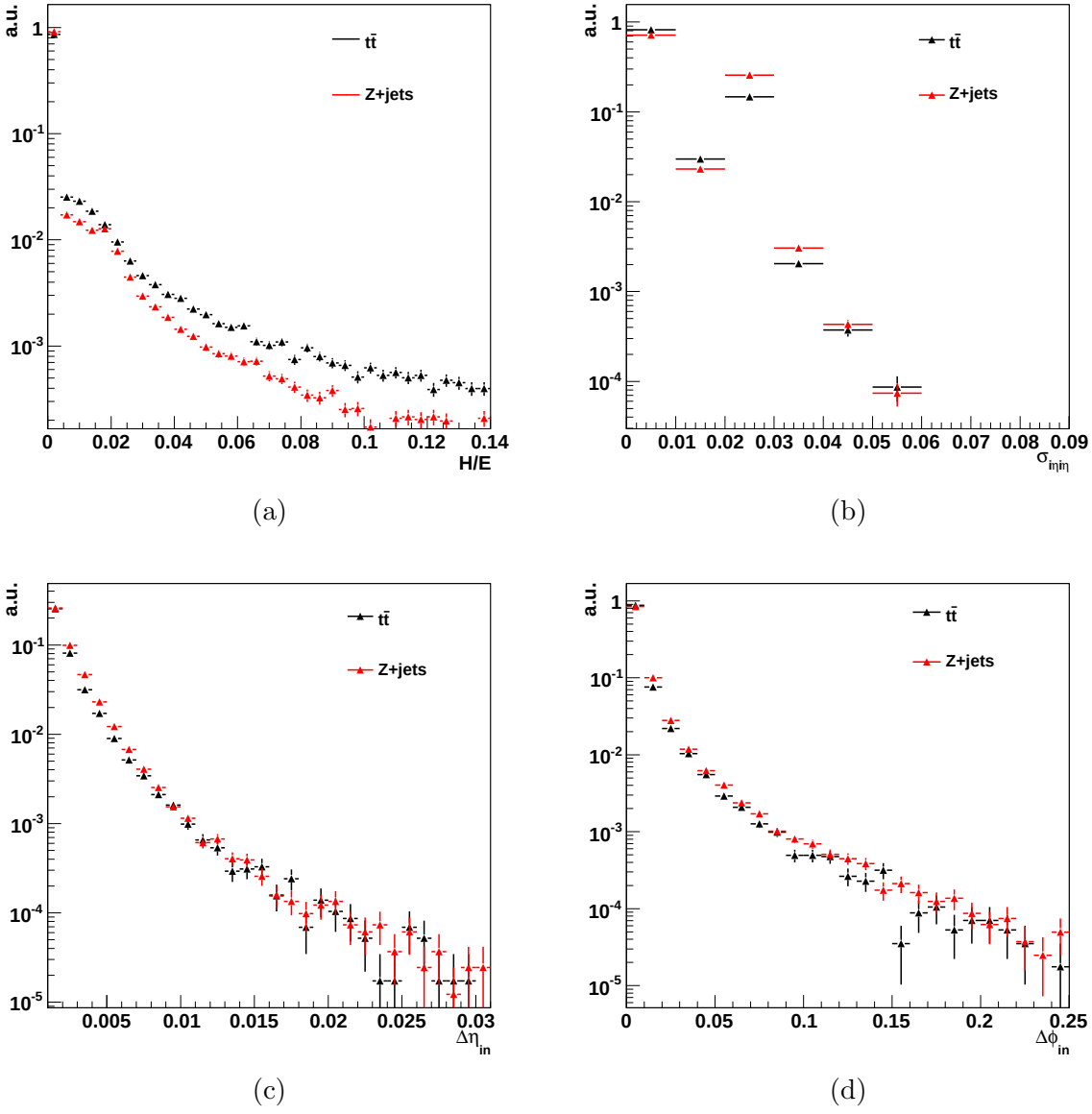
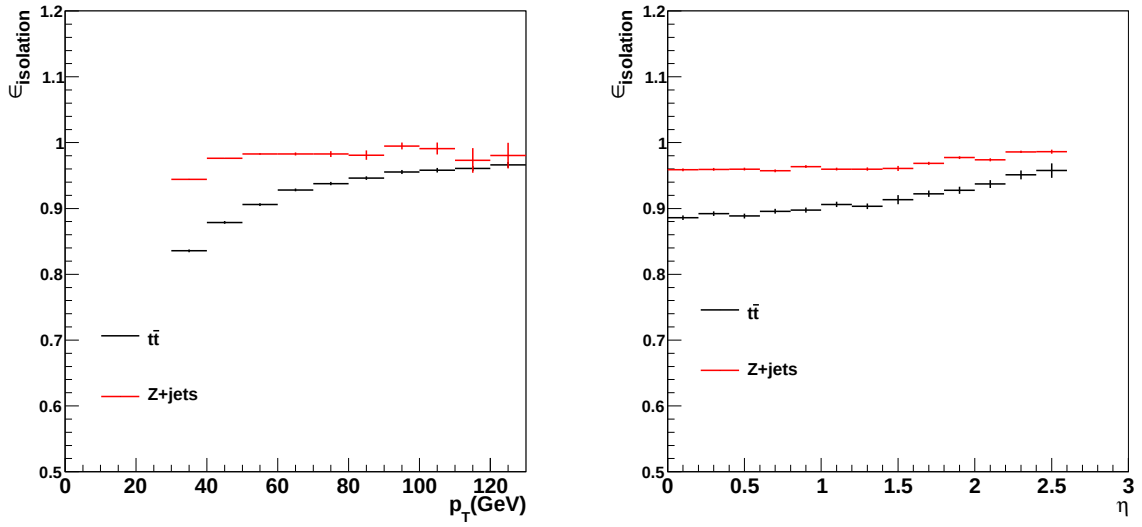
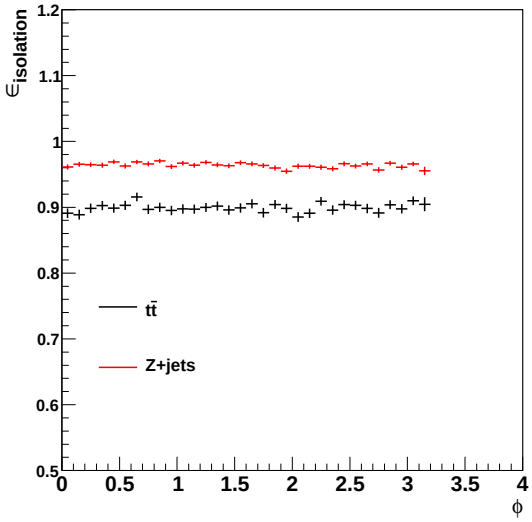


Figure 4.16: The electron identification variables in $Z \rightarrow ee$ and semi-electron final state of the $t\bar{t}$ events.



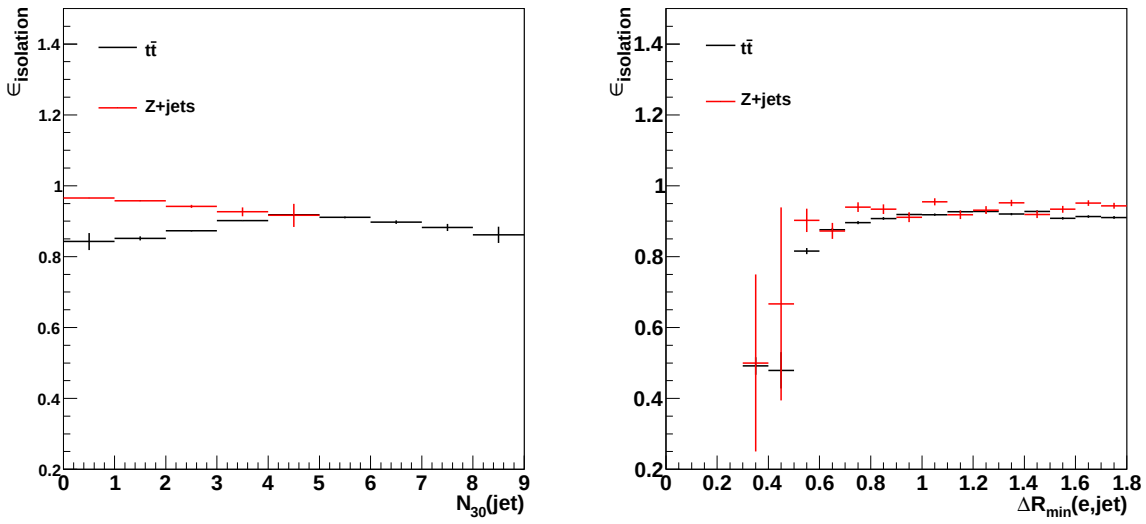
(a)

(b)



(c)

Figure 4.17: The differential electron isolation efficiency in $Z \rightarrow ee$ and semi-electron final state of the $t\bar{t}$ events.



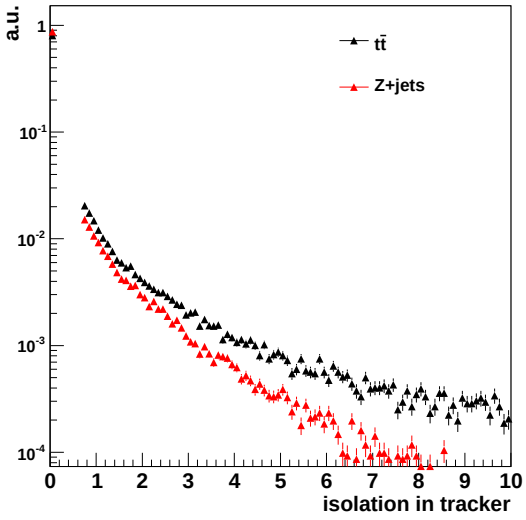
(a) (b)

Figure 4.18: The electron isolation efficiency in $Z \rightarrow ee$ and semi-electron final state of the $t\bar{t}$ events in the bins of the jet multiplicity (a) and the minimum distance between the jets and the electron (b).

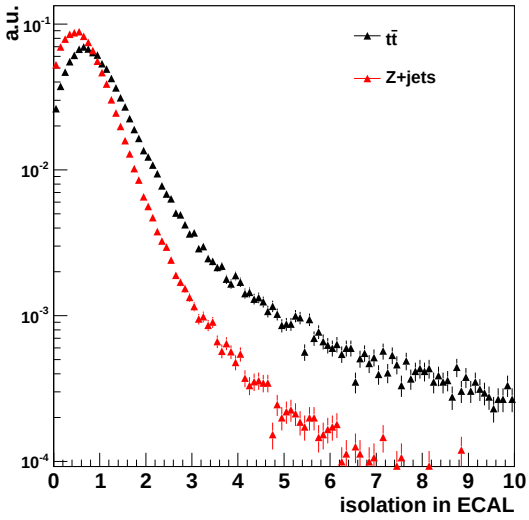
difference in the identification efficiencies, $\Delta\epsilon_{\mathrm{id}}^{\mathrm{Z,t\bar{t}}}$, is small enough to be neglected. For the case of isolation, a difference of $\Delta\epsilon_{\mathrm{iso}}^{\mathrm{Z,t\bar{t}}} \approx 6\%$ is observed. This information [164] is used in [25, 163] to incorporate the systematic uncertainty arising from different event topologies and properties in $t\bar{t}$ and $Z \rightarrow ee$ processes.

4.3 Jet reconstruction

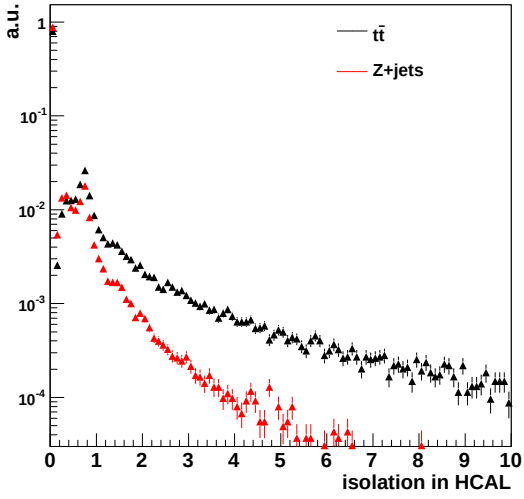
The jet reconstruction in CMS is of great importance since almost every physics process at the LHC contains jets of charged and neutral particles in the final state. The information from different parts of the detector can be combined to serve as input for the jet reconstruction algorithms. While the ECAL and HCAL energy deposits in the form of CaloTowers (see Section 2.2.2) are used in the formation of **calorimeter jets**, the well measured tracks (see Section 2.2.1) are the building elements of the **track jets** [166]. The **jet-plus-track**'s [167] exist in between, exploiting the excellence of the tracking system to improve the p_T resolution and response of the calorimeter jets. Finally to reconstruct the jets in the context of the particle flow event reconstruction [152], the information from all CMS subdetectors results in the formation of the whole particle content of the event. The charged and neutral particles are grouped into the **particle flow jets**[168] using the dedicated algorithms.
Independent from the input, the outcome of the jet algorithms is expected to remain unchanged if for example the energy carried by a single particle is split between two collinear particles (collinear-safe requirement). Moreover, adding soft particles should



(a)



(b)



(c)

Figure 4.19: The tracker (a), ECAL (b) and HCAL (c) isolation quantities for the electron in $Z \rightarrow ee$ and the semi-electron final state of $t\bar{t}$ events.

not spoil the stability of the jet finding results (infrared-safe).
Having evolved in time, currently in CMS the *Seedless Infrared-Safe cone* (SISCone) algorithm [169] with the opening angle 0.5 (0.7), the **Fast** κ_T algorithm [170] with the recombination parameter 0.4 (0.6) and the **Anti-κ_T** algorithm [171] with the recombination parameter 0.5 (0.7) are used. The performance of different jet algorithms are studied and compared in CMS for the CPU time usage and the optimal values for their parameters are also looked for [172]. The Anti-κ_T algorithm with the recombination parameter 0.5 is the most common method in the experiment and it is used to reconstruct the calorimeter jets for the analysis presented in this thesis. Hence the algorithm is briefly reviewed.

4.3.1 The Anti-κ_T jet algorithm

The algorithm is an extension to the κ_T jet formation method, assigning a list of energy dependent distances to each entity that here is a CaloTower

$$d_{iB} = \frac{1}{k_{\mathrm{Ti}}^2}, \tag{4.12}$$

$$d_{ij} = min(\frac{1}{k_{\mathrm{Ti}}^2}, \frac{1}{k_{\mathrm{Tj}}^2})\frac{\Delta_{\mathrm{ij}}^2}{R^2}. \tag{4.13}$$

In Equation 4.12, d_{iB} is defined between the entity i and the direction of the colliding particles while d_{ij} in Equation 4.13 is between the constituents i and j. The variable k_{Ti} is the transverse energy of the entity i and the quantity Δ_{ij} is the i-j distance in the y-ϕ plane

$$\Delta_{\mathrm{ij}}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2. \tag{4.14}$$

The recombination parameter R is taken 0.5 in the current discussion. The algorithms proceeds by looking for the smallest 'd' and if it is a d_{ij}, combining the entities i and j. For those steps where d_{iB} is the smallest, the entity i is considered as a jet and is removed from the list. The procedure continues until no entity is left. The algorithm is infrared-safe since the soft constituents tend to combine with the harder ones, i.e. if $k_{\mathrm{Ti}} \gg k_{\mathrm{Tj}}$, the variable d_{ij} is governed by k_{Ti} and the entity j is absorbed by i. The algorithm is also collinear-safe because for the collinear entities with similar energies, the smallness of Δ_{ij} helps the d_{ij} to be small enough to combine the two constituents. Besides the jet energy, an important property for a reconstructed jet is its direction which plays a significant role in the b-jet identification methods. Following the energy recombination scheme, E-scheme, the 4-momentum of the participating CaloTowers are added to find the jet 4-momentum. This results in massive jets. In the E_T-scheme massless jets are produced by equating the p_{T} of the jet to the sum of the E_{T}'s of the CaloTowers. The jet η (ϕ) position is the energy weighted sum of $\eta^{CaloTower}$ ($\phi^{CaloTower}$) [173]. The energy scheme has been used for the 2010 CMS data and simulated analyses [174] and hence in this thesis.

4.3.2 Jet energy corrections and resolutions

Although the CaloTowers are already cleaned from noises and passed the selection scheme in Table 2.1, the resulting jets may be different from the real collimated group of particles that has to be described. The jet energy can be affected by extra interactions either from pile-up or underlying events. In addition due to the complex shape and composition of the detector, its response to similar jets is not uniform over the whole η range. Moreover, the p_T of the jet, its flavor and other properties like the electromagnetic energy fraction can influence the measured energy by the detector.
The jet energy correction in CMS is factorized into seven sequential levels:

Offset which corrects the energy for pile-up and possible electronic noise (L1).

Relative η which makes the response uniform in pseudorapidity relative to a control region, $|\eta| < 1.3$ (L2).

Absolute $\mathbf{p_T}$ that is extracted in bins of jet p_T for the jets in the $|\eta| < 1.3$ region. Since it is applied after the relative (η) correction, the response for the control region can be generally used in the whole η range (L3).

EMF that corrects the variation in jet response with the electromagnetic energy fraction (L4).

Flavor correction which is intended to correct the jets regarding to their flavor. The detector response is higher for e.g. quark jets than for gluon jets since they fragment into higher momentum particles. For the specific case of b-flavor jets, data-driven methods are developed to extract the correction factors within the $t\bar{t}$ events (L5).

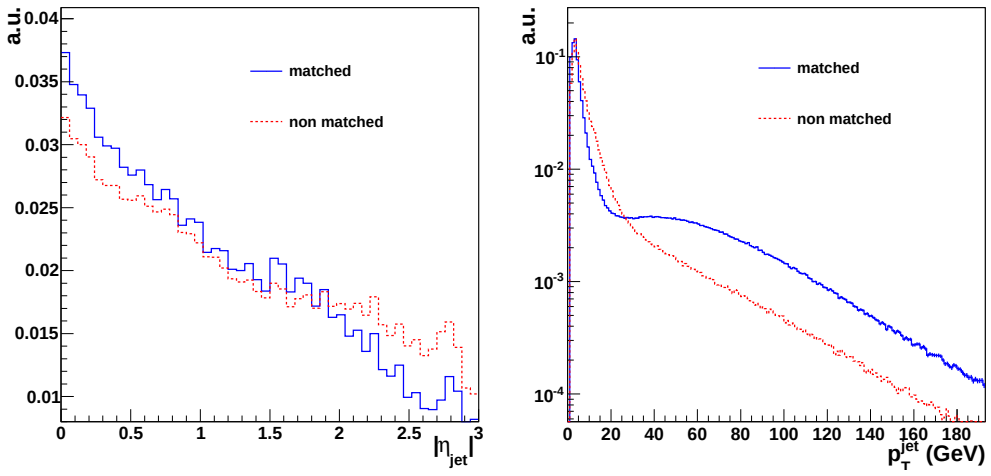
Underlying event which corrects for the energy of the underlying events that contributes in the jet energy estimation (L6).

Parton level correction that is based on the comparison with the partons energy using the information at the generator level (L7).

Among all steps, only the first three corrections are mandatory and can be extracted either from simulation or from data. The corrected jet energy is obtained as

$$E_{corr} = (E_{raw} - offset) \times CF(rel : \eta) \times CF(abs : p_T). \tag{4.15}$$

The jet momentum resolution which is an indicator for the precision of the jet p_T measurement is another important subject in the jet studies in CMS. For different types of jets, both the jet energy correction factors and the momentum resolution are studied within the 2010 dataset of pp collisions [174] using data-driven techniques. It is shown for all jet types that the total uncertainty on the jet energy scale is constrained to 3% and smaller for $p_T > 50\,\text{GeV}$ and $|\eta| < 3.0$. Based on the studies in [174], for the analysis presented in this thesis the relative (η) and absolute (p_T) corrections are applied on both data and simulation while the offset correction seems to be necessary for the jets in real pp collisions. The effect of the jet energy scale systematic uncertainty



(a) (b)

Figure 4.20: The jet $|\eta|$, (a), and corrected p_T, (b), distributions for the jets in the semi-electron final state of $t\bar{t}$ event. The "(non) matched" category containes the jets (not) matching to the quarks generated in decay products of the event.

on the final results is also investigated where a rather conservative variation of 10% is considered.

Figure 4.20 shows the $|\eta|$ distribution as well as the distribution of the corrected transverse momentum for the jets in the semi-electron final state of $t\bar{t}$ events where the jets are divided into "matched" and "non matched" categories. The "matched" category containes the jets that are matched with the quarks from the semi-electron final state of $t\bar{t}$ better than $\Delta R = 0.3$ in the $\eta - \phi$ plane. The jets failing this requirement are grouped into the "non matched" category. The matched jets are found to be slightly more central and have a larger transverse momentum.

Beside the good energy estimation, it is also necessary to estimate the jet direction properly. The space resolution of the jets plays an important role in for example the di-jet mass analyses and in particular in the jet flavor identification algorithms. For a given property, s_{jet}, the resolution can in general be defined as the width of

$$\frac{s_{jet} - s_{gen}}{s_{gen}}, \tag{4.16}$$

where s_{gen} is the same property of the object from the generator level which is associated to the reconstructed jet by the ΔR matching. The reconstructed jets can be matched either to partons at the parton level or to *GenJets*.

The decay products of the proton collisions become stable after hadronization. These stable particles are used as inputs to the jet algorithm for a *GenJet* before they interact with the detector material or bend in the magnetic field. The GenJets are built with the same reconstruction algorithms as the simulated jets. Hence a comparison between the GenJets and the simulated jets would result in a better understanding of the magnetic field and the detector effects on the jets of particles.

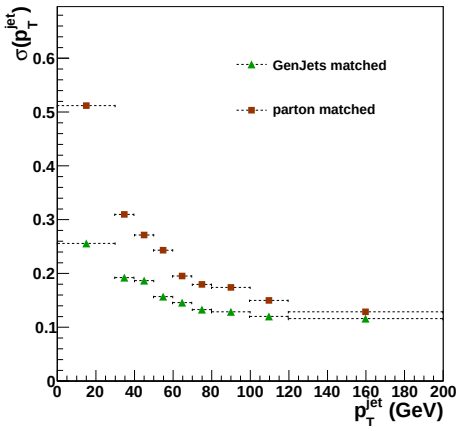


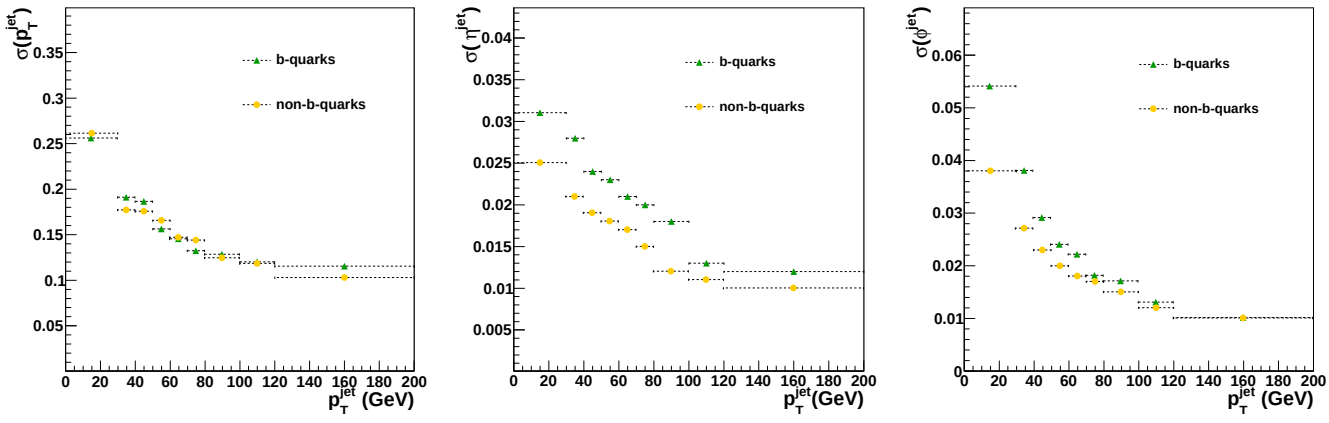
Figure 4.21: The p_T resolution of the jets originated from b-quarks in the semi-electron final state of $t\bar{t}$ events. The deviation from the p_T of the original quark is compared with the deviation from the p_T of the associated GenJet. To obtain the resolution in each p_T bin, the distribution of the relative p_T difference as in Equation 4.16 is fitted with a Gaussian and the width is taken as the resolution.

Figure 4.21 illustrates the p_T resolution of the jets originated from b-quarks in the semi-electron final state of $t\bar{t}$ events. A better resolution is achieved when the jet transverse momentum is compared with the p_T of the associated GenJet. For the b-flavored jets, a fraction of energy may be carried by the neutrino which is produced in the leptonic decay of B-mesons. This energy is lost when the energy of the simulated jet and the energy of the GenJet is calculated while it is present in the total energy of the original parton. In the calorimeter jets, the same happens for the particles interacting not so strongly with the calorimeters such as muon. Such effects influence not only the energy but also the direction of the simulated jet. Therefore, to study the jets momentum and space resolutions, GenJets are used.

Figure 4.22 shows the resolution of the p_T, η and ϕ variables of the jets in the semi-electron final state of $t\bar{t}$ events where the resolutions for b-flavored and non-b-flavored jets are plotted separately. The distribution explained by Equation 4.16 is fitted with a Gaussian and the width is taken as the resolution in each plot. It can be seen that the jets with higher p_T are reconstructed with a better resolution. The reason is that the charged particles with lower momentum in the jet are bent more by the magnetic field. Hence, their energy may not be included in the jet total energy calculation. The difference between the b- and non-b-flavored jets is not so significant for the p_T resolution while it is considerable for the angular resolutions specially at lower p_T's.

4.3.3 Jet identification variables for $t\bar{t}$ analyses

In addition to the usual electronic noise, there are other sources of unphysical energies that might appear in the CMS calorimeter, including the occasional malfunctions of the detector electronics. A wide set of jet identification variables are studied in CMS to reject these fakes [175]. These identification criteria are applied on top of the noise



(a) (b) (c)

Figure 4.22: The jet p_T-, η- and ϕ-resolutions (a,b,c respectively) of the four leading jets in the $t\bar{t}$ events.

rejection presented in Table 2.1 which is imposed on CaloTowers before the jet reconstruction. In the 2010 $t\bar{t}$ analyses, the following jet identification variables are used where the cuts are so that the resulting jet selection efficiency is close to one:

f_{em} or the electromagnetic energy fraction is specially important since if it is too low, meaning negligible amount of energy in ECAL, the measured energy is most probably coming from HCAL noise. Hence a lower threshold of $f_{em} > 0.01$ is applied. One can in addition ask for an upper bound since for $f_{em} \approx 1$, the object can be an electron faking a jet.

N90$_{hits}$ is the minimum number of calorimeter hits containing 90% of the jet energy. Physical jets tend to fire many hits so in $t\bar{t}$ analyses the jets are requested for $N90_{hits} > 1$.

f_{HPD} is the energy fraction belonging to the hottest HPD readout in the HCAL (see Section 2.2.2 for HPD definition). If this fraction is $f_{HPD} > 98\%$, it means the readout channels around the hottest one are not necessarily fired so the signal tends more to be a readout noise rather than a real jet.

N$_{CaloTowers}$ is the total number of CaloTowers assigned to the jet which is in particular larger for more energetic jets. The jets are required to have $N_{CaloTowers} > 5$ in the analysis developed in this thesis.

In Figure 4.23, the jet identification variables are plotted for the matched and non matched jets in semi-electron $t\bar{t}$ events where a clear difference between the matched and non matched jets is observed. The matched jets tend to spread their energy thus having higher $N90_{hits}$ and firing more CaloTowers. The electromagnetic energy fraction is high enough for the matched jets where they also fire more HCAL readout channels, leading to lower values for f_{HPT}.

Looking at the evolution of the identification variables as a function of the jet p_T

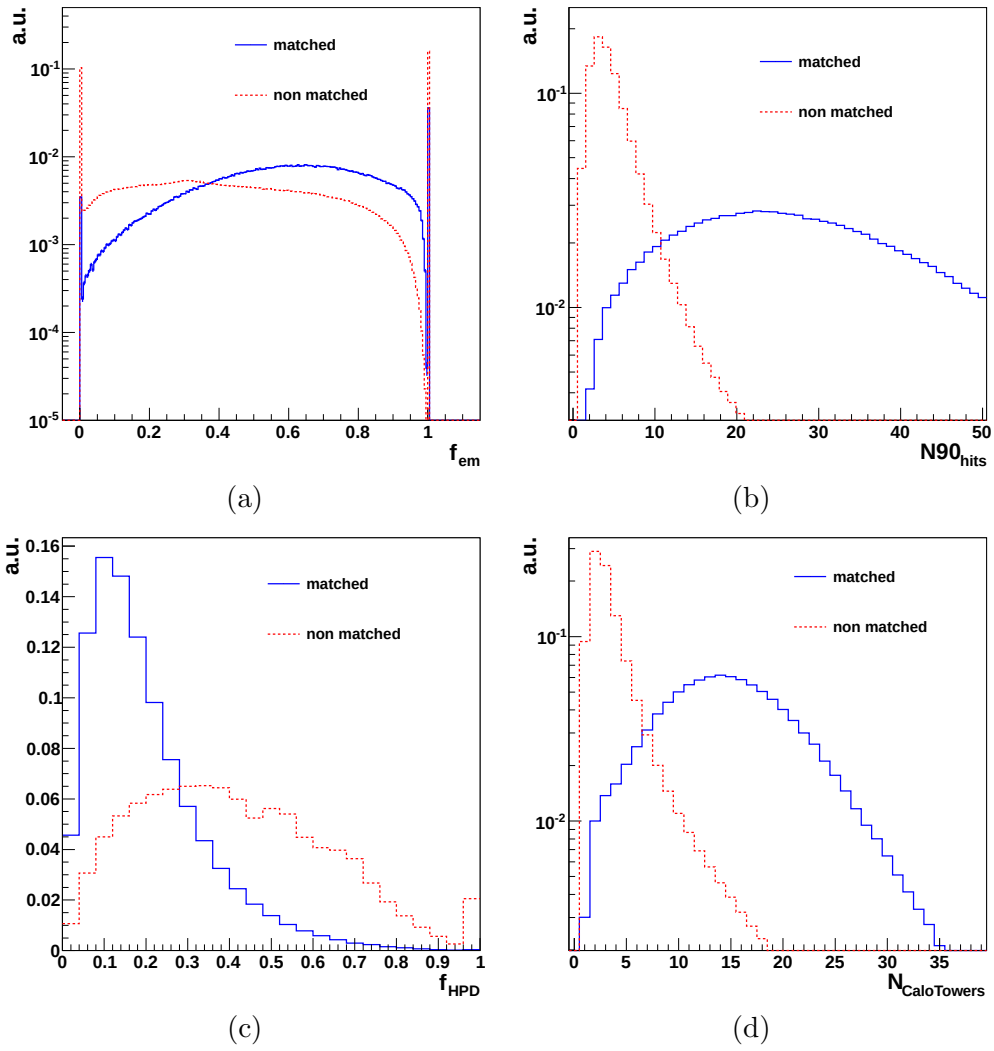
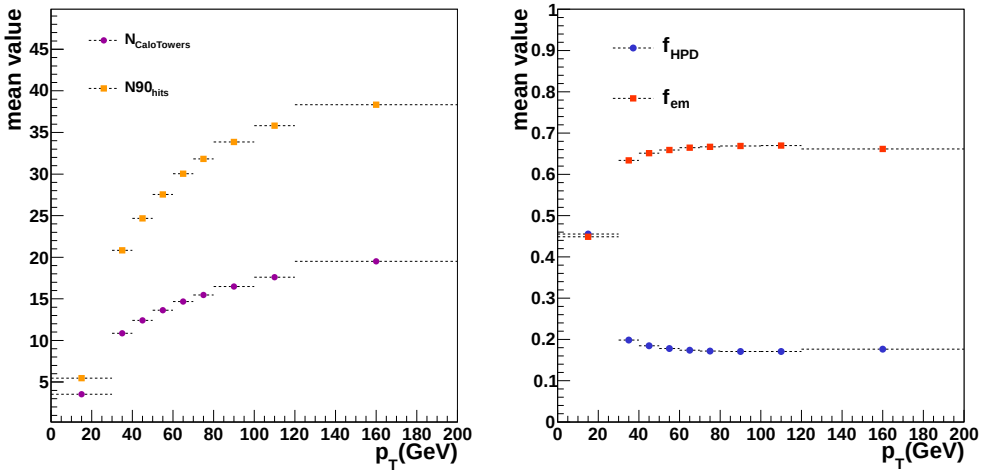


Figure 4.23: The distribution of the jet identification variables for the four leading jets in $t\bar{t}$ events: electromagnetic fraction (a), $N90_{hits}$ (b), f_{HPD} (c) and $N_{CaloTowers}$ (d).



(a) (b)

Figure 4.24: The mean value of $N_{CaloTowers}$ and $N90_{hits}$, (a), together with the mean value of f_{HPD} and f_{em}, (b), in the bins of the jet p_T for the jets in the semi-electron final state of $t\bar{t}$ events .

in Figure 4.24, one can deduce that a cut of $p_T > 30\,\text{GeV}$ can already reject the jets with poor identification variables in $t\bar{t}$ events. However, selecting the jets according to the identification variables is still favorable to keep the well defined jets as much as possible specially in the presence of the background processes other than $t\bar{t}$. The effect of the jet identification variables is studied in more detail in Chapter 5 from the event selection point of view.

4.4 b-jet identification algorithms

A considerable effort in the CMS experiment is devoted to the identification of b-flavored jet of hadrons, introduced in Chapter 3, since they are essential to characterize various Standard Model and New Physics channels, e.g [176, 177].
The b-tagging algorithms exploit the distinct properties of the decay of b-hadrons which are the product of b-quark fragmentation:

- The relatively long life-time of b-hadrons, $\tau \approx 1.5\,ps$ ($c\tau \approx 450\,\mu s$) [3], introduces a displaced vertex at the point of decay, *secondary vertex*, that can be reconstructed relying on the excellence of the CMS inner tracking system,

- The final state of the b-hadrons decay contains on average 5 tracks[7], all having a sizable impact parameter (IP) with respect to the primary vertex.
  The IP can be calculated either in the transverse plane or in 3D due to the good resolution in z, provided by the pixel tracker in CMS.

[7] An example is $\bar{B}_s \rightarrow D_s^+ l^- \nu_l \rightarrow \pi^+ K^+ K^- l^- \nu_l$, in which the B-meson decays via a charmed hadron and produces 5 charged particles.

- One of the charged tracks among the b-hadron decay products is a soft lepton in 19% of the time per lepton family if both direct and cascade decays are taken into account. Since the b-quark is much more massive than what it decays to, its decay products including the possible lepton have a larger transverse momentum with respect to the jet axis.

It is therefore obvious that tracks are the most powerful ingredients for b-tagging not only for their usage in the secondary vertex finding procedure but also for their IP that is already a discriminator against the non-b-flavored jets. The reconstructed tracks (see Section 2.2.1) with $p_T > 1.0\,\text{GeV}$ are further asked to fullfil some other quality requirements [178]. To reconstruct the vertex in the jet the tracks seem to be originated from the secondary vertex are find and undergo the dedicated hard-assignment vertex fit, namely the Trimmed Kalman Fitter (TKF) [89]. While the basic idea of the vertex reconstruction is similar to what presented in Section 2.2.1, in secondary vertex finding the vertex candidates are filtered against the primary vertex.
On top of the TKF, the Tertiary Vertex Track Finder is applied to find the additional

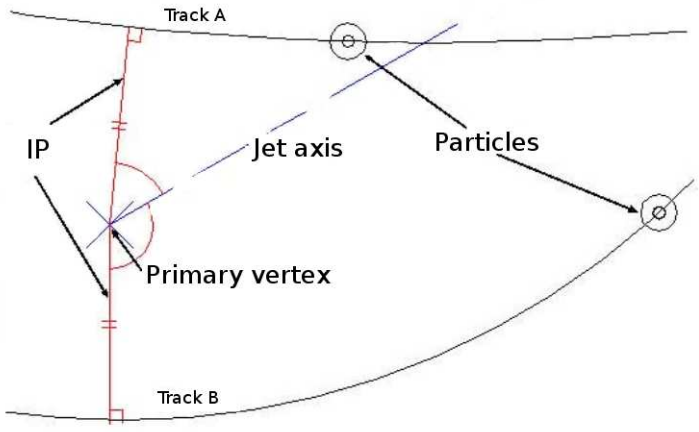


Figure 4.25: A schematic representation of the tracks impact parameter in jets.

tracks from the b-c-decay chain [179].
The track and vertex reconstruction provide the inputs for the b-tagging algorithms which are basically divided into three categories: IP based algorithms, vertex based methods and the algorithms based on finding a soft lepton within the jet. Each algorithm supposedly gives a unique output, the "discriminator", by which one can estimate the flavor of the jet.

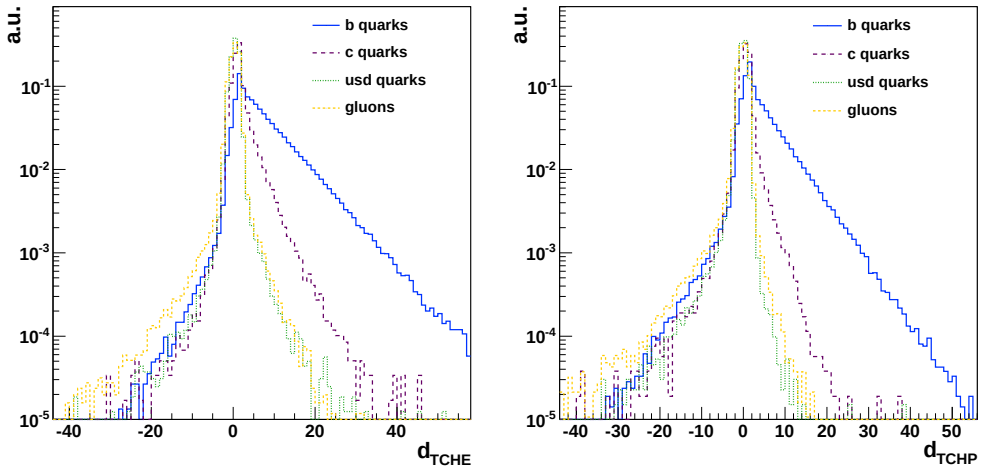
4.4.1 Track impact parameter based tags

The geometrical interpretation of the impact parameter for single tracks is illustrated in Figure 4.25. The IP segment here is defined as a vector from the primary vertex toward the tracks. The sign of the scalar product between the jet axis and the IP segment provides useful information about the jet flavor. For the example shown in Figure 4.25, the signed IP of the track A is positive where it is negative for the track

B.
While for the short lifetime decays the sign is randomly changed, for weakly decaying b-hadrons it is mostly positive. Therefore the signed IP can be used as a discriminar to recognize the b-qurak jets. Moreover, since the IP and its uncertainty, σ_{IP}, can be at the same oreder, the signed IP significance, IP/σ_{IP}, seems to be a better b-tagging observable.

Track Counting algorithms

In the so-called *Track Counting* algorithms a jet is identified as b-flavored if it contains at least N tracks with the signed IP/σ_{IP} greater than some threshold, S. The parameter N is fixed to $N = 2$ for the efficient b-tagging (Track Counting High Efficiency) where $N = 3$ is a better choice for making a purer b-jet selection (Track Counting High Purity). Having the N fixed, a continues distribution is obtained by looking at the IP significance of the N'track where the tracks are ordered by the signed IP/σ_{IP}. Because of its simplicity, the *Track Counting High Efficiency* algorithm has been widely used in 2010 data analyses in the CMS experiment. It is also the chosen algorithm for the method developed in this thesis. Figure 4.26 shows the discriminator distribution of



(a) (b)

Figure 4.26: The distribution of the track counting b-tag discriminators for the jets in $t\bar{t}$ events with the high efficiency, (a), and high purity, (b), parametrization.

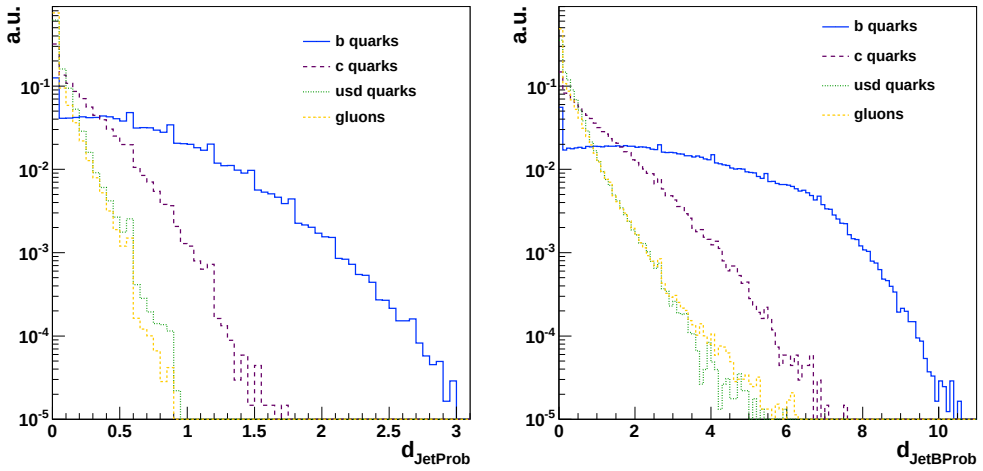
the track counting algorithms for b-jets, c-jets and other types of (u, s, d and gluon) jets in the $t\bar{t}$ events. The distributions are more populated in the positive large values for the b-quark jets than for the light jets. The S parameter is actually a cut on the disctriminator value which indicate the tightness of the b-jets selection. It can be seen that for given cut on the positive discriminator values, a purer b-quark jet will be obtained via the High Purity algorithm.

Jet probability algorithms

A track by track probability (P_{tr}) can be defined based on the impact parameter significance. The probability density function can be extracted from the tracks with negative IP significance, hence the tracks from b-hadron decays are accordingly given less probability. The probability that all tracks in the jet are compatible with the primary vertex is defined as

$$P_{jet} = \Pi \cdot \sum_{\mathrm{n=0}}^{\mathrm{N-1}} \frac{(-ln\,\Pi)^n}{n!} \tag{4.17}$$

where $\Pi = \prod_{i=1}^{N} P_{tr}(i)$. The discriminator of the *jet probability* (JP) algorithm is $-log(P_{jet})$ which gives higher values for b-jets. In the same circumstance the *jet B probability* (JBP) algorithm takes the four most displaced tracks to estimate how likely it is that the jet is b-flavored. The discriminator is defined as $\frac{-1}{4} \times (log(P_{\mathrm{jet}}) + log(P_{\mathrm{jet}}^{4\,\mathrm{trk}}))$ where $P_{\mathrm{jet}}^{4\,\mathrm{trk}}$ is related to the four mentioned tracks. The choice of four tracks is driven



(a) (b)

Figure 4.27: The distribution of the jet probability b-tag discriminators for the jets in $t\bar{t}$ events with the jet probability, (a), and the jet B probability, (b), definition.

by the average reconstruction efficiency for the tracks in jets, 80%, times the average number of tracks in b-jets, $\sim$5. The distribution of the jet probability discriminators are illustrated in Figure 4.27 where the behavior of b-, c- and $udsg$-jets in $t\bar{t}$ events are compared. The gluon jets and the uds-quark jets seem to behave similarly. They both are accomulated at low b-tag values while the b-flavored jets are distributed in the higher ranges. The c-quark jets are found to be in between.

4.4.2 Secondary vertex tags

There are two different approaches to identify b-jets using the secondary vertex. Prior to the algorithms, the jets are categorized based on the presence of a secondary vertex:

Jets with a *RecoVertex* contain at least one qualified secondary vertex. A *PseudoVertex* is created for the jets in which no reconstructed secondary vertsex is found. Two tracks with IP significance greater than 2 are needed to create such vertex. Jets do not fit in any of the mentioned categories are labeled as *NoVertex*.
The *simple secondary vertex* algorithm looks for at least one *RecoVertex*, returns no discriminator otherwise; hence its maximum efficiency is limited to the probability of finding a vertex in the jets with a weakly decaying b-hadrons. This algorithm can also have the *high efficiency* (SSVE) and *high purity* (SSVP) versions. Where for the former the number of track associated to the secondary vertex is $N_{trk} \geq 2$, for the latter, $N_{trk} \geq 3$ is required. The significance of the 3D flight path is the discriminator value for this algorithm.

The *combined secondary vertex* (CSV) algorithm is more sophisticated since it uti-

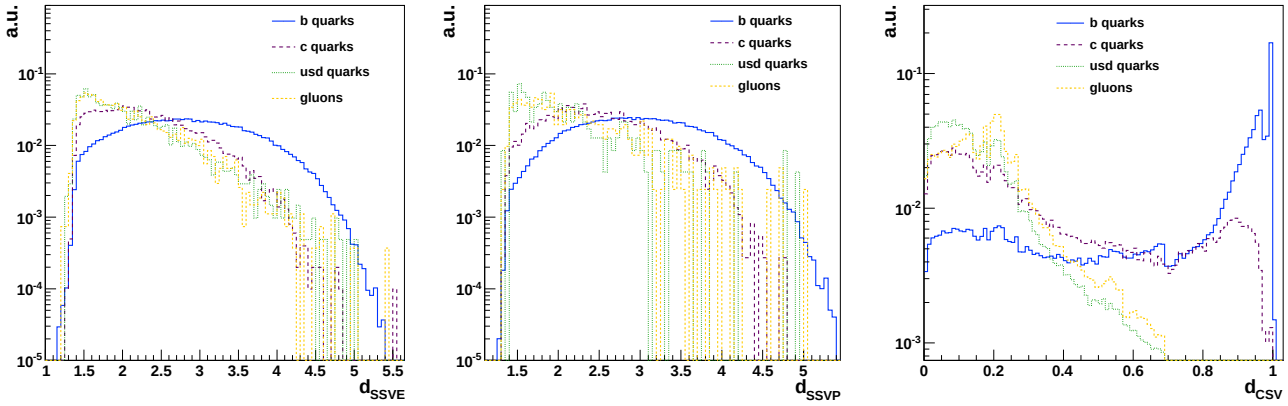


Figure 4.28: The distribution of the high efficiency, (a), and high purity, (b), simple secondary vertex b-tag discriminators together with the b-discriminant of the combined secondary vertex algorithm, (c), for the jets in $t\bar{t}$ events.

lizes additional variables [178] to provide the discriminator value even for the *Pseudo* and *NoVertex* cases. The variables are combined in a single discriminator using the Liklihood Ratio technique,

$$\mathcal{L}^{b,c,q} = f^{b,c,q}(\alpha) \times \prod_i f_\alpha^{\mathrm{b,c,q}}(x_i). \tag{4.18}$$

In Equation 4.18 $\alpha(= 1, 2, 3)$ denotes the vertex category, q stands for the jets other than b- and c-jets and x_i are the individual variables. The $f^m(\alpha)$ is the probability for the flavor m to fall into the α vertex category and $f_\alpha^{\mathrm{m}}(x_i)$ is the probability density function of x_i variable in the α category for the flavor m. To account for the differences between c-jets and $udsg$-jets, the final discriminator is defined as

$$d = f_{BG}(c) \times \frac{\mathcal{L}^b}{\mathcal{L}^b + \mathcal{L}^c} + f_{BG}(q) \times \frac{\mathcal{L}^b}{\mathcal{L}^b + \mathcal{L}^q}, \tag{4.19}$$

where $f_{BG}(c) + f_{BG}(q) = 1$ since f_{BG} denotes the c- and q-content of non-b-jets. The b-discriminators for the secondary vertex algorithms are shown in Figure 4.28 where the

distributions for b- and non-b-jets in the $t\bar{t}$ sample are illustrated. While all discriminators show different behaviors for the b-quark and non-b-quark jets, a clear distinction is obtained for the combined secondary vertex algorithm.

4.4.3 Soft lepton tags

The presence of a lepton in the weak decay of b-hadrons can be complemented with some other variables to create a b-discriminator. CMS has developed the soft lepton b-tagging algorithms both for the muons and the electrons [81] where the muon b-taggers are more common. The *soft muon by* $p_{T_{rel}}$ method (SMPt) relies on the angle between

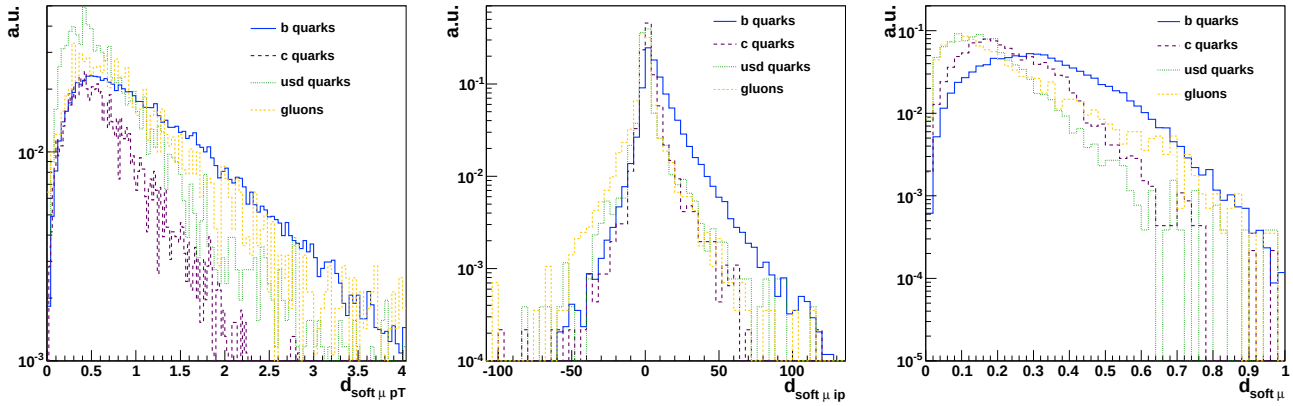


Figure 4.29: The distribution of the soft muon based b-tag discriminators for the jets in $t\bar{t}$ events. The $p_{T_{rel}}$ discriminator, (a), the muon IP significance, (b), and the discriminator from the neural network which is a combination of variables, (c), are presented.

the muon and the jet axis which tends to be close to 90°. The $p_{T_{rel}}$ variable is defined as the p_T of muon with respect to the jet axis. The larger $p_{T_{rel}}$ leads to the higher purity in the b-jet selection. In the *soft muon by IP significance* algorithm (SMIP) the discriminator is the IP significance of the muon when it is positive. In case of more muons in the jet, the one with highest significance is taken in both algorithms.
The distance between the muon and the jet axis in the $\eta - \phi$ plane together with the lepton momentum to the jet energy ratio, are combined to the $p_{T_{rel}}$ and the impact parameter significance by a neural network. The b-quark jets from the $t\bar{t}$ events and the non-b-quark jets from the QCD multi-jet samples are used for the training. The resulting discriminator is known as the *soft muon* (SM) b-tagging algorithm. Figure 4.29 shows the distribution of the soft muon b-discriminators for jets in the semi-electron final state of $t\bar{t}$ events where the jets are categorized according to their flavor.

4.4.4 The performance of the b-tagging algorithms

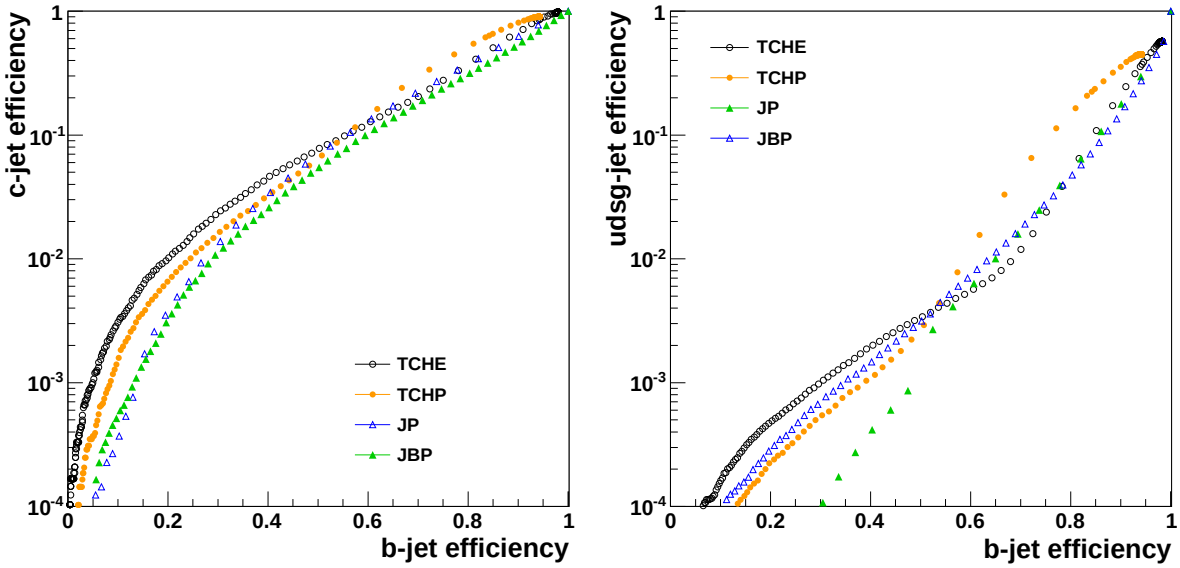
The performance of the b-tagging algorithms is examined by looking at their power in recognizing true b-jets. While it happens that some true b-jets are not accepted by the

algorithm, jets with an origin different from b-quarks can be tagged as b-jets. Hence one can define the efficiency, ϵ_b, and the mis-tag rate, $\bar{\epsilon}_b$, of a b-tagging algorithm

$$\epsilon_b = \frac{N^{\text{accepted}}_{\text{true b-jets}}}{N_{\text{true b-jets}}}, \qquad \bar{\epsilon}_b = \frac{N^{\text{accepted}}_{\text{true q-jets}}}{N_{\text{true q-jets}}}, \tag{4.20}$$

where q denotes either the gluons or non-b-quarks. The quantity $N^{accepted}$ in Equation 4.20 is the number of jets whose b-discriminator value exceeds some threshold. The thresholds are chosen to maximize the efficiency for a certain mis-tag rate. Generally speaking, three cut values referred to as *working points* are defined per algorithm for different analysis purposes: loose, medium and tight where the latter gives the highest purity or the lowest mis-tag rate.

To illustrate the performance of the b-tagging algorithms in terms of the efficiency and

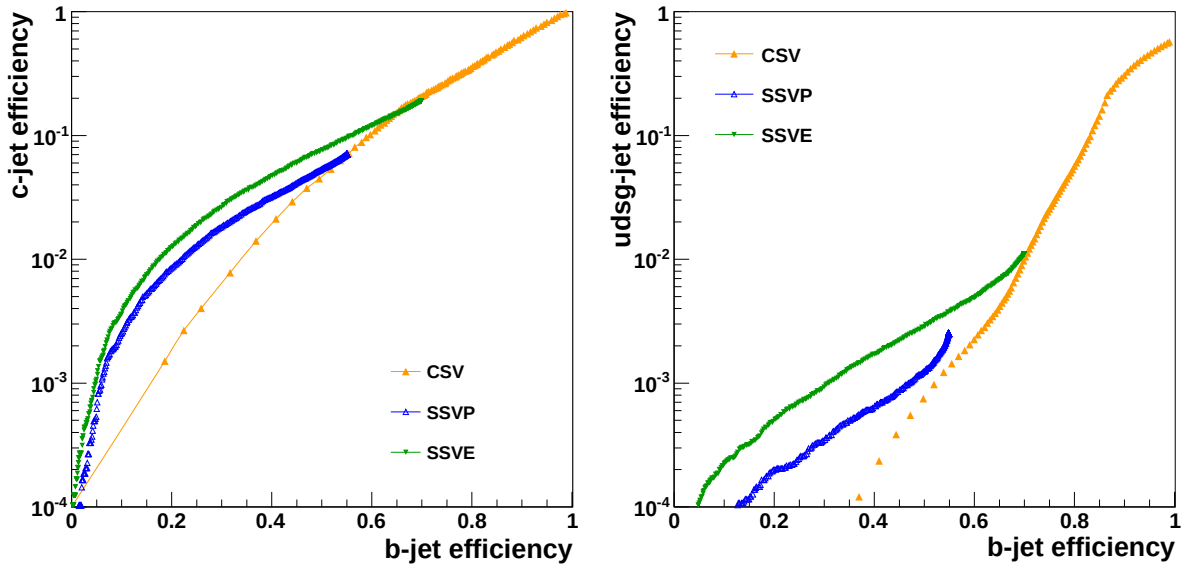


(a) (b)

Figure 4.30: The efficiency of the IP based b-tagging algorithms for c-quark jets, (a), and $udsg$-jets, (b), versus the efficiency for b-quark jets. Jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ are selected within semi-electron $t\bar{t}$ events.

the mis-tag rate, jets in the semi-electron $t\bar{t}$ sample are classified based on their origins into b-jets, c-jets, and $udsg$-jets where g stands for the jets with a gluon origin. Jets are required to have a $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$. Considered as heavy with respect to gluons and uds-quarks, the jets originating from c-quarks can have some properties in common with the b-jets. Therefore it makes sense to investigate the efficiency of the b-tagging algorithms for c-jets separately. Figure 4.30 demonstrates the efficiency of the IP based b-tagging algorithms for non-b-jets versus the efficiency for b-jets. All algorithms show a higher mis-tag rate for c-jets. The JBP algorithm gives a lower mis-tag rate for c-jets at almost any b-jet efficiency. The algorithm also works fine in rejecting the light (u, d, s) quark and gluon jets over a wide range of b-jet efficiencies.

The TCHP algorithm has a good performance for tight b-jet selections, i.e. $\epsilon_b < 30\%$. The performance of the vertex based algorithms are shown in Figure 4.31. It can be



(a) (b)

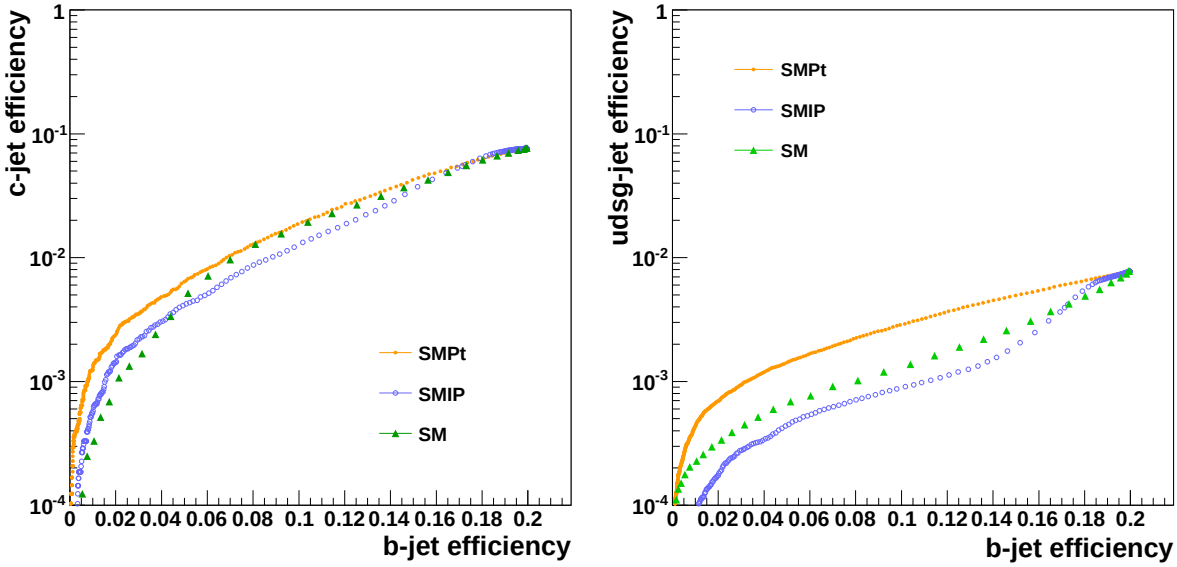
Figure 4.31: The efficiency of the vertex based b-tagging algorithms for c-quark jets, (a), and $udsg$-jets, (b), versus the efficiency for b-quark jets. Jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ are selected within semi-electron $t\bar{t}$ events.

seen that the b-jet efficiency for the SSVE (SSVP) algorithm is limited to about 70% (55%) reflecting the limited probability of finding a vertex in a b-jet. The CSV algorithm gives a lower mis-tag rate for both c- and $udsg$-jets over a wide range of the b-jet efficiency. For the b-jet efficiencies below 55%, the SSVP algorithm is more performant than the SSVE one, as expected. As illustrated in Figure 4.32, the b-jet efficiency of the soft muon taggers is constrained to $\lesssim 20\%$ due to the branching fraction of the B meson in soft muons. The SMIP algorithm seems to perform better in non-b jet rejection for almost all b-jet efficiencies.
The TCHE b-tagging algorithm is chosen as the baseline for the analysis developed in this thesis.

4.4.5 Methods to investigate the b-tagging performance

Different methods have been developed to study the performance of the b-tagging algorithms where the intention is to apply the whole measurement consistently on the collision data itself [180, 181].

Negative tags for $\bar{\epsilon}_b$ estimation : The method is developed for the mis-tag rate estimation of the uds-quark jets and gluon jets which are expected to give a similar distributions for the positive and negative tag valuess [182].



(a) (b)

Figure 4.32: The efficiency of the soft muon taggers for c-quark jets, (a), and $udsg$-quark jets, (b), versus the efficiency for b-quark jets. Jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ are selected within the semi-electron $t\bar{t}$ events.

From an inclusive jet data sample, a non-b-jet enriched sample is made by selecting jets with the negative tag value. Assuming that the final mis-tag rate is to be estimated on a positive b-discriminator value of $d = d^*$, the efficiency of the negatively tagged jets, $\epsilon^{(-)}_{data} = \frac{N_{accepted}}{N_{all}}$, is measured at the $d = -d^*$ working point in data.

If the negatively tagged jet sample in data was not contaminated by b- or c-quark jets, $\epsilon^{(-)}_{data}$ would already be an estimate for the mis-tag rate of uds-quark and gluon jets. To account for such contaminations, $\epsilon^{(-)}_{MC}$ which is defined like $\epsilon^{(-)}_{data}$ but on a simulated, negatively tagged jet sample together with the mis-tag rate for true non-b-jets, $\bar{\epsilon}_{MC}$, at $d = d^*$ are extracted from simulation. Finally under the assumption of

$$\frac{\bar{\epsilon}_{data}}{\bar{\epsilon}_{MC}} = \frac{\epsilon^{(-)}_{data}}{\epsilon^{(-)}_{MC}}, \tag{4.21}$$

$\bar{\epsilon}_{data}$ can be estimated. The negative discriminator in the track counting methods is obtained by reordering the tracks in the jet in such a way that the track with the largest negative IP is the first. For the vertex based algorithms, the discriminator becomes negative if the secondary vertex is reconstructed upstream with respect to the primary vertex.

Estimation of ϵ_b using $p_{T_{rel}}$: This method is applied on the di-jet events that have a muon close to one of the jets, relying on the fact that the jets in di-jet processes are from the same flavor [183]. While the $p_{T_{rel}}$ spectrum is fitted to the sum of the

expected templates for the light, b- and c-quarks in order to extract the fractions of b- and non-b-jets in the sample, the efficiency of the b-tagging algorithm is measured on the second jet after subtracting the non-b-jet contributions. Apart from the light jet content which can also be estimated from data, other templates are taken from simulation. The data-driven template of light jets is obtained from an inclusive jet sample for which any *high purity* track in the jet is taken as a muon candidate.

System8 method to estimate $\epsilon_{\mathbf{b}}$ and $\bar{\epsilon}_{\mathbf{b}}$: Two weakly correlated b-taggers and a sample containing muons within jets are the inputs for the system8 method where the jets passing a dedicated b-tag cut form the tagged sample [184]. Writing the flavor content for the combination of the samples and b-taggers makes a system of 8 equations in which the correlation factors are takenfrom simulation. Numerically solving the 8 equations, the efficiency and the mis-tag rate of the two b-tagging algorithms are estimated simultaneously.

Measurement of $\epsilon_{\mathbf{b}}$ in $\mathbf{t\bar{t}}$ events : A method to estimate the b-tagging efficiency within top-quark pair events in the both semi- and di-lepton final states is developed in [185]. For the semi-lepton channel, the events go through a dedicated $t\bar{t}$ selection and the hadronically decaying top-quark ($t \to Wb \to qq'b$) is reconstructed using a Likelihood ratio technique. One of the jets in the reconstructed top-quark needs to be tagged as a b-jet to purify the sample. The ϵ_b is estimated on the remaining jet in the event which is supposedly the b-jet from the leptonically decaying side, $t \to Wb \to l\nu b$. Given the number of tagged jet as $N_{tag} = N_{\rm tag}^{\rm b} + N_{\rm tag}^{\rm non-b}$, one can obtain

$$\epsilon_b = \frac{1}{x_b}[x_{tag} - \bar{\epsilon}_b(1 - x_b)], \tag{4.22}$$

where x_b is the content of true b-jets in the sample, derived from simulation. The fraction of tagged jets, x_{tag}, together with the mis-tag rate can be measured on data.

Chapter 5

Measurement of the b-tagging efficiency with $t\bar{t}$ events

The aim of this chapter is to develop a method to estimate the b-tagging efficiency, introduced in Chapter 4, using $t\bar{t}$ events. The method has already been checked on the semi-muon final state [186] and the focus in this thesis is on the semi-electron decay channel. From the theoretical point of view, these two channels are the same in particular at the TeV scale since both the electron and the muon in the final state are in the massless regime. However from the detector side, as described in Chapters 2 and 4, these two objects are quiet different. Although the electron reconstruction efficiency in CMS is high enough, the muon objects make much clearer signatures in the detector specially thanks to the dedicated muon system.
Prior to the b-tagging efficiency measurement, the top quark events need to be discriminated from the huge amount of background processes by means of sequential selection requirements. Considering the final state of interest, $t\bar{t} \rightarrow \bar{b}bqq'\nu e$, one basically looks for events containing a single electron and "at least" four jets to account for the extra radiation jets associating the process. The identification criteria on the physics objects, explained in Chapter 4, are complementary to the selection sequence to make the sample as pure as possible. The event selection strategy together with the selection performance are addressed in Section 5.1.
The event selection results in a sample enriched with the top-like events. But it is still challenging to search for the b-jets among at least four jets present in the event. Hence, building a tool to make a non-biased b-jet sample is crucial. The b-jet sample gets even more purified if one can subtract the non-b-jets mistakenly entered the sample. In Section 5.3 it is explained how a pure b-jet sample is prepared and how the b-tagging efficiency measurement is performed on the sample with the least possible bias. The evaluation of the systematic uncertainties on the method is presented in Section 5.5.
Finally, Section 5.6 is devoted to the first application of the method in the semi-electron channel on the 7 TeV data collected by the CMS experiment in 2010.

5.1 Selection of the $t\bar{t}$ candidates in the semi-electron channel

To restrict the huge size of the background, QCD multi-jets in particular, events are pre-selected before they go for the final filtering. The event selection procedure in the simulation based analysis here follows the base line defined for selecting the events from the pp collision products within the top quark analysis group in CMS [1]. The pre-filtering contains the HLT (see Section 2.2.4) requirement, the request for the presence of a good primary vertex and some loose cuts on the electron and the jet candidates. This is followed by the event selection by looking for an isolated, high-p_T electron where events with extra electrons or muons are vetoed afterward. Finally the request for at least four high-p_T jets completes the picture of the $t\bar{t} \to \bar{b}bqq'\nu e$ process. This selection chain is applied on the signal and background simulated samples [2] summarized in Table 3.2.

5.1.1 The HLT and pre-filtering requirements

All events are asked to pass the trigger path labeled as `HLT_Ele15_SW_L1R` for which an electron trigger object with $E_T > 15\,\text{GeV}$ is looked for as explained in Section 2.2.4. The Level-1 trigger seed for this path, labeled as `L1_SingleEG8`, is based on the presence of a supercluster with the corrected energy of $E_T^{corr} > 8\,\text{GeV}$. The `SW` in the name of the trigger path denotes for the startup conditions applied on the electron pixel-matching window. The trigger efficiency is 93% for signal while for the QCD background it is about ∼26%.
At least three jets with $p_T > 15\,\text{GeV}$ and $|\eta| < 3.0$ together with at least one electron with $p_T > 20\,\text{GeV}$ and $|\eta| < 2.5$ have to be present in the event passing the HLT criterion. More than 88% of the triggered signal events have the desired "electron and jets" signature where ∼73% of the QCD multi-jets are rejected. The pre-filtering is completed by looking for at least one primary vertex of interaction which is away from the center of the detector in z-direction not more than 15 cm. On the transverse plain it is expected to be in a circle around the beam line with $\rho < 2.0\,\text{cm}$ radius. The vertex should not be labeled as `Fake` and has to have ndof > 4, as defined in Equation 2.5. The efficiency of finding a good primary vertex is close to one.
As is summarized in Table 5.1, the pre-selection efficiency is high enough for the processes with real electrons while it considerably reduces the size of the QCD background.

5.1.2 Electron selection and the extra lepton vetos

At the first step of the event selection, events are requested for the presence of exactly one electron with the following properties which are already defined in Chapter 4:

[1] The event selection criteria in 2010 has been evolving with time. Different versions can be found here: `https://twiki.cern.ch/twiki/bin/view/CMS/TopLeptonPlusJetsRefSel_el`.

[2] The samples indicated as *Spring10* in Table 3.2 are used to develop the method for the b-tagging efficiency where the *Fall10* samples are employed in the first application of the method on the collision data.

	$t\bar{t}$ (signal)	$t\bar{t}$ (others)	single-top	W+jets	Z+jets	QCD
pre-selection efficiency	82%	34%	35%	20%	29%	6.3%

Table 5.1: The pre-filtering efficiency for different simulated samples contributing to the analysis. The size of the QCD sample is reduced by $\sim 94\%$ where only 18% of the signal sample is rejected.

- $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$. The supercluster of the electron candidate needs to be out of the barrel-endcap transition region of the ECAL, $1.4442 < |\eta^{sc}| < 1.566$. While the choice of the η-range is driven by the detector acceptance, the high-p_T cut is justified considering the relatively high energy carried by the electron from the W boson in the $t\bar{t}$ events (see e.g. Figure 4.13).

- $relIso < 0.1$ to reduce the rate of accepting the electrons within jets.

- Identified as an electron according to WP70 requirements (Table 4.1). With the identification requirements it is checked if the electron candidate has a consistent signature in the ECAL and the tracker system. The individual electron shower shape in the ECAL which is affected by the bremsstrahlung photon emissions is also considered.

- The two dimensional impact parameter with respect to the primary vertex in the transverse plain has to be $d_0(p.v.) < 200\,\mu m$. The electron candidate with a small $d_0(p.v.)$ comes most probably from the main interaction point rather than processes like photon conversion.

- Not coming from the photon conversion regarding the partner track veto: $|\Delta cot(\Theta)| < 0.02$ and $|Dist| < 0.02$. Two parallel tracks can be an indication for the photon conversion as explained in Section 4.1.3.

Normalized to $100\,pb^{-1}$ integrated luminosity, Figure 5.1 shows the distribution of the transverse momentum for the leading electron (ordered by p_T) in different simulated samples. The clear cut of $p_T > 20\,\text{GeV}$ is due to the pre-selection. The choice of the leading electron for illustration is because of the fact that events for which the leading electron does not pass the p_T threshold are rejected anyway. Due to the huge amount of the QCD multi-jets, the p_T distribution is plotted with and without the QCD sample. Hence the distributions for processes other than QCD multi-jets are better visible.
The distributions are dominated by the electrons from the two challenging background processes, W+jets and QCD multi-jets as expected. While for the electrons in the multi-jet events no special behavior is seen, the real electrons in W boson decay carry an amount of energy which is on average close to half of the rest mass of the W boson.
Figure 5.2 illustrates the discriminating power of the identification and isolation for

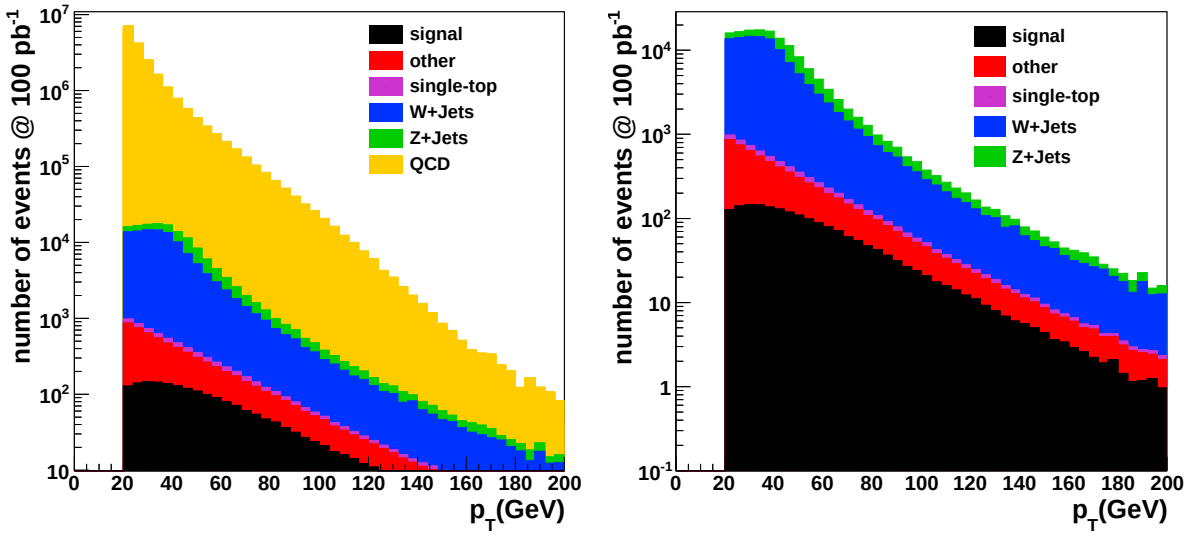


Figure 5.1: The p_T distributions of the electron. While the left plot contains all the signal and background processes, the right plot is without QCD multi-jets contribution. The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.

the electrons where the results of all identification requirements are combined in a boolean variable. The distribution of the impact parameter, $d_0(p.v.)$, of the leading electron is shown in Figure 5.3. It can easily be seen that most of the electrons in QCD processes are not coming from the primary vertex of the interaction. The number of selected electrons is shown in Figure 5.4. While asking for at least one prompt electron significantly reduces the QCD contamination as well as a good fraction of the W+jets events, tightening the requirement to the presence of "exactly" one selected electron helps suppressing the Z+jets background contribution.

Within the events containing exactly one prompt electron, it is requested for the prompt electron candidate to not come from the photon conversion. Figure 5.5 (a) depicts the fraction of the conversion electrons in different samples according to the presence of a partner track. The plot shows that the partner track veto for the prompt electron candidate is effective in rejecting QCD multi-jets, even after the cut on $d_0(p.v)$.

Another property of the electron candidate which is worth studying is the electron seeding information discussed in Section 4.1. Since the p_T of the isolated electron in the $t\bar{t}$ event is relatively high, it is expected to reach the ECAL material and to be recognized by the ECAL driven seeded reconstruction algorithms. On the other hand, the electrons in the jets and the low-p_T electrons are mostly recovered only by the tracker driven seeding algorithms.

In Figure 5.5 (b) the seeding information of the electrons passed all the mentioned requirements is illustrated. If the electron is seeded with both approaches, it is flagged as ECAL driven. It can be seen that even after all other qualification requests, rejecting the tracker driven electrons still suppresses the QCD background. Therefore, the electrons are requested to be flagged as ECAL driven.

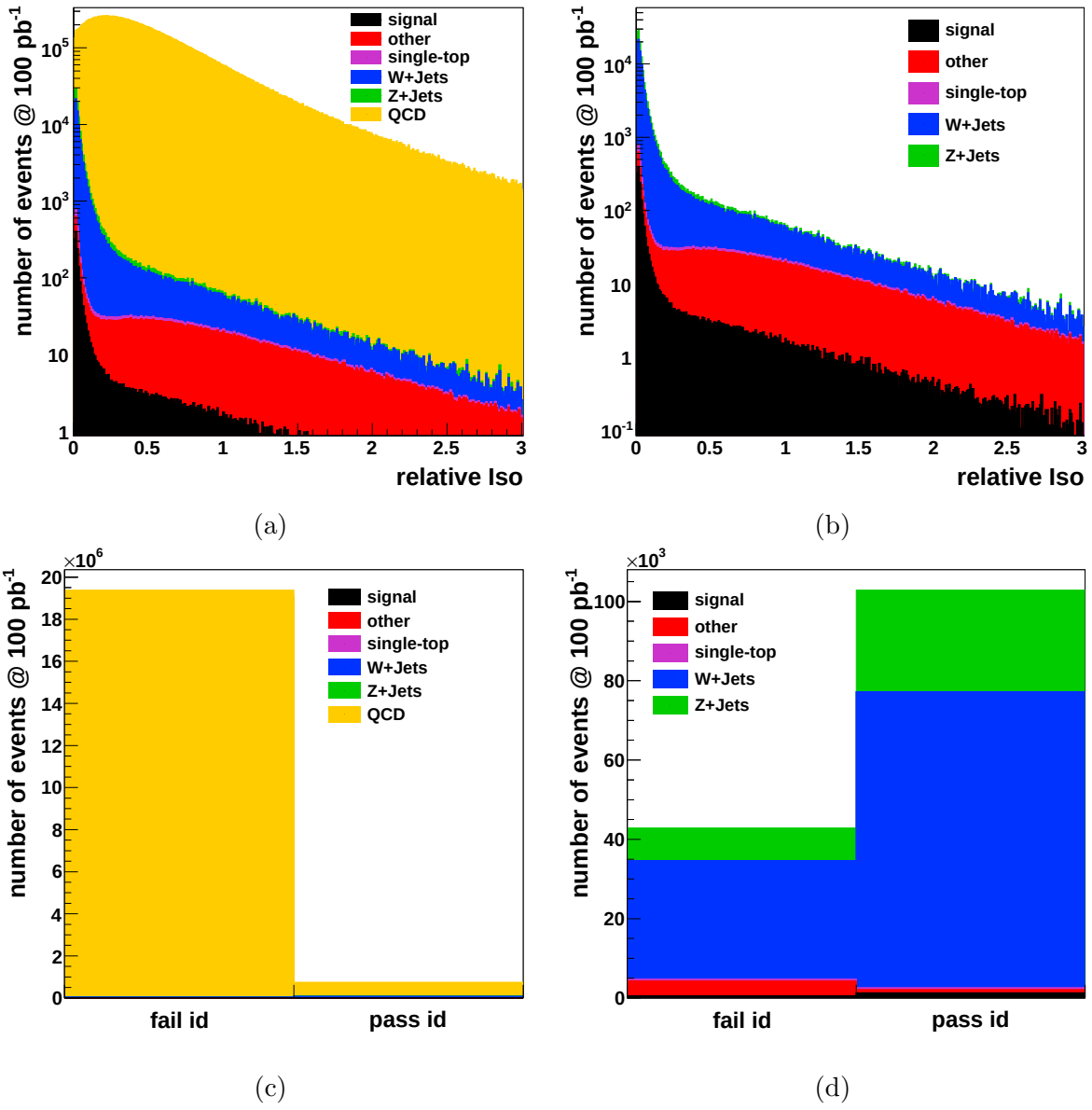
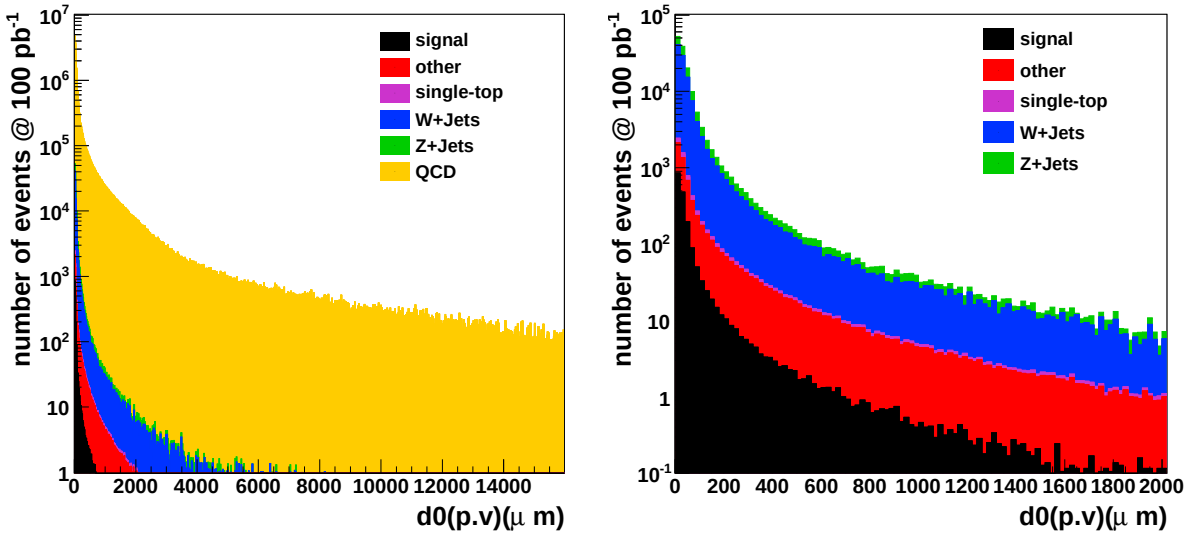
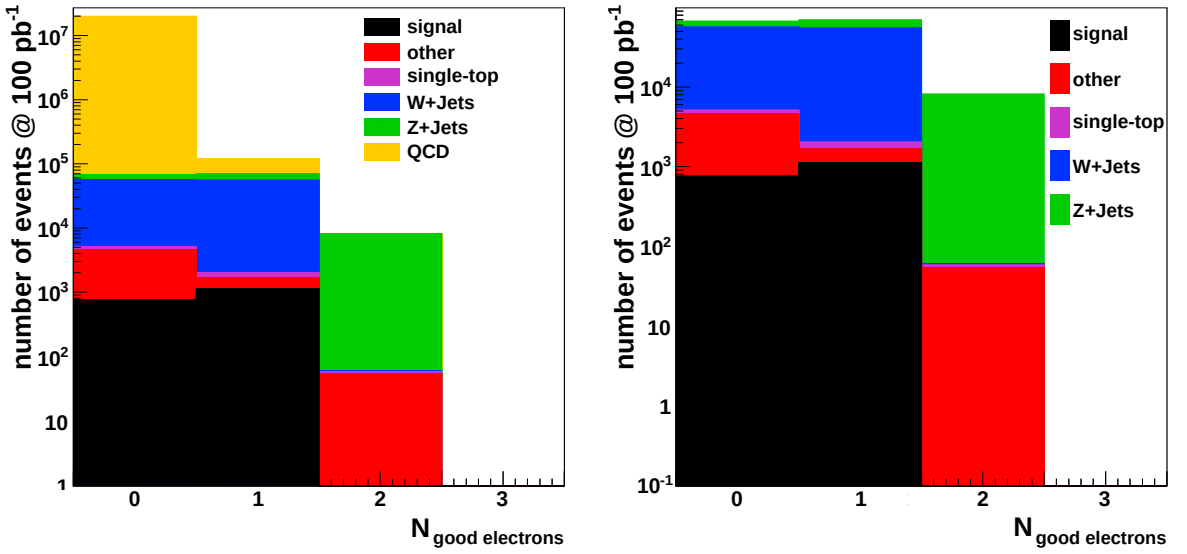


Figure 5.2: The relative isolation variable, (a) and (b), and the WP70 identification requirements combined in a single boolean, (c) and (d), for the electron within the $t\bar{t}$ and different background processes. The plots on the left column, (a) and (c), include the QCD multi-jet contribution where (b) and (d) are without the QCD component. The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.



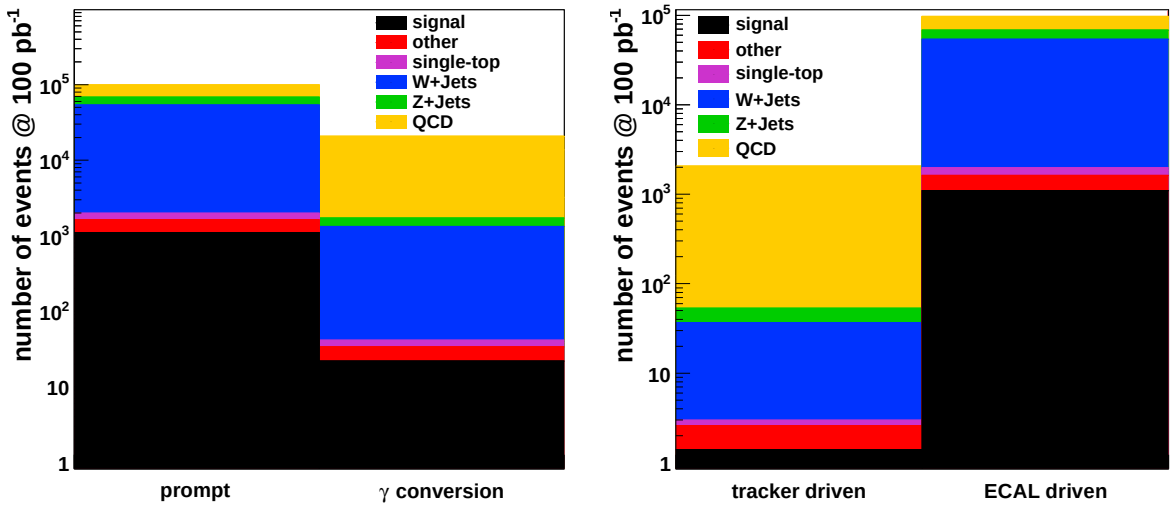
(a) (b)

Figure 5.3: The transverse impact parameter w.r.t to the primary vertex in the $t\bar{t}$ and different background processes with (a) and without (b) the QCD multi-jet contribution. The histograms are normalized to $100\,pb^{-1}$ integrated luminosity.



(a) (b)

Figure 5.4: Number of selected electrons in the $t\bar{t}$ and different background processes with (a) and without (b) the QCD multi-jet contribution. The histograms are normalized to $100\,pb^{-1}$ integrated luminosity.

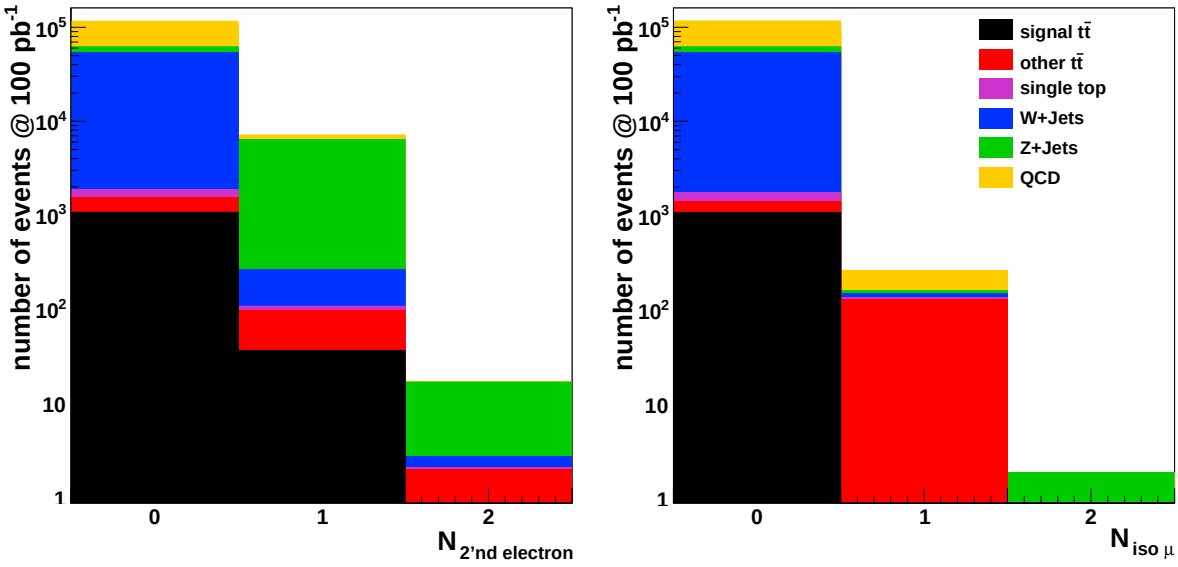


(a) (b)

Figure 5.5: The indicator for the existence of a partner track as a sign of conversion (a) and the seeding information for the electrons (b) in the $t\bar{t}$ and different background processes. The plot in (b) is made from the electrons in the "prompt" bin of distribution (a). The histograms are normalized to $100\,pb^{-1}$ integrated luminosity.

Vetoing extra electrons and muons

Although there is a strict limit on the number of prompt electrons, there is still a chance to find looser electrons in the event. The main physics process providing such a signature is $Z \rightarrow ee$. The di-electron final state of $t\bar{t}$ or the final state with one electron and one τ-lepton for which the τ-lepton decays leptonically, $t\bar{t} \rightarrow \bar{b}be\nu_e\tau\nu_\tau$, are also other possibilities. To avoid these contributions, the event is rejected if it contains extra electrons with $p_T > 20\,\text{GeV}$ and $|\eta| < 2.4$ (ECAL gap excluded). The extra electrons need to fulfill the WP95 identification requirements as listed in Table 4.1. According to Figure 5.6 (a), a significant amount of events with the second loose electron belongs to the Z+jets process. The second electron veto reduces the Z+jets contribution by a factor of a bit less than 50% where only about 12% of $t\bar{t}$ events in other final states are rejected. Since the second electron requirements are loose, there is a possibility for an electron in jets to be selected. This leads to a small signal rejection of 4%.
The presence of a prompt muon in the event is an indication for $t\bar{t} \rightarrow \bar{b}be\nu_e\mu\nu_\mu$ decay channel which is suppressed by the muon veto. Events are rejected if a muon candidate with $p_T > 20\,\text{GeV}$, $|\eta| < 2.4$ and $reliso < 0.05$ is found. The muon candidate needs to meet the additional criteria of $\chi^2 < 10$, $d_0(p.v.) < 0.02\,\text{cm}$ and the number of valid hits ≥ 11. Figure 5.6 (b) illustrates different background contributions discarded by the muon veto where the rejection power for other $t\bar{t}$ decay channels is about 27%. The presenece of Z+jets events with two isolated muons implies that the probability of finding a prompt electron in $Z \rightarrow \mu\mu$ processes is not zero although it is rather small ($\sim 0.1\%$).



(a) (b)

Figure 5.6: The number of loose electrons other than the prompt candidate (a) and the number of prompt muons (b) within the $t\bar{t}$ and different background processes containing exactly one prompt electron. The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.

As it can be seen in Figures 5.6 (a,b), about 1% of the W+jets events are also removed by the lepton veto due to the presence of the fake leptons. Almost no signal event is rejected by the muon veto.

5.1.3 Jet selection requirements

The jet content of the event is the final part investigated in the event selection. Before applying any cut on the jets, one need to make sure that the objects are not electrons mimicking the jet features. This can happens since the electron and jet reconstruction algorithms are run individually and are isolated from each other. Hence the same energy deposits in the calorimeters might be used by both algorithms. If the same energy clusters succeed to be integrated in a jet and to be used in the electron supercluster formation, the energy will be double counted, the jet collection will be contaminated by the electron contribution.
To avoid such confusion, different solutions are proposed in CMS, e.g. [187]. The approach which is used by PAT (see Section 2.3.1) at the analysis level is to remove the well-identified electron candidates from the jet collection i.e. for each well-identified electron candidate, the closest jet is rejected if $\Delta R(jet, electron) < 0.3$. In the analysis presented here, the ECAL driven electrons fulfilling the electron selection requirements other than the partner track veto are considered as the so-called "well-identified" electrons.
As it is shown in Section 4.3, the jets in $t\bar{t}$ events have relatively high energy and are

mostly in the central part of the detector. It is assumed that the four quarks in the semi-electron final state of the $t\bar{t}$ event are energetic enough to make the four leading jets (ordered by p_T) after hadronization and at the reconstruction level. The four leading jets with the corrected $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ are accepted to go through the other selection steps. The correction includes the relative η and absolute p_T calibrations explained in Section 4.3.2. Figure 5.7 shows the distributions of the kinematic

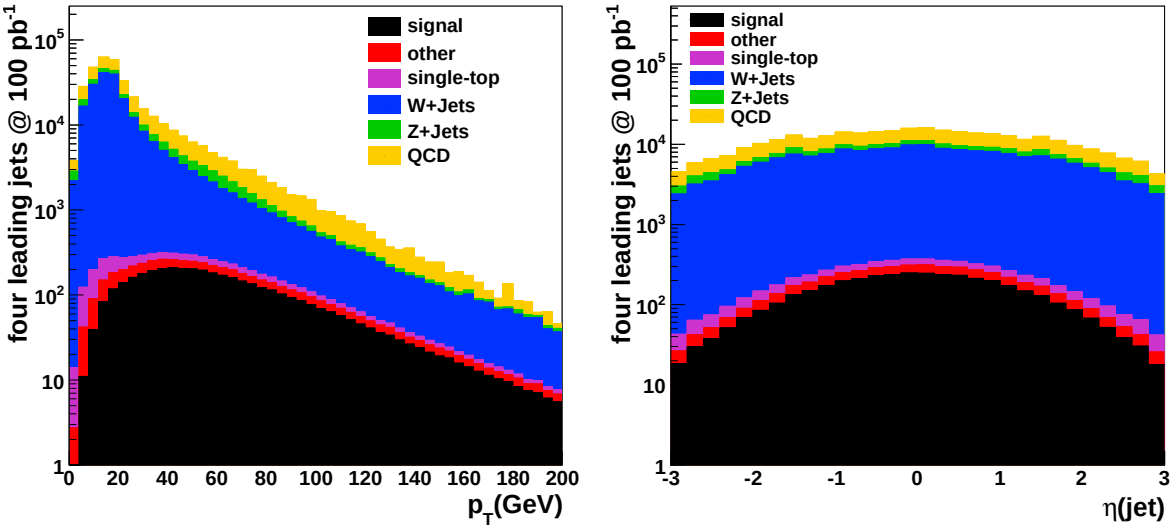
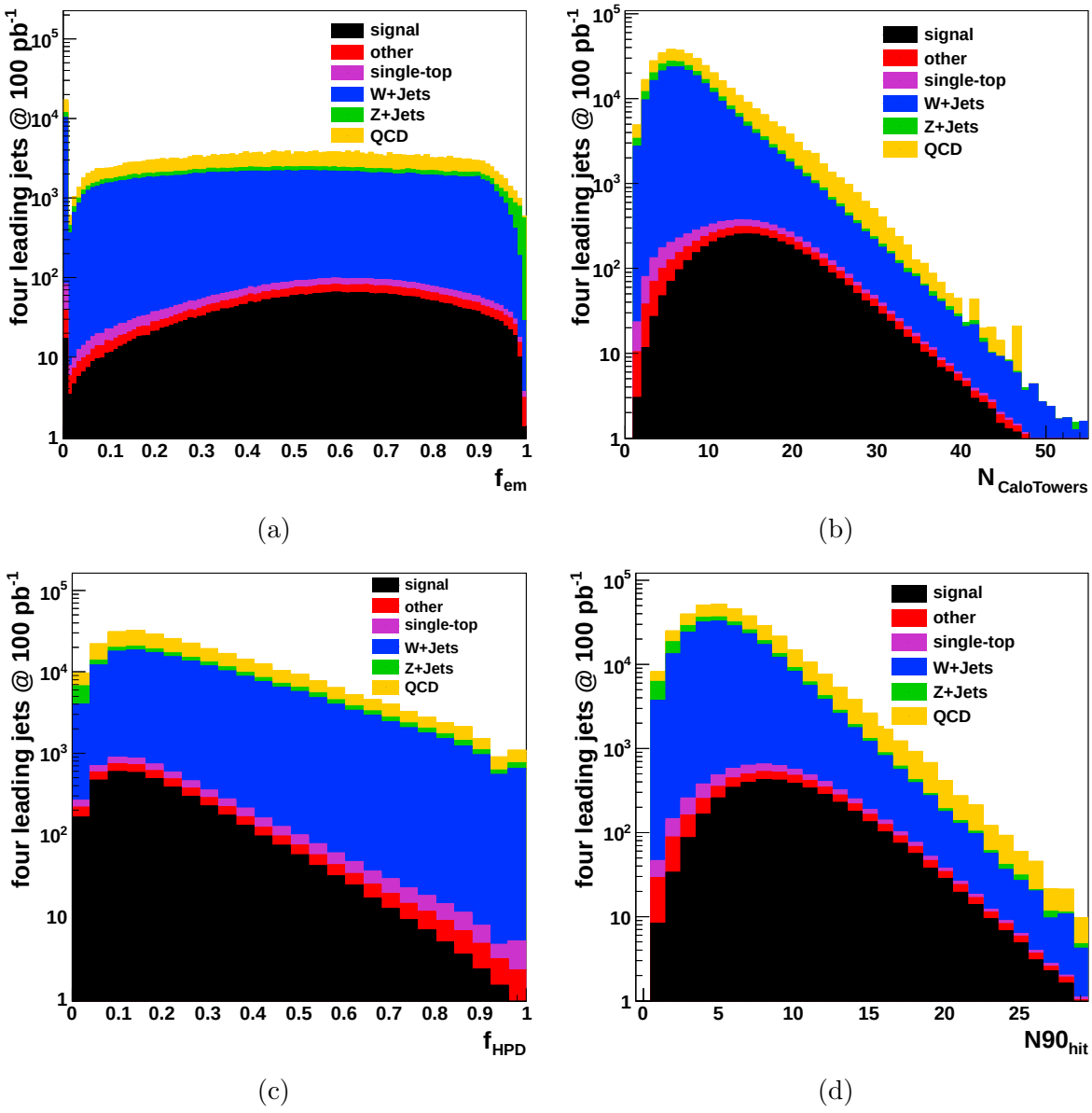


Figure 5.7: The p_T and $|\eta|$ distributions of the four leading jets within the $t\bar{t}$ and different background processes survived the lepton veto. The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.

properties for the four leading jets in different samples contributing in the analysis. All contributions are normalized to $100\,pb^{-1}$ integrated luminosity. Although the QCD contamination has largely been reduced by the electron selection, there are still QCD multi-jet events in the whole range of p_T and η of the jets. The most challenging background now is the W+jets process. This background has a prompt electron from the W boson decay hence it survives the electron selection. This is in particular important to decrease the W+jets fraction because it contains mostly non-b-jets and can influence the final analysis which aims to measure the b-tagging efficiency.

Extra conditions can be imposed on the jets based on the identification variables introduced in Section 4.3.3. The jet identification variables of the four leading jets are illustrated for the signal and the background contributions in Figure 5.8. The jets have already passed the kinematic requirements. Comparing to processes like W+jets, the $t\bar{t}$ jets are better identified. They are firing more CaloTowers, distinguishable from HCAL readout noise (lower f_{HPD}) and are with reasonable electromantic energy fraction (see Section 4.3.3).

The proposed jet identification requests in top quark analysis group are $f_{em} > 0.01$ to reject the possible HCAL noises, $N90_{hits} > 1$ to accept mostly the physical jets and $f_{HPD} < 0.98$ to accept the real jets rather than the noise in the HCAL readout. How-



(a) (b) (c) (d)

Figure 5.8: The jet identification variables of the four leading jets in the $t\bar{t}$ and different background processes survived the lepton veto: electromagnetic energy fraction (a), number of calotowers (b), f_{HPD} (c) and $N90_{hits}$ (d). The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.

ever looking at the distributions in Figure 5.8, it makes sense to study the effect of the other variable, $N_{CaloTowers}$. Besides, the accumulation of the Z+jets events in $f_{em} \approx 1$ indicates the presence of the electrons in the jet collection even after the jet-electron cleaning explained before.
To find an effective combination of the jet identification variables, $N_{CaloTowers} > 4$ and $f_{em} < 0.9$ are added to the list of proposed identification cuts. The lower limit on the f_{em} is enhanced to 0.05 and is fixed. For the rest, the effect of each individual cut on the signal over background ratio is investigated.
Two definitions of signal and background are considered: The first, scenario I, is the usual definition in which all processes other than $t\bar{t} \to \bar{b}bqq'\nu e$ are considered as background; In another view, since one is interested to prepare a b-jet sample at the end for the b-tagging efficiency measurement, all processes containing real b-jets are put on the signal side. It means the signal sample contains all $t\bar{t}$ and single-top processes while the rest of the samples form the background contribution (scenario II). The small fraction of b-jets in QCD, Z+jets and W+jets is neglected.
Table 5.2 summarizes the study for both I and II scenarios where the S/B ratio in

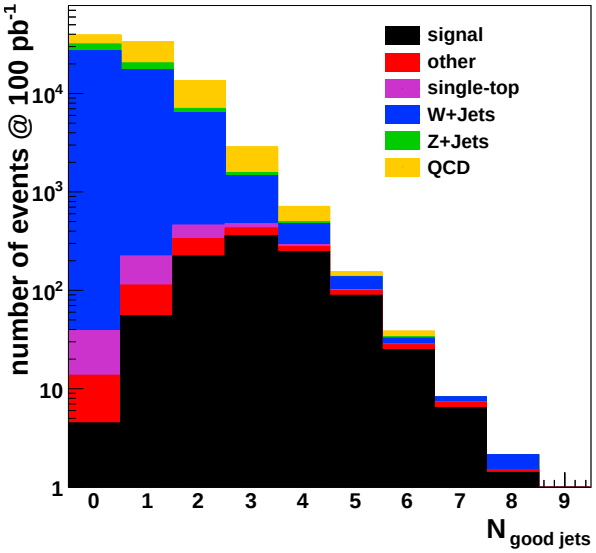
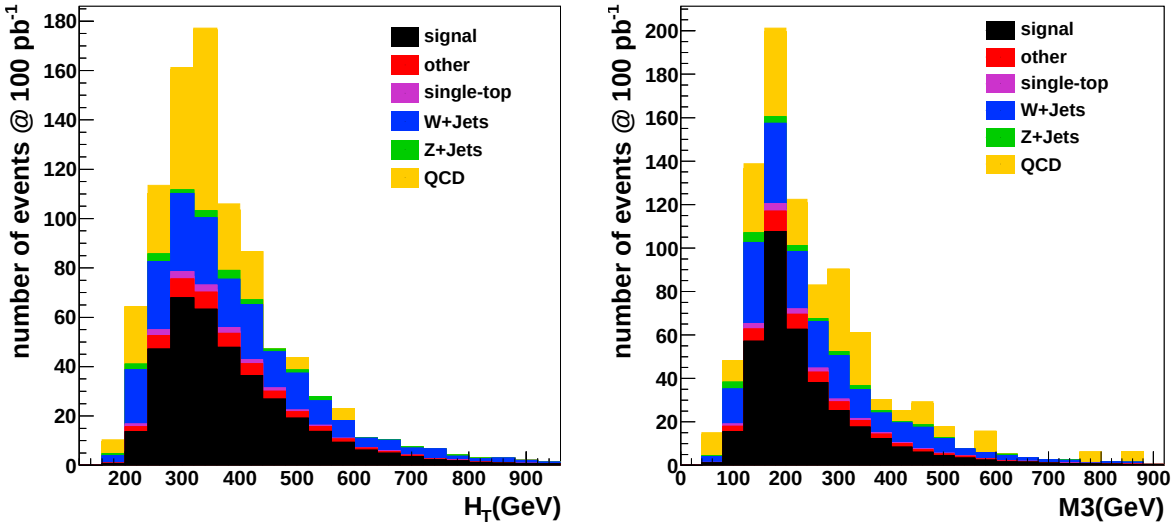


Figure 5.9: The number of selected jets in the event within the $t\bar{t}$ and different background processes. The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.

each case is calculated after asking four leading jets to meet the relevant criteria. It seems that the f_{HPD} and $N90_{hits}$ cuts are too loose and have almost no effect on the S/B ratio at least for the simulated samples. The effect of $f_{em} < 0.9$ request is more recognizable for scenario II since this cut mainly reduces the Z+jets contribution as mentioned. The cut on $N_{CaloTowers}$ helps to keep more qualified jets at higher energies. Finally, events are selected if their four leading jets with $p_T^{corr} > 30\,\text{GeV}$ and $|\eta| < 2.4$ pass the $N_{CaloTowers} > 4$ and $0.05 < f_{em} < 0.9$ requirements. It leads to a S/B of 1.17 (1.71) for scenario I (II).
Figure 5.9 shows the number of selected jets for different processes. It can be seen that requiring $N^{jets}_{selected} \geq 4$ suppresses most of the challenging W+jets contribution.

Identification cut	No Cut	$f_{HPD} < 0.98$	$N_{CaloTowers} > 4$	$N90_{hits} > 1$	$f_{em} < 0.9$
S/B, scenario I	1.01	1.01	1.14	1.02	1.15
S/B, scenario II	1.51	1.52	1.66	1.52	1.70

Table 5.2: The effect of different individual jet identification cuts on the S/B ratio in $100\,pb^{-1}$ integrated luminosity for two definitions of the signal and background, I and II.



(a) (b)

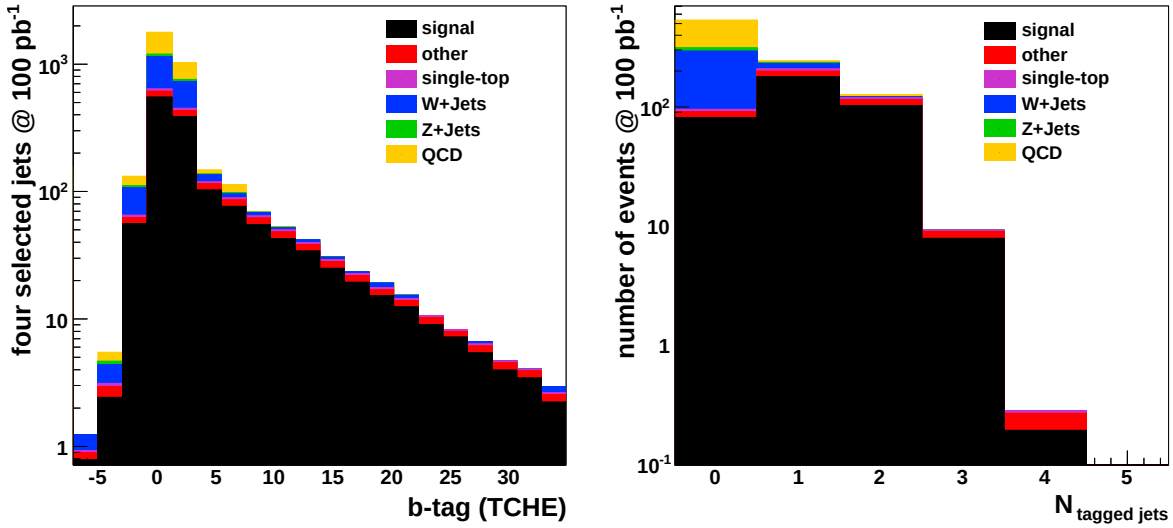
Figure 5.10: The H_T (a) and the $M3$ (b) distributions for the $t\bar{t}$ and different background processes after asking for at least four well-identified jets. The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.

The H_T variable which is the scalar sum of the p_T of the four leading jets together with the $M3$ variable are plotted in Figure 5.10. The $M3$ quantity is the invariant mass of the vectorial sum of the three out of the four leading jets that are giving the highest p_T among all possible combinations. This variable is an estimator for the mass of the hadronically decayed top quark. The cut flow of the event selection is summarized in Table 5.3.

The final row is the result of the same selection but without the ECAL driven criterion for the prompt electron. It seems that the seeding information can suppress part of those QCD multi-jets events that even survive the jet selection. At the end, the number of W+jets events is high with respect to the number of $t\bar{t}$ (signal). In the following, it is investigated how to deal with this background process in an efficient way.
Dedicated simulation/filtering has been done for the processes like Wc and Vqq (see Section 3.4 for definitions) where the samples have overlap with the W(Z)+jets samples. Hence, to avoid the event double conting these samples are not used in the analysis while their expected number of events in different event selection steps are presented in Table 5.4.

5.1.4 Possibilities to suppress the W+jets background

A powerful tool to eliminate the remaining backgrounds after the final event selection is to ask for the existence of at least one b-jet among the four leading jets. This request has a significant effect since the jets in the W+jets and QCD multi-jet processes are mostly originated from the gluons or the quarks other than b-quarks. Figure 5.11 shows



(a) (b)

Figure 5.11: Within the selecetd events according to Table 5.3, the b-tag discriminator distribution of the Track Counting High Efficiency b-tagging algorithm (a) of the four leading jets and the number of jets recognized as b-jets accordingly for $d_{TCHE} > 4$ (b) are shown. The distributions are normalized to $100\,pb^{-1}$ integrated luminosity.

	$t\bar{t}$ (signal)	$t\bar{t}$ (others)	single-top	W+jets	Z+jets	QCD
initial	2330	1.342×10^4	3280	3.131×10^6	3.048×10^5	5.608×10^8
pre-selection	1915	4520	1148	6.369×10^5	8.912×10^4	3.560×10^7
exactly one prompt electron	1124	548.8	454.2	3.133×10^5	3.366×10^4	7.485×10^4
conversion veto	1102	536.6	444.4	3.064×10^5	3.276×10^4	4.806×10^4
ECAL driven request	1101	535.4	443.9	3.062×10^5	3.272×10^4	4.580×10^4
second electron veto	1065	473.1	432.3	3.060×10^5	2.063×10^4	4.563×10^4
muon veto	1065	342.5	422.9	3.059×10^5	2.059×10^4	4.555×10^4
at least 4 qualified jets	373	45.94	17.74	228.4	25.03	131.6
at least 4 qualified jets with no ECAL driven demand	373	45.94	17.74	228.4	25.03	160.1

Table 5.3: The cut flow table for different processes contributing to the analysis. The ECAL driven requirement on the electron candidate can reduce the number QCD multi-jets surviving even the jet selection by 18%. Additional treatment might be needed for the W+jets background which is at the same order of the signal at the end. The numbers are normalized to $100\,pb^{-1}$ integrated luminosity.

	Vqq	Wc
initial	6500	6.06×10^4
pre-selection	1969	6792
exactly one prompt electron	895.9	3903
conversion veto	871.2	3816
ECAL driven request	870.3	3813
second electron veto	764.2	3800
muon veto	754.3	3781
at least 4 qualified jets	15.65	16.44

Table 5.4: The cut flow table for the Vqq and Wc processes which are not used in the analysis to avoid the possible event double counting. The numbers are normalized to $100\,pb^{-1}$ integrated luminosity.

the b-tag discriminator distribution for the Track Counting High Efficiency b-tagging algorithm (see Section 4.4.1), together with the number of jets tagged as b-jet accordingly. The jet is recognized as b-jet if the algorithm discriminator exceeding the value of four, $d_{TCHE} > 4$.
From the distribution (a), it is visible that both W+jet and QCD multi-jet events are accumulated in the negative and low values of the discriminator. This behavior is expected for non-b-jets as explained in Section 4.4.1. Therefore with "at least one

	$t\bar{t}$ (signal)	$t\bar{t}$ (others)	single-top	W+jets	Z+jets	QCD
at least one b-jet	291.1	36.18	12.65	26.72	4.5	9.84

Table 5.5: Number of events after requesting for at least one b-jet among the four leading jets for different processes contributing the analysis. The numbers are normalized to $100\,pb^{-1}$ integrated luminosity.

b-jet" requirement most of these backgrounds are rejected (distribution (b) in the same figure). The impact of the b-jet selection on the other processes can be found in Table 5.5. The S/B ratio (scenario I) is enhanced by a factor of 4 after b-tagging.
Despite of the promising effect of b-tagging on the background rejection, this requirement can make the final analysis complicated. Since the ultimate goal is to measure the b-tagging efficiency, one need to be careful in preparing the b-candidate jet sample to not include the tagged jets. Otherwise, the efficiency measurement can be biased. Therefore it is useful to look for alternative variables with a similar rejection power.
A long list of kinematic variables and variables describing the topology of the event has been studied to find properties with a similar discriminating power against the background events while keeping a reasonable amount of the signal events. In all cases, a cut values for the variable is chosen equating the signal sample efficiency with the efficiency achieved by b-tagging, i.e. $\epsilon_{signal} \sim 78\%$. The efficiency of the W+jets background for this cut value is compared to the ϵ_{W+jets} of the b-tagging which is about 12%.
As an example, the result for the H_T variable is shown in Table 5.6. The cut on the H_T is fixed to 293 GeV for which the same signal efficiency as b-tagging is achieved. The efficiency of the W+jets background is however $\sim$ 73%, means the signal and background events are similarly rejected.

Although the studied variables other than b-tagging do not succeed in effectively rejecting the W+jets background and hence will not be used in the event selection, further studies presented in Section 5.3.3 confirm that this background contamination can be dealt with in another way. It needs to be mentioned that with further refinements during the analysis, the QCD multi-jets contribution does not have a big influence either. Therefore, all events surviving the jet selection are used in the final analysis.

	$t\bar{t}$ (signal)	$t\bar{t}$ (others)	single-top	W+jets	Z+jets	QCD
$H_T > 293\,\text{GeV}$	290.5	35.88	12.94	167.1	17.44	87.14

Table 5.6: An example of the event properties investigated to be alternatively used instead od b-tagging: Number of events after applying $H_T > 293$ selection for different processes contributing the analysis. The numbers are normalized to $100\,pb^{-1}$ integrated luminosity.

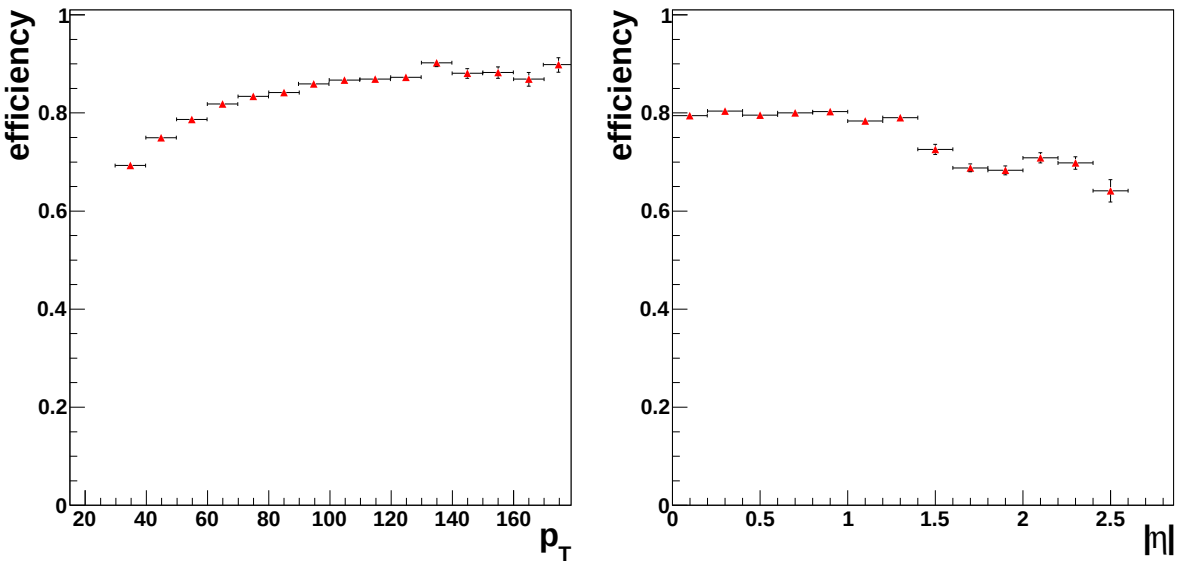
5.1.5 The selection performance for the signal $t\bar{t}$ sample

About 16% of signal events survive the whole selection chain. However, not all of the objects by which the event is selected represent the true physics objects at the generator level. The prompt electron and the four leading jets in the event make the event topology and are the inputs for the rest of the analysis. Hence it is worth investigating if each of these selected objects are truly coming from the $t\bar{t}$ decay in the semi-electron decay mode.

- **Electron selection performance:** This subject is studied in detail in Section 4.2 with slightly different selection criteria. The same approach with the adjusted selection requirements is used here to find the electron selection performance. Events entering this study have survived the preselection criteria as explained in Section 5.1.1.
  The efficiency, ϵ_e, is defined as the fraction of the true electrons[3] with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$ (gap excluded) which pass the rest of the electron selection criteria including the track partner veto and the ECAL driven seeding requirements. The probability for a true electron to be selected is on average $\sim 78\%$. This number seems consistent with what is presented in Section 4.2 if one considers the independence of isolation and identification efficiency, i.e. $\epsilon(id, iso) = \epsilon_{id}.\epsilon_{iso}$, as well as the effect of the selection based on the seeding and conversion rejection. As illustrated in Figure 5.12 (a) ϵ_e increases as a function of p_T^e. From Figure 5.12 (b) one can deduce that the electron selection is more performant in the barrel than in the endcap. The electron selection results in a sample with more than 99% purity where the purity is the number of true electrons divided by the total number of electrons in the selected $t\bar{t}$ sample.

- **Jet selection performance:** The jet selection performance can be spoiled mainly due to presence of the radiation jets. For the final state radiations, $q \to q'g$, the quark after radiation, q', and the associated gluon are either included in the same jet or end up in two jets with the directions usually different

[3] Electron candidates matched with the generated electron from W boson decay. The matching is defined as having a separation less than 0.3 in the η-ϕ plane.



(a) (b)

Figure 5.12: The electron selection efficiency as a function of p_T (a) and $|\eta|$ (b) the electrons in the signal $t\bar{t}$ sample. The selection is applied on the electrons with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$ (gap excluded).

than the original quark, q. This usually results in jets with lower p_T. The initial state radiations can be on the other hand hard enough to be reconstructed as one of the four leading jets. Figure 5.13 shows the p_T distribution of four quarks initiated from the $t\bar{t}$ decay together with the initial state radiated partons. It seems that the radiated partons are most of the time harder than the softest quark.

Now the question is that how well the kinematic and the jet identification requirements are able to pick events with four leading jets created by the quarks in the semi-electron mode of the $t\bar{t}$ decay. It is in particular interesting to know this performance on those signal events that already passed the electron selection and the lepton veto.

The four leading jets in each event are asked for the $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ before they are divided into quark-jets and radiation jets [4]. Events are rejected if any of the four leading jets fail the kinematic requirements. This already reduces the size of the signal sample by $\sim$45%.

The quark-jets are defined as the jets matched with any type of quarks produced in the semi-electron $t\bar{t}$ decay. The only parameter for matching is the jet-quark distance in η-ϕ plane. In the first step of matching, the quarks are ordered by p_T. Then the jet having the minimum separation from the first parton is assigned to it and removed from the jet list if $\Delta R_{min} < 0.3$. The procedure continues with

[4] Due to the jet-electron cleaning, the fraction of fake jets in the jet sample is negligible; $f_{fake\,jets} \approx 2 \times 10^{-5}$.

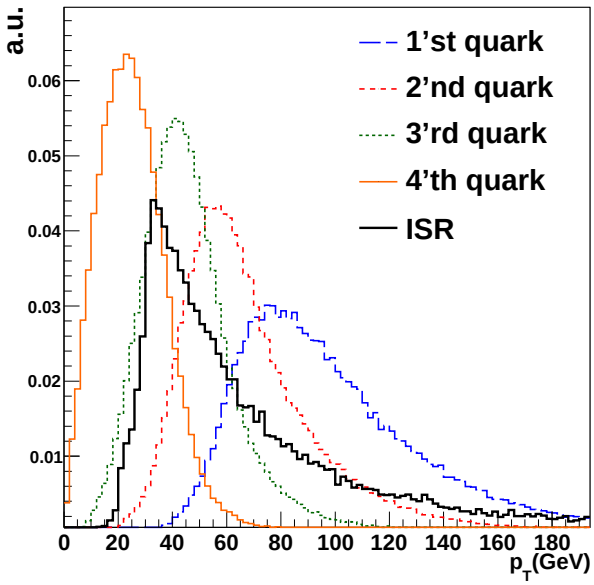


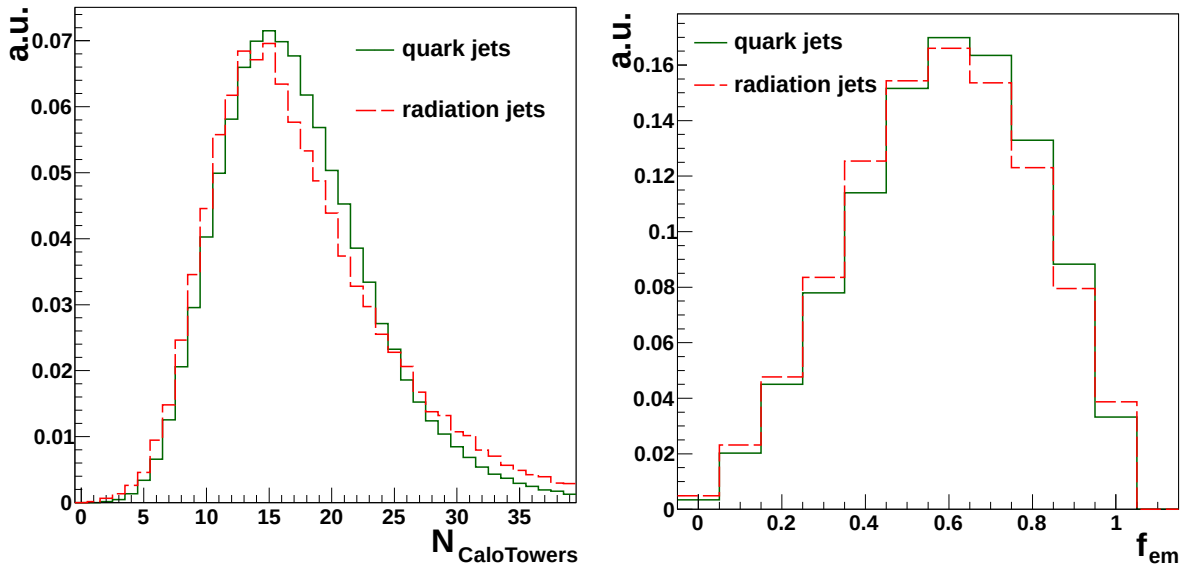
Figure 5.13: The transverse momentum of the four quarks generated in the semi-electron mode of the $t\bar{t}$ decay together with the p_T of the radiated partons in ISR processes in the same events.

the next quark in the parton list until all partons are associated to a jet.
The two identification parameters, the number of CaloTowers and the electromagnetic energy fraction, are compared in Figure 5.14 for the two jet categories where a very similar behavior is observed for the jets in both groups.

According to what is shown in Figure 5.14, the identification cuts will not discriminate between the radiation jets and the quark-jets. The inclusive identification efficiency is about 80% for both. Thus for the $t\bar{t}$ jet sample in which the jets are required to have $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$, the radiation content ($\frac{N_{rad\,jets}}{N_{tot\,jets}}$) which is about 25% does not change after applying the jet identification requirements.

5.2 The event topology reconstruction

The b-tagging efficiency has a general definition, given in Equation 4.20. This definition implies that the efficiency has to be measured on jets whose origin is a b-quark. While in the simulation based analyses it is always possible to check the origin of a jet, on data it is not possible. For the data driven ϵ_b measurement methods introduced in Section 4.4.5, an important subject is to find the true b-jets via indirect ways.
The same challenge exists for the evaluation of ϵ_b within top quark events. Since the b-tagging algorithms exploit the b-jet features as explained in Section 4.4, the b-jet finding is to be performed without making use of these b-jet characteristics to avoid a possible bias on the ϵ_b measurement. Hence the solution lies in the kinematics and the topology of the event.
The $t\bar{t}$ event contains one b-jet per top quark in the final state, assuming $|V_{tb}| = 1$.



(a) (b)

Figure 5.14: The number of CaloTowers (a) and the electromagnetic energy fraction (b) of the four leading jets reconstructed in the semi-electron $t\bar{t}$ final state. Jets are divided in two categories based on their origin.

In the semi-electron channel, there are two non-b-jets from a W boson decay. Hence the invariant mass of this di-jet system is expected to be close to the world average m_W (or the value introduced in the simulation). If one of the b-jets has the same top quark origin as the two non-b-jets, the 3-jet combination would give an invariant mass near the world average m_{top} (or the value used for simulation). The difference between the invariant mass of the tri(di)-jet combination and the mass of top quark (W boson) would have a clear meaning if it is compared to the expected resoultion $\sigma_{m_{top}}$ (σ_{m_W}) on the top quark (W boson) mass. So one can define an estimator to evaluate the correctness of the reconstructed top quark hypothesis as:

$$\chi^2 = \left(\frac{m_{j1j2} - m_W}{\sigma_{m_W}}\right)^2 + \left(\frac{m_{j1j2j3} - m_{top}}{\sigma_{m_{top}}}\right)^2 . \tag{5.1}$$

Among the 12 possibilities for the 3-jet combinations out of four leading jets, the one which minimizes the χ^2 in Equation 5.1 is assumed to match the hadronically decayed top quark. The remaining jet is supposedly coming from the leptonic side and is taken as the b-jet candidate. Therefore a jet sample, enriched in b-content, can be formed without using directly the b-jet characteristics.
To obtain the values for the mass constraints and resolutions in Equation 5.1, the four leading jets are matched with partons in the simulated sample of semi-electron $t\bar{t}$ events. The W boson of the hadronic side of the event is reconstructed with the two jets matched to quarks produced in the true W boson decay. A jet matched to a b-quark is combined with the reconstructed W boson if they both come from the same top quark decay.

The invariant mass distributions of the reconstructed top quark and W boson are fitted to a Breit-Wigner function which is convoluted with a Gaussian to account for the random detector effects:

$$(BW * G)(x) \equiv \int_{\text{Range}} BW(x-y)G(y)\, dy. \tag{5.2}$$

In Equation 5.2, G and BW denote the Gaussian and the Breit-Wigner functions, respectively. The mass distributions and the fitted functions together with the fit pa-

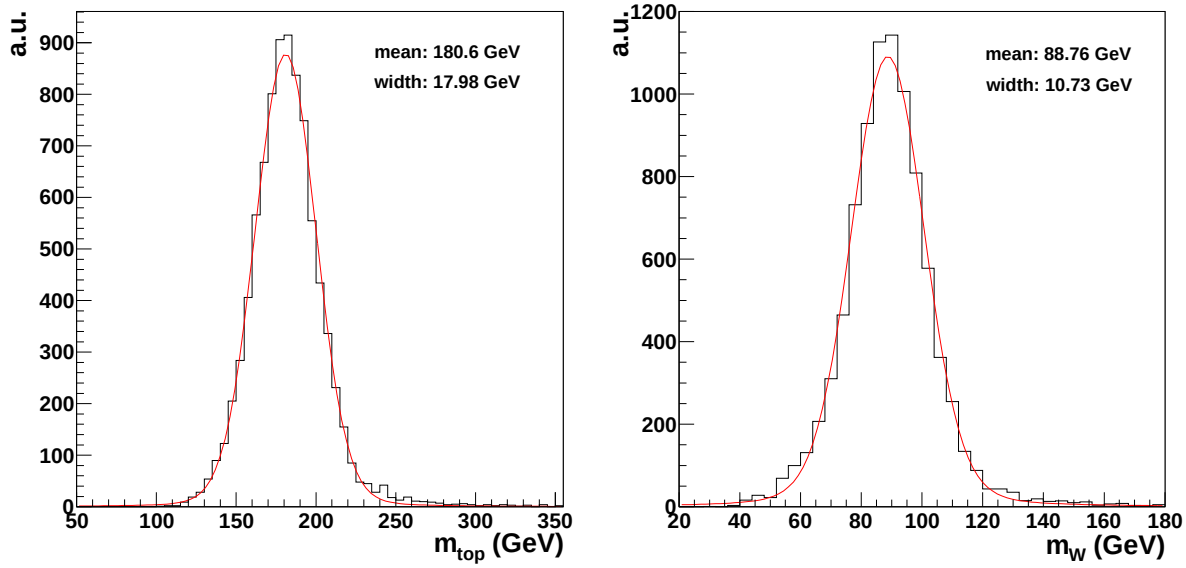
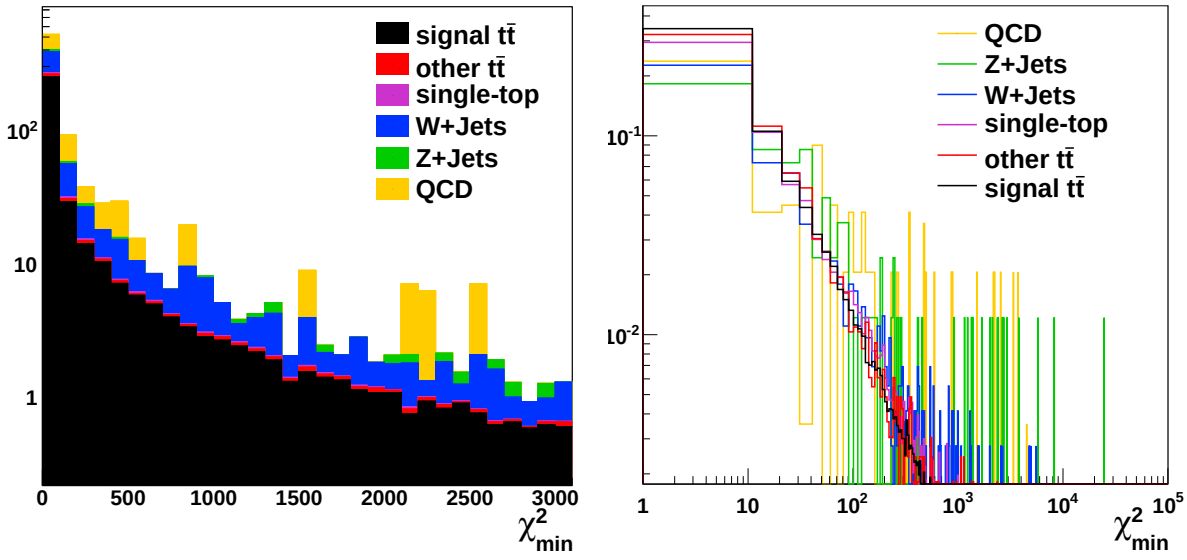


Figure 5.15: The mass distribution of the reconstructed top quark (left) and W boson (right) using the four leading jets matched with the quarks present in semi-electron final state of the $t\bar{t}$ events. The fit function is a Breit-Wigner convoluted with a Gaussian as in Equation 5.2. The values indicated on the plots are the mean and the width of the convoluted function, used for the constraints in Equation 5.1.

rameters are shown in Figure 5.15. Instead of the world average for the mass and width, the fitted values of the mean (mass) and width (resolution) of the convluted function are taken for both top quark and W boson.
Both mass and width depend on the jet energy scale correction. Although the jets in this analysis are calibrated with the Level 2 and Level 3 correction factors, they do not reproduce the world average values for the top quark and W boson mass. Hence for the sake of consistency within the method, this approach is preferred in the simulation based analysis and even in the data analyses since the jet energy scale calibrations are not perfect.
The χ^2_{min} distribution for the signal sample and different background processes other than QCD is illustrated in Figure 5.16 (a) for an integrated luminosity of $100\, pb^{-1}$. The lower values are mostly occupied by the signal process as expected.
Normalized to unity in Figure 5.16 (b), the shapes of the χ^2_{min} distribution for the



(a) (b)

Figure 5.16: The χ^2_{min} distribution for the signal and background processes: all samples normalized to $100\,pb^{-1}$ integrated luminosity (a); all samples normalized to unity for the shape comparison (b).

physics processes present in the analysis are compared. While in the W(Z)+jets and the QCD multi-jets events the χ^2_{min} values are distributed almost randomly, the processes containing a top quark follow similar shapes. Consequently, the tail is longer for processes with no top quark produced in the hard interaction.
Although the χ^2_{min} distribution shows a more reasonable behavior for the signal events compared to background events, there is still a large tail for the jet combinations in the semi-electron final state of $t\bar{t}$ which can prevent the χ^2-method to work properly. The source and the influence of such combinations will be discussed in the context of the performance of the topology reconstruction.

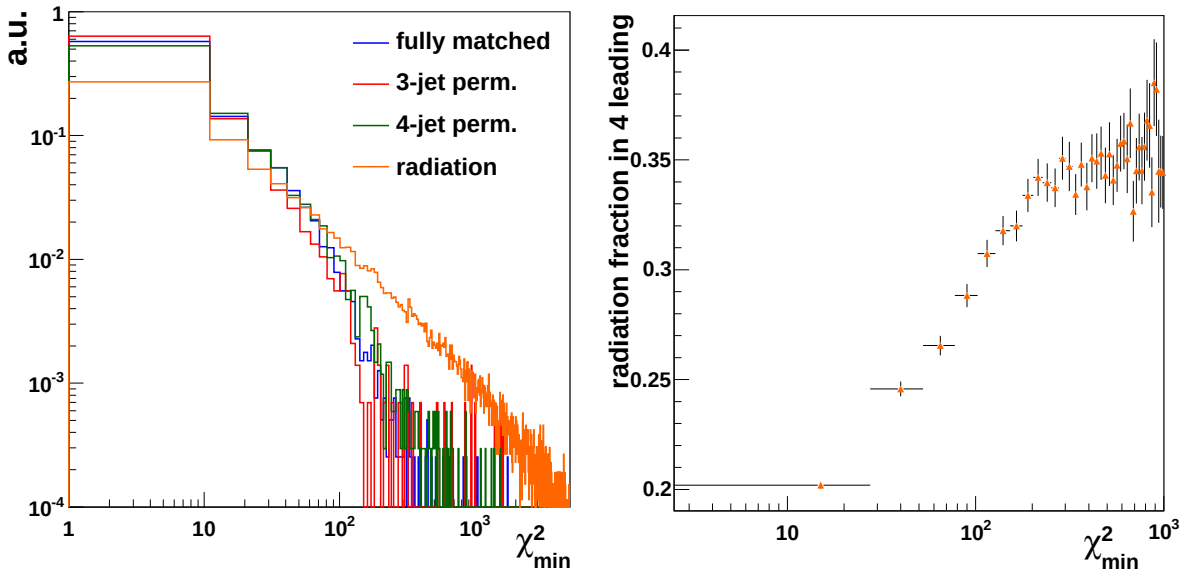
5.2.1 The performance of the topology reconstruction

The performance of the topology reconstruction can be studied from different aspects. One can check the probability of representing the true hadronically decayed top quark for the jet combination with the χ^2_{min}. Following the main goal of the analysis which is the b-tagging efficiency measurement, it is also interesting to know the probability for the remaining jet out of the four leading jets to be originated from a b-quark. While the first approach is relevant for the processes containing top quark, the second can be extended to any type of events containing jets. In the current study, the first aspect of the topology reconstruction performance is investigated within the semi-electron $t\bar{t}$ events where the second one is looked at in both signal and background processes.

Investigating the χ^2_{min}-method in the signal sample

Within the signal sample one would not expect the large tails for the χ^2_{min} distribution as what was shown in Figure 5.16. However, as discussed in Section 5.1.5, the probability of a radiation jet to appear among the four leading jets even after the whole selection is ∼25%. The jets can therefore be misassociated to the W boson and/or top quark, due to the presence of the radiation jets even in the best combination. This results in large χ^2_{min} values and entails the χ^2_{min} distribution.

For those events where no radiation jet is present among the four leading jets, the



(a) (b)

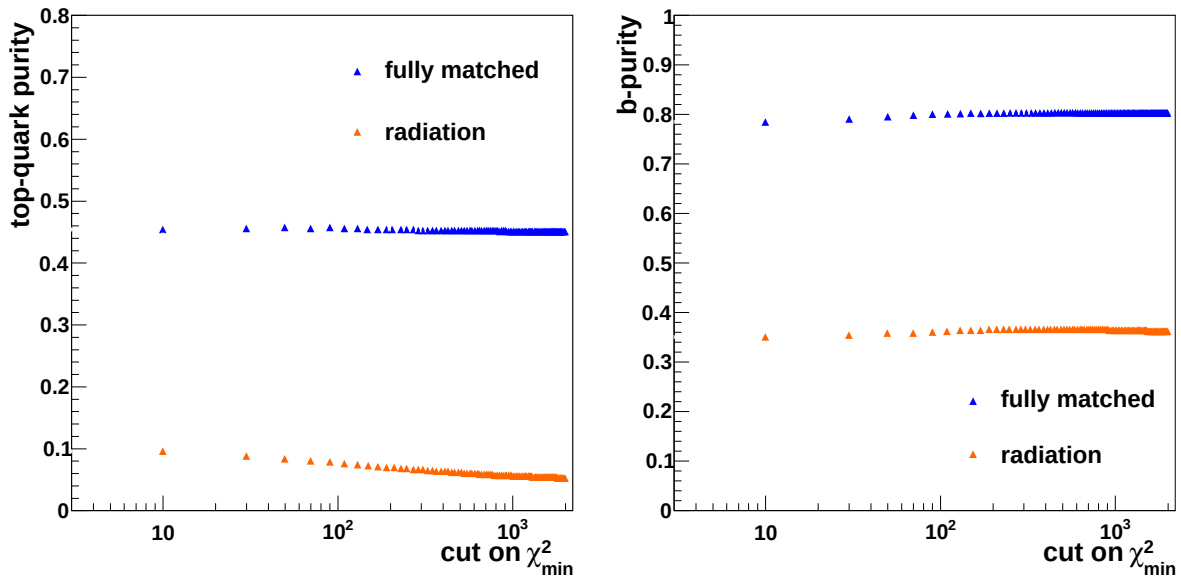
Figure 5.17: The distribution (a) of χ^2_{min} for the signal events with four leading jets matched to the generated quarks and for those with at least one radiation jet among the four leading jets. The former is split based on the correctness of the reconstructed top quark. The radiation fraction (b) among the four leading jets as a function of χ^2_{min}.

misassociation of the quark-jets in the best combination (with the least χ^2 value) can be split into three categories. While in the first category, the category with the "fully matched" combinations, the jets associated to the W boson are matched with the non-b-quarks and the third jet is matched to the b-quark from the hadronic side of the event, in the second category these three jets are allowed to commute among each other ("3-jet permutations"). The third set contains the rest of the possible quark-jet combinations ("4-jet permutations"). The permutation between the jets making the W boson is ignored in all cases. Another category that is complementary to the three sets mentioned is the combinations from events in which "at least" one radiation jet exists among the four leading jets. Figure 5.17 (a) illustrates the distribution of χ^2_{min} for all categories. The effect of the wrong associations is small while the best combinations (those with the minimum χ^2 value) in the radiation category extend the χ^2_{min} distribution toward large values. Beside the distributions, Figure 5.17 (b) shows

the profile of the radiation fraction among the four leading jets, $f_{Rad} = \frac{N_{Rad.Jets}}{N_{4leadings}}$, as a function of χ^2_{min}. The fraction f_{Rad} [5] is on average about 20% at small χ^2_{min} values. It increases as a function of χ^2_{min} where at very large χ^2_{min} values it becomes flat, and on average about 35%.
Despite of the radiation effects, the χ^2 method is indeed performing fine in reconstructing the hadronically decayed top quark. This becomes clearer if one compares the purity obtained by the method to the probability of picking the correct combination through the random selection of three jets out of four. Including all permutations other than the W boson jet components, this random probability is $\frac{1}{12} \approx 8\%$ within the signal events where no radiation jet exists among the four leading jets. The χ^2 method enhances this purity to ∼45%. This is actually the fraction of events with no radiation for which the best combination is "fully matched".
For the events containing "at least one" radiation jet among the four leadings, one can still construct the "fully matched" combination if "only one" radiation jet is present and the rest three jets are coming from the hadronic decay of top quark. For cases with more radiation jets there is no possibility to make the "fully matched" combination. The χ^2 method returns the "fully matched" combination only in less than 10% of the time if at least one radiation jet is found within the four leadings. Applying a cut on the χ^2_{min} value, χ^2_{cut}, and rejecting events for which the best combination has a value of $\chi^2_{min} > \chi^2_{cut}$ can increase a bit the purity for events with radiation while it has almost no effect on the combinations from no-radiation events. Figure 5.18 (a) shows the purity as a function of a cut on the χ^2_{min} value for events with and without radiation jets.
Instead of looking at the reconstructed top quark, one can be interested in looking at the remaining jet, as is already explained. The purity of the jet sample made from the remaining jet, called from now on the leptonic b-candidates, is defined according to the flavor of their original partons. A random selection within the events with no radiation jet among the four leadings results in 50% purity where the outcome of the method has a purity of ∼80%. The method gives lower purity, ∼35%, for the combinations in the events with radiation.

Figure 5.18 (b) illustrates the purity of the b-candidate sample, the b-purity, as a function of a cut on the χ^2_{min} value for both types of events. The different feature with respect to the previous case is the slightly increasing trend of the purity. The reason is that the combinations which fail the "fully matched" requirements have on average higher values of χ^2_{min}. They may however pass the b-purity criterion on the leptonic side which does not care about the jet-quark matching details in the hadronic top sector. In another words, if the χ^2 method selects the jets from the hadronically top quark decay, the remaining jet is the b-quark jet from the leptonic side of the event. Hence the b-purity of the b-candidate jet sample increases. For the same combination on the other hand, the jets associated to the hadronic top quark may be in a wrong permutaion. Therefore the same combination that enhances the b-purity on the leptonic side can fail the "fully matched" criterion on the hadronic side. A similar situation happens when the b-quark jet from the hadronic side of the

[5] Depending on the number of radiation jets among the four leading jets, the radiation fraction will by definition have a value of 1/4, 1/2, 3/4 or 1.



(a) (b)

Figure 5.18: The top quark (a) and the b- (b) purity as a function of a cut on χ^2_{min} for the signal events with four leading jets matched to the generated quarks and for those with at least one radiation jet in the four leading jets.

event remains out of the three jet combination made by means of the χ^2 method. Regarding the b-purity, an upper bound on the χ^2_{min} does not seem necessary.

The χ^2_{min}-method performance in the presence of background processes

From the distribution in Figure 5.16 (b), it can be seen that the χ^2_{min} distribution for the background processes entails to huge values which are not of physics interest. Hence, it is worth investigating how much these non-physical tails influence the b-purity of the jet sample.

Due to the large cross section and the relatively limited statistics available for the simulated QCD multi-jets samples, the events in these samples are given large weights to estimate this background in a data of $100\,pb^{-1}$ of integrated luminosity. The shape of the QCD multi-jet distributions is therefore not reliable with few number of events remaining after the full top quark event selection. In fact, mainly the QCD multi-jets from the tail of the distributions survive the top quark selection. Hence the shapes made by such events have large uncertainties. For these reasons, the QCD multi-jet contribution is not shown for the plots in the rest of the analysis. The effect of QCD multi-jet events on the total systematic uncertainty is evaluated at the end (see Section 5.5).

Figure 5.19 is the purity of the leptonic b-candidate jet sample as a function of a cut on the χ^2_{min} value. Normalized to $100\,pb^{-1}$ of integrated luminosity, the signal and all background processes except the QCD multi-jets, contribute in building the b-candidate

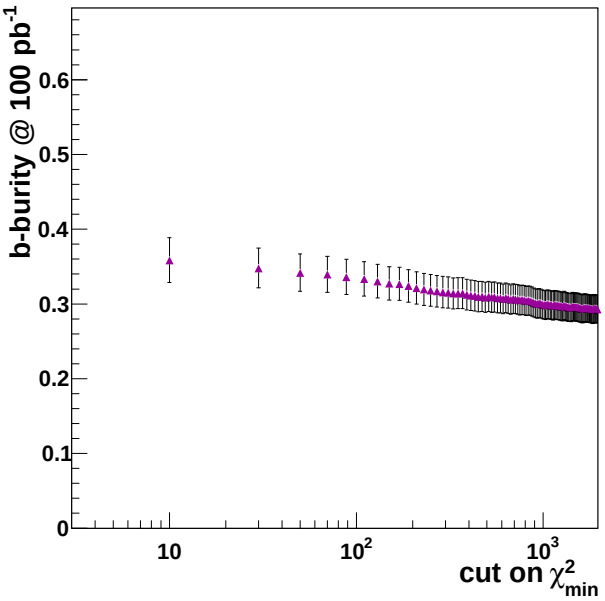


Figure 5.19: The b-purity as a function of a cut on χ^2_{min} for the signal and background processes except the QCD multi-jets. The purity is calculated for $100\,pb^{-1}$ integrated luminosity.

jet sample. The b-purity decreases at large values of χ^2_{min} since there the contamination of the background processes is higher. By the way, the difference in the purity is only about 5% over the whole range where the overall purity is ∼30%. From the b-purity point of view, there can be some motivation to put an upper limit on the χ^2_{min} value in the presence of backgrounds. However as it will be discussed later, only a subset of the b-candidate jet sample will be considered for the efficiency measurement for which the non-b-quark contamination including radiation jets is successfully subtracted. Therefore, no upper limit is now applied on the χ^2_{min} value. The average b-purity is reduced to ∼22% if the QCD contribution is considered. With the QCD multi-jets included, the trend of the b-purity versus the cut on χ^2_{min} remains the same, of course with a big uncertainty.

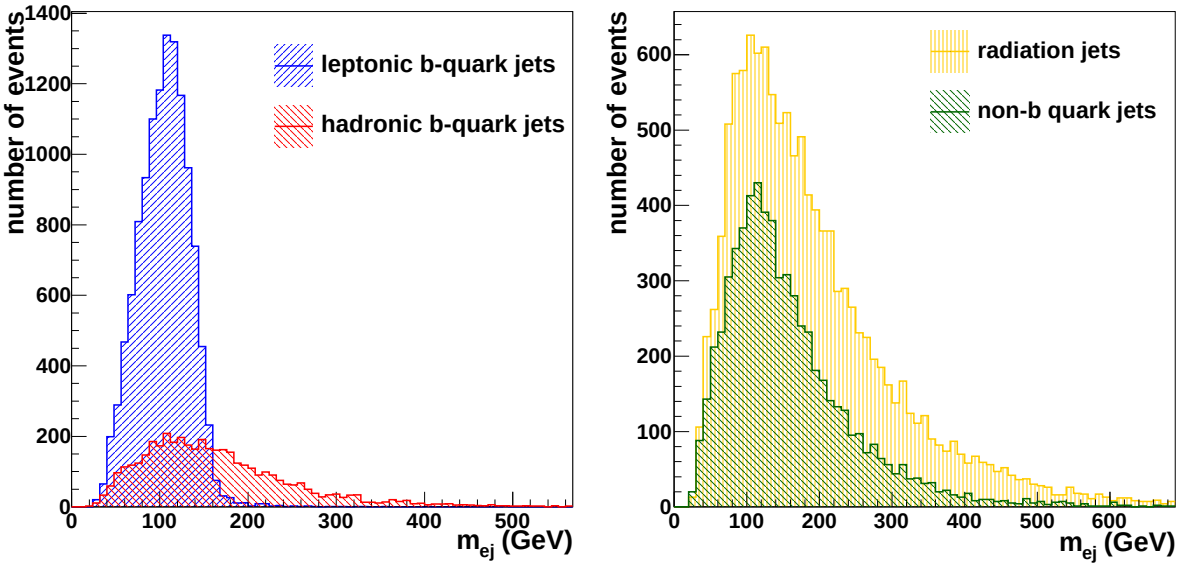
5.3 The b-tagging efficiency estimation

The indirect b-jet selection is achieved by combining three jets in the event within a χ^2 method to reconstruct the hadronically decayed top quark and letting the remaining jet [6], the leptonic b-jet candidate, end up in a so-called leptonic b-candidate jet sample. The b-candidate jet sample in signal events has a b-purity of $\sim$ 35% (radiation included) while its purity changes to $\sim$ 30% in the presence of background processes. The b-candidate sample is not pure enough to serve as the input for the b-tagging efficiency measurement. The kinematic properties of the event are exploited once more, this time to purify the b-candidate jet sample.
The key variable is the invariant mass of the leptonic b-jet candidate and the electron

[6] In the whole analysis the focus is on the four leading jets, ordered by p_T.

from the W boson decay on the leptonic side of the event ($t \to bW \to b\,e\nu_e$), denoted as m_{ej}. Based on the existing correlation between the electron and the b-quark coming from the same top quark and the conditions imposed by the ν_e helicity, a special distribution is expected for the m_{ej} variable. It has a relatively sharp drop somewhere between $m_{ej} = 150\,\text{GeV}$ and $m_{ej} = 160\,\text{GeV}$, with the true b-jets from the leptonically decayed top quark accumulated below this drop. This is derived and explained in detail in Appendix B.

To clarify the idea, Figures 5.20(a,b) demonstrate the m_{ej} variable in the



(a) (b)

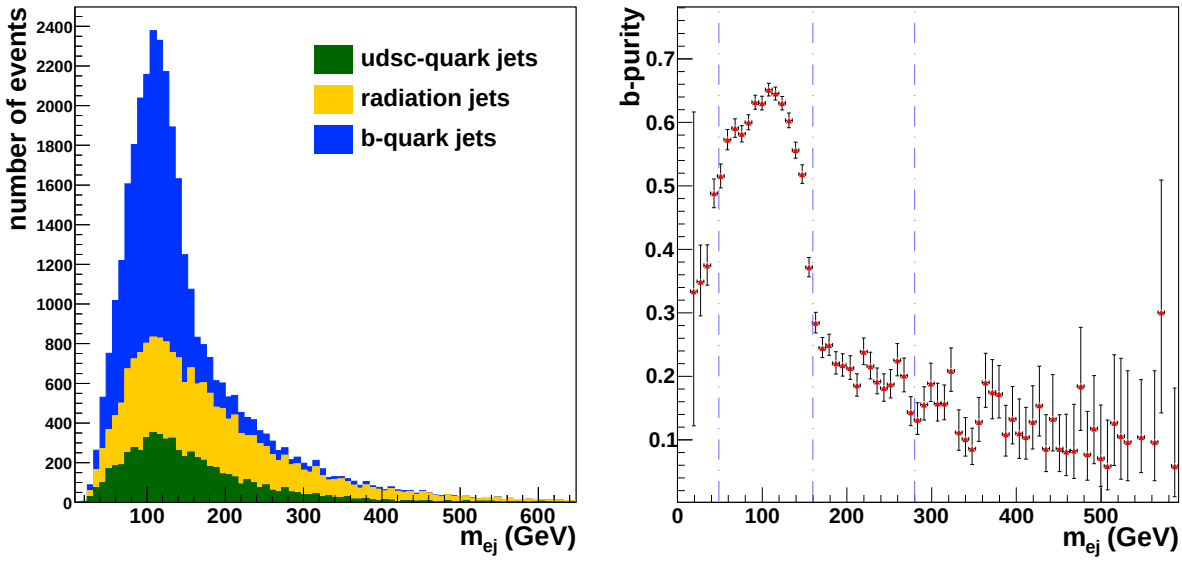
Figure 5.20: The m_{ej} distribution for the b-candidate jet sample in the signal events, the b-quark jet (a) and the non-b-qark jet (b) subsamples.

semi-electron $t\bar{t}$ process where the b-candidate jet sample is divided into the b-quark [7] and non-b-quark jet subsamples. For the b-quark jet subsample, the contribution of b-quark jets from the leptonic and hadronic side of the events are shown separately. In the non-b-quark jet subsample also, the jets are further factorized based on their quark or radiation origin. It can be seen that only the b-quark jets from the leptonic side of the events drop relatively sharply at $150\,\text{GeV} < m_{ej} < 160\,\text{GeV}$ and the other distributions are characterized by their wider bulks and longer tails.
Figure 5.21 (a) shows the m_{ej} distribution for the b-candidate jet sample in the signal events where the contributions from b-quark and non-b-quark jets are shown in different colors. It can be seen that the bulk of the distribution is indeed dominated by the b-quark jets while there are more non-b-quark jets in the tail region. The b-purity as a function of m_{ej} is shown in Figure 5.21 (b) for the signal events, providing the two distinct regions: the **b-dominated**(Left) in the bulk area and the **b-depleted** (Right)

[7] In the rest of the analysis, the jets matched to a b-quark at generator level are called b-quark jets.

in the tail region.

In this analysis, the 50 GeV$< m_{ej} <$ 160 GeV range is taken for the b-dominated part



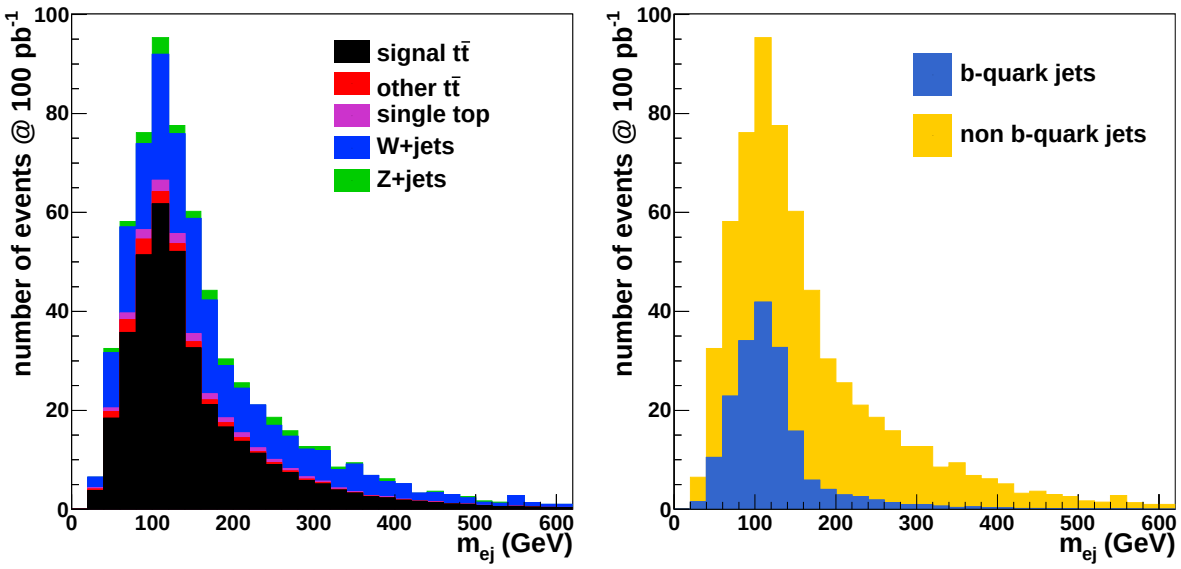
(a) (b)

Figure 5.21: The m_{ej} distribution for the b-candidate jet sample in the signal events, (a), separated according to jets origin. The b-purity of the same jet sample as a function of m_{ej}, (b), indicating the b-dominated and the b-depleted regions.

where the 160 GeV$< m_{ej} <$ 280 GeV interval is considered as the b-depleted region. The b-purity is ∼59% in the b-dominated area which is higher than 35%, the average purity of the whole b-candidate jet sample. Lower than the average purity, the b-purity falls down to ∼22% in the b-depleted part.

Although it might seem reasonable to get focused on the b-dominated part of the jet sample for the b-tagging efficiency measurement, one needs to deal with the non-b-quark jet contamination in this area. Handling the non-b-quark jet contamination in the b-dominated jet sample which will be discussed in the next section, is more crucial if the background processes are considered since they introduce more impurities in the b-candidate jet sample.

Normalized to 100 pb^{-1} of integrated luminosity, Figure 5.22 (a) shows the m_{ej} distribution for the signal and the background processes. The contribution of the W+jets process that contains mostly non-b-quark jets is considerable as expected from the event selection results. These events are distributed almost everywhere and enlarge the tail at high m_{ej} values. The same distribution, when categorizing the jets according to their origin is shown in Figure 5.22 (b). The b-quark jet content in the b-dominated area (∼ 39%) is still larger than the average purity (∼ 30%) and is more than three times higher than the purity in the b-depleted region (∼ 11%).



(a) (b)

Figure 5.22: The m_{ej} distribution for the b-candidate jet sample in the signal and background events, (a). The same distribution factorizing the jets according to their origin, (b). Distributions are normalized to 100 pb^{-1} integrated luminosity.

5.3.1 Purifying the b-candidate sample and the first results

The estimation of the b-tagging efficiency, ϵ_b, relies in practice on the *"shape"* of the b-discriminator distribution of the b-quark jets, since the efficiency ϵ_b is basically obtained by calculating the *"fraction"* of the b-quark jets passing a cut on the b-discriminator value. Within the b-candidate jet sample and for a given b-tagging algorithm, the distribution of the b-discriminator is the superposition of the b-tag distributions for b- and non-b-quark jets in the sample. Hence to extract the b-discriminant shape of b-quark jets, one needs to subtract the non-b-quark jets part of the distribution,

$$\Delta_b^S = \Delta_{all}^S - \Delta_{non-b}^S, \tag{5.3}$$

where Δ_x^S stands for the b-tag distribution of x-jets in the S jet sample and the subtraction is performed bin by bin.
Since one is interested to estimate the b-tagging efficiency using a b-dominated jet sample, S in Equation 5.3 is basically corresponding to a jet sample dominated by b-quark jets. The idea of splitting the b-candidate jet sample into two parts based on the m_{ej} value, is very helpful here. First becuase the b-dominated jet sample is formed using the jets in the Left region. Hence, one can rewrite Equation 5.3 for $S = L$ where L denotes the jet sample in the Left area,

$$\Delta_b^L = \Delta_{all}^L - \Delta_{non-b}^L. \tag{5.4}$$

Moreover, if there is no strong correlation between the m_{ej} variable and the value of the b-discriminant, the *"shape"* of Δ_{non-b} in general remains stable from the Right to the

Left m_{ej} region. Therefore, Equation 5.3 can be written for the b-dominated jet sample where the *"shape"* of the b-discriminator for the non-b-quark jets, Δ^L_{non-b}, is obtained from the jets in the b-depleted sample, Δ^R_{non-b}. Regarding the different statistics in the Left and Right regions, the Δ^R_{non-b} needs to be scaled to match the Δ^L_{non-b}. Hence Equation 5.3 takes the form of

$$\widehat{\Delta}^L_b = \Delta^L_{all} - F \cdot \Delta^R_{non-b}, \tag{5.5}$$

where the scale factor, F, is defined as

$$F = \frac{N^L_{non-b}}{N^R_{non-b}}. \tag{5.6}$$

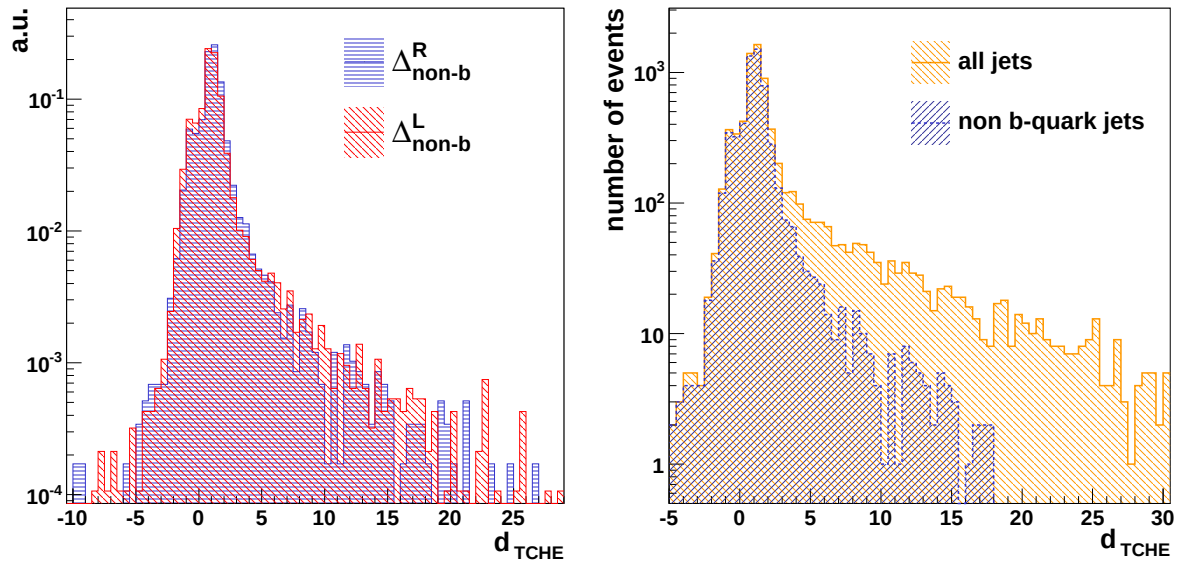
The $\widehat{\Delta}^L_b$ notation is used to emphasize that what is obtained from Equation 5.5 is an estimator for Δ^L_b.
Finally, the b-tag distribution of non-b-quark jets in the b-depleted sample, Δ^R_{non-b}, can be approximated by Δ^R_{all} regarding the low b-purity in the b-depleted jet sample. Therefore, the Δ^L_b distribution can finally be estimated as

$$\widehat{\Delta}^L_b = \Delta^L_{all} - F \cdot \Delta^R_{all}. \tag{5.7}$$

Figure 5.23 (a) demonstrates how the discriminator of the Track Counting High Efficiency algorithm [8] (TCHE) evolves for non-b-quark jets from the Right to the Left region. The b-discriminator shape for non-b-quark jets is stable over the desired range of the m_{ej} values, indicating that the m_{ej} and the b-discriminant are indeed not too correlated. Figure 5.23 (b) shows what to be neglected to make the approximation of $\Delta^R_{non-b} \approx \Delta^R_{all}$ valid. This approximation is in fact driven by the desire to make this part of the method independent from the generator level information. In real data, the origin of the jets is not known and it is not possible to extract the Δ^R_{non-b} distribution. The shape of these two distributions (Δ^R_{non-b} and Δ^R_{all}) are however not identical as it can be seen in the figure. While the bulk of Δ^R_{non-b} distribution is well described by this approximation, the shapes differ in the tail. The difference comes from the contribution of b-quark jets with their high b-tag values. It should be noted that the approximation ignores at most 22% of the jet content in the Right region which is equal to the b-purity in the b-depleted jet sample.

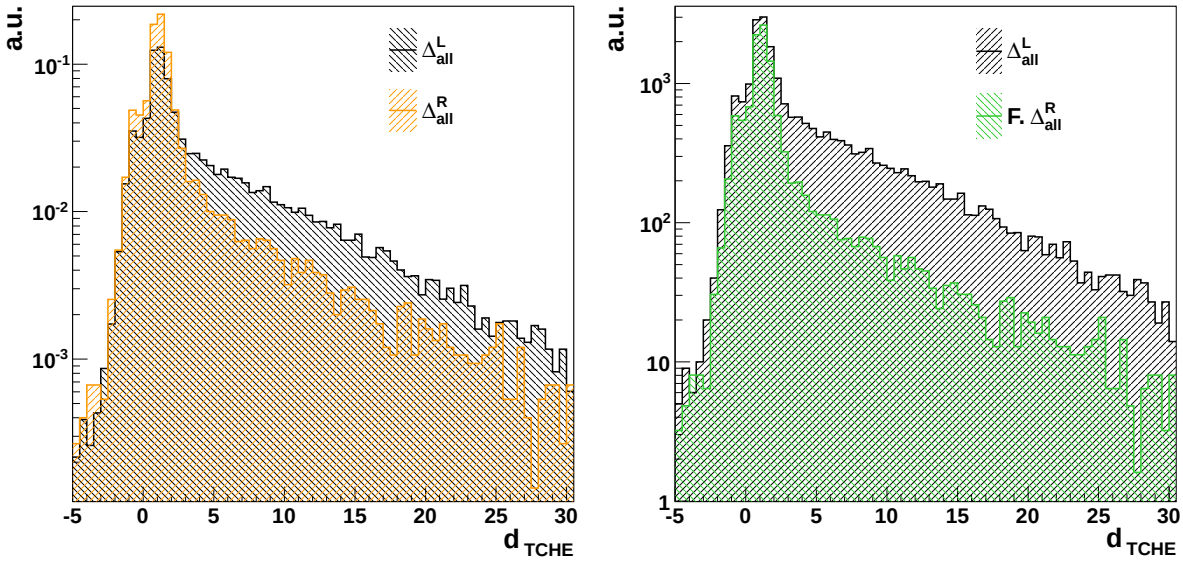
Although the Δ^R_{all} distribution contains b-quark jets, the number of b-quark jets which are removed from the Left region by the Left-Right subtraction is relatively small. In Figure 5.24 (a) the shapes of Δ^R_{all} and Δ^L_{all} distributions are compared. The larger tail in Δ^L_{all} implies the larger b-quark jet content in the Left area, as expected.
To estimate the amount of b-quark jets in the b-dominated jet sample (Left region) which are removed by the subtraction, the Δ^L_{all} and the $F.\Delta^R_{all}$ distributions are shown in Figure 5.24 (b). The scale factor F is evaluated using the information from simulation but it can also be derived from the data itself as detailed in Section 5.3.5. Within the semi-electron final state of $t\bar{t}$ events, $F = 1.607$.

[8] This b-tagging algorithm, denoted as TCHE, is used in the rest of the analysis to present the development of the method.



(a) (b)

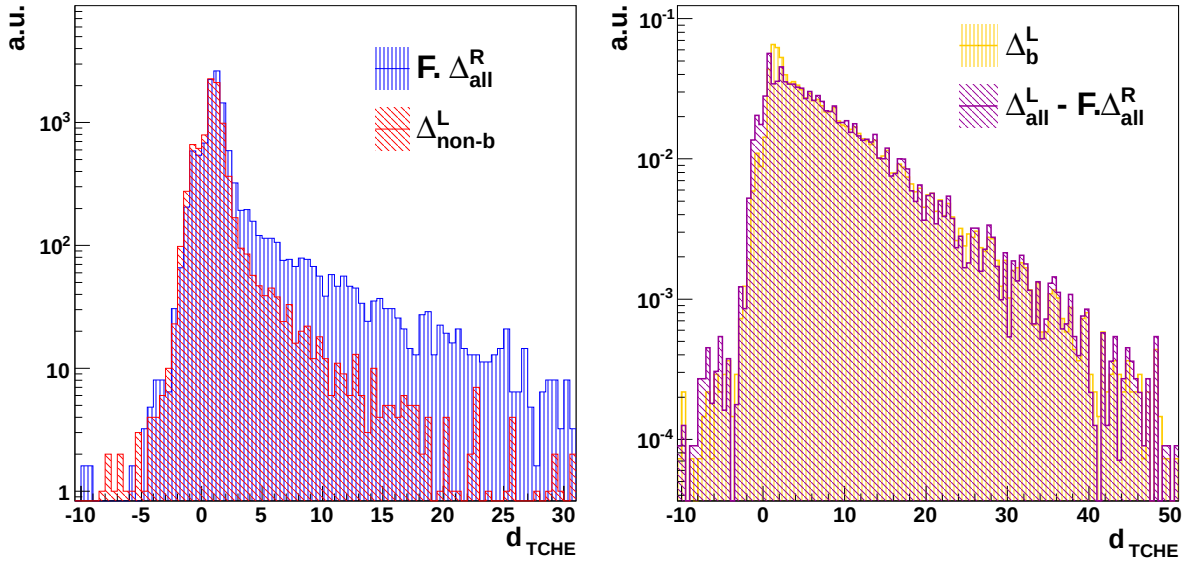
Figure 5.23: The TCHE b-discriminator distribution for the non-b-quark jets in the b-dominated and b-depleted subsamples within the signal events, (a). The shape of the TCHE b-discriminant for non-b-quark jets and all jets in the b-depleted jet sample within the signal events, (b).



(a) (b)

Figure 5.24: The "normalized" TCHE b-discriminator distribution for the all jets in the b-dominated and b-depleted subsamples within the signal events, (a). The TCHE b-discriminant distribution for the b-dominated and the "scaled" b-depleted jet samples within the signal events, (b).

The distributions in Figure 5.24 (b) are not normalized to give an insight about the real amount of b-jets that will be subtracted. According to the figure, there will remain a considerable amount of jets with high b-tag values after the subtraction, $\Delta^L_{all} - F.\Delta^L_{all}$. Figure 5.25 (a) on the other hand, compares the $F \cdot \Delta^R_{all}$ to the b-discriminator distribution of the non-b-quark jets in the b-dominated sample, Δ^L_{non-b}. The use of the scale factor, F, is successful while the difference in the tail could have already been expected because of the approximation made in the b-depleted jet sample, Figure 5.23.
In Figure 5.25 (b) the subtracted distribution of b-discriminant (the right hand side of Equation 5.7) is compared to the b-tag distribution of the b-quark jets (left hand side of Equation 5.7) within the b-dominated jet sample. The high b-tag values in Δ^R_b are well described by the method where at low positive values, the number of the jets are underestimated.
Having the subtracted sample, i.e. the $\widehat{\Delta}^L_b$ distribution, the b-tagging efficiency, $\widehat{\epsilon_b}$, can

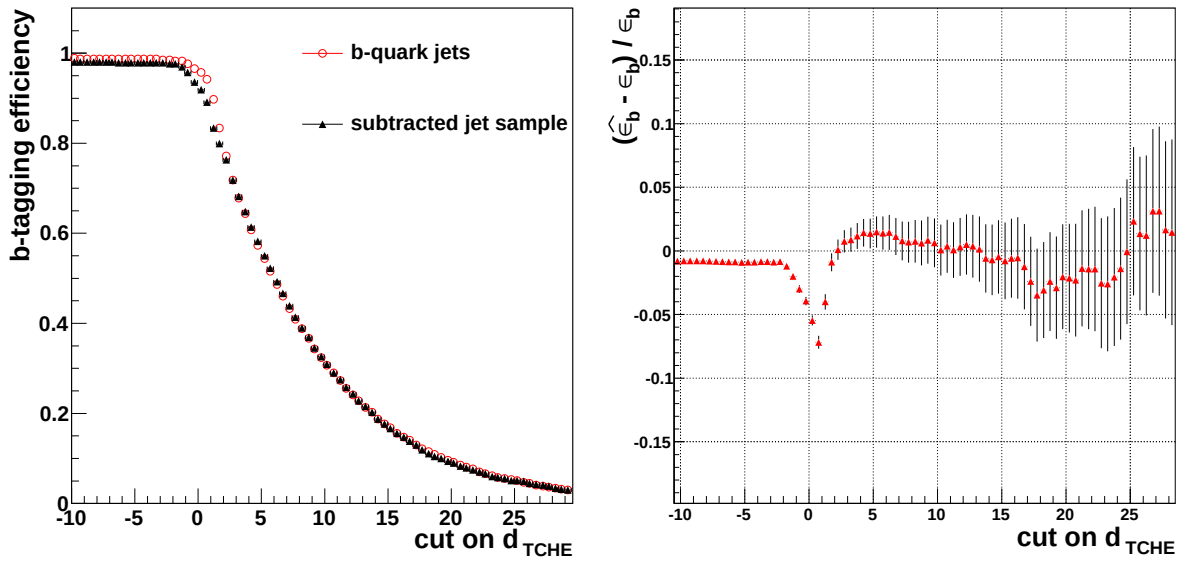


(a) (b)

Figure 5.25: The scaled b-tag distribution of the jets in the Right region, $F \cdot \Delta^R_{all}$, compared to the non-b-quark b-tag distribution in the Left area, Δ^L_{non-b}, (a). The shape of subtracted b-tag distribution according to Equation 5.7 compared to the b-tag distribution shape of the b-quark jets in the b-dominated jet sample, (b).

be estimated as a function of a cut on the b-discriminator value (d_{TCHE}). To evaluate the performance of the method in terms of the b-tagging efficiency, one can compare $\widehat{\epsilon_b}$ with ϵ_b, the b-tagging efficiency of the b-quark jets in the b-dominated jet sample which is derived from the Δ^L_b distribution. A good agreement is observed between the outcome of the method and the true b-jet identification efficiency except at low positive b-tag values, Figure 5.26 (a). The discrepancy is translated into $\Delta\epsilon_b = \frac{\widehat{\epsilon_b} - \epsilon_b}{\epsilon_b}$ in Figure 5.26 (b) to give a picture of the relative difference between the method and the

true efficiency. The bias at the beginning comes from the non-tagable[9] jets with the artificial b-tag value of -100. The contribution of this kind of jets is small, $\sim 1\%$.



(a) (b)

Figure 5.26: The b-tagging efficiency as a function of a cut on the b-discriminator value within the b-dominated jet sample of signal events compared to the efficiency for the b-quark jets in the sample sample, (a). The relative difference between the distributions in (a) to evaluate the discrepancy, (b).

5.3.2 Reconsidering the m_{ej} and the b-discriminant correlation

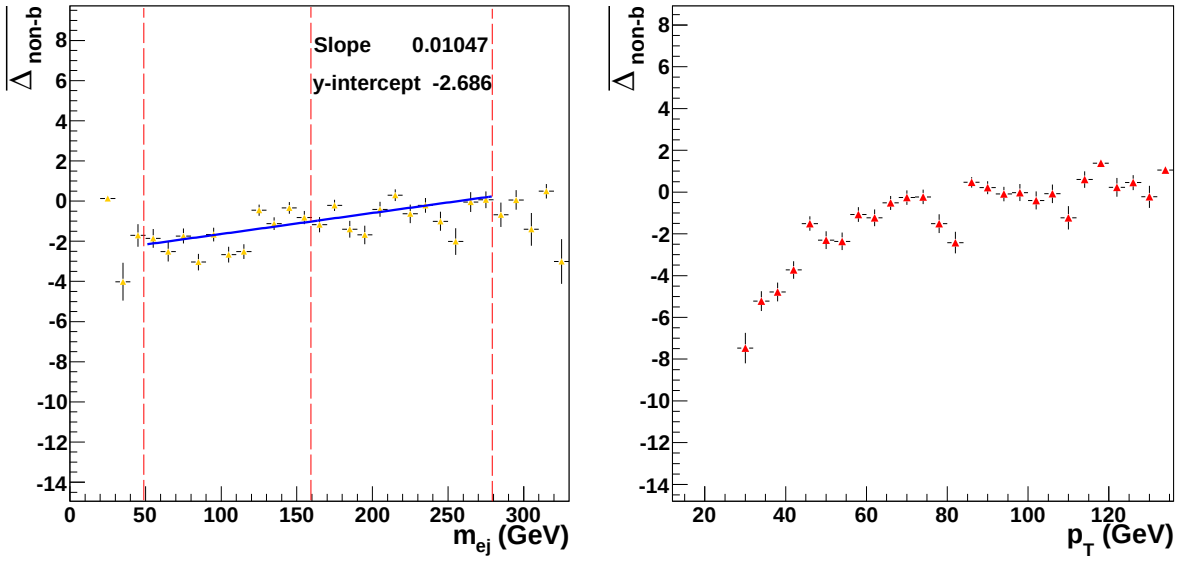
To investigate the observed dip in Figure 5.26, the assumptions behind the method have to be revisited. The dip happens at low d_{TCHE} values which are more accumulated by the non-b-jets. Therefore, it makes sense to reinvestigate the estimation of the non-b-jet content in the b-dominated sample.

In Figure 5.23, the shapes of Δ_{non-b} in the Left and the Right regions were compared where it has been concluded that the m_{ej} and the d_{TCHE} are not correlated too much. Hence one could have estimated the Δ^{L}_{non-b} from the same distribution in the Right region.

Since the direct correlation between d_{TCHE} and m_{ej} is not strong enough to be visible in the shape comparison, the mean value of the b-tag discriminant in the of m_{ej} is plotted in Figure 5.27 (a) where a straight line is fitted to demonstrate the increasing trend.

Although at a first glance the existing relation between the m_{ej} and the b-discriminator

[9] The TCHE algorithm needs at least two selected tracks in a jet, i.e. compatible with the jet axis. The jet is considered as non-tagable otherwise.



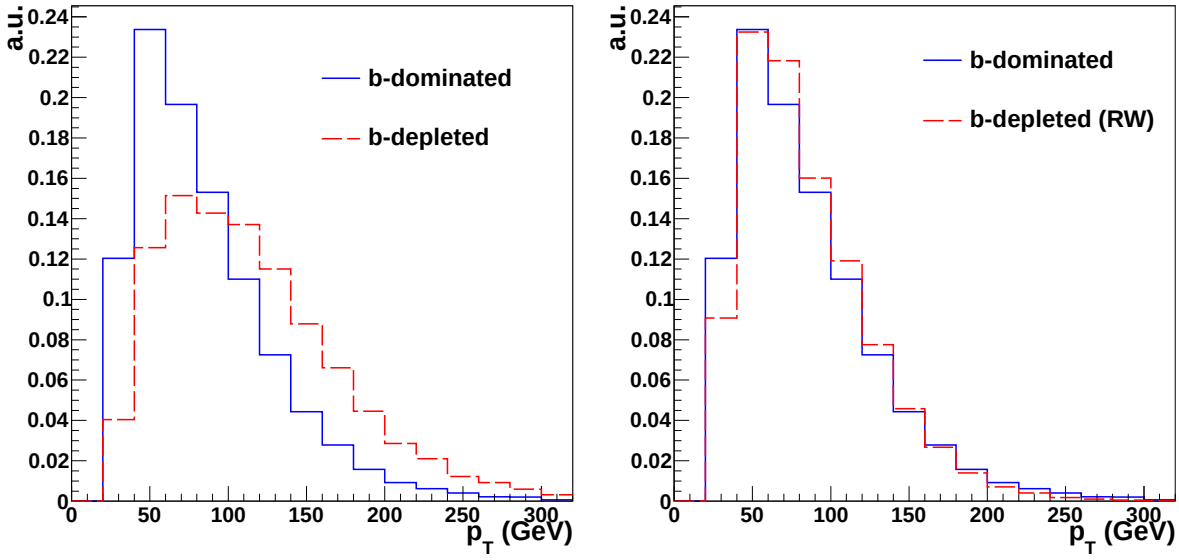
(a) (b)

Figure 5.27: The mean value of the TCHE b-discriminator in the bins of m_{ej} (a) and p_T (b) for non-b-quark jets in the b-candidate jet sample for the semi-electron $t\bar{t}$ events. The vertical dashed lines in (a) are the boundaries by which the b-depleted and the b-dominated subsamples are constructed.

may seem not so obvious, the correlation between the jet p_T and the d_{TCHE} which is shown in Figure 5.27 (b) clarifies it. This figure illustrates the mean value of the d_{TCHE} distribution for the non-b-quark jets in the b-candidate jet sample which is increasing as a function the jet p_T. The jets with higher p_T on the other hand, tend to give higher m_{ej} values hence the slightly increasing behavior is also observed in Figure 5.27 (a).
To resolve such a correlation, the jets in the Right region are given a weight according to their p_T. To extract the weight, the p_T distribution of the jets in the b-dominated subsample is divided by the p_T distribution of the jets in the b-depleted jet sample. The divided histogram is fitted to an arbitrary function and this function gives the p_T-dependent weight given to the jets in the Right region. Thereafter the p_T distribution of the jets in the b-depleted subsample resembles the one for the jets in Left region. Figure 5.28 compares the p_T distribution of the jets in the b-depleted and b-dominated subsamples before and after reweighting. The distributions after reweighting seem to agree much better.

The influence of the p_T reweighting on the $m_{ej} - d_{TCHE}$ correlation is shown in Figure 5.29. It can be observed that the slope of the increasing trend has been reduced compared to the case before p_T reweighting shown in Figure 5.27 (a).

The Δ^R_{all} in Equation 5.7 is replaced by the b-tag distribution of the reweighted jets, $\Delta^R_{all,rw}$, and the b-tagging efficiency is calculated in the same way as before. Fiqure 5.30 shows the estimated efficiency $\widehat{\epsilon_b}$ after the reweighting procedure together with the relative difference between the $\widehat{\epsilon_b}$ and the true b-tagging efficiency of the b-quark jets. The dip has certainly been reduced by the p_T reweighting method.



(a) (b)

Figure 5.28: The comparison between the p_T distributions of the jets in the b-dominated and the b-depleted jet sample before (a) and after (b) p_T reweighting. The jet samples are made withing the semi-electron $t\bar{t}$ events.

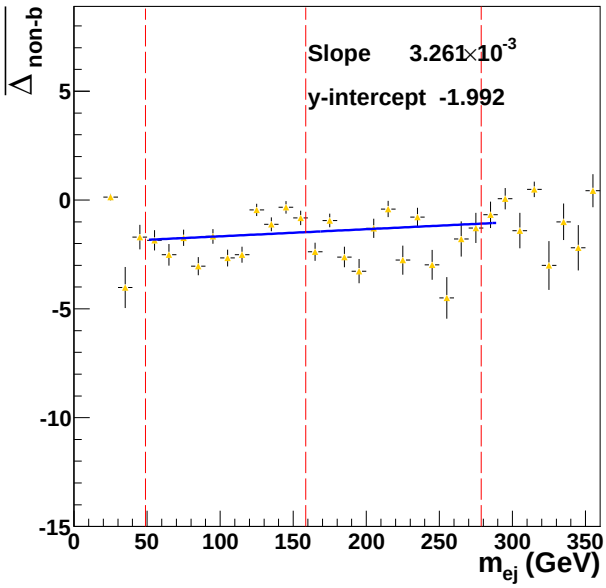
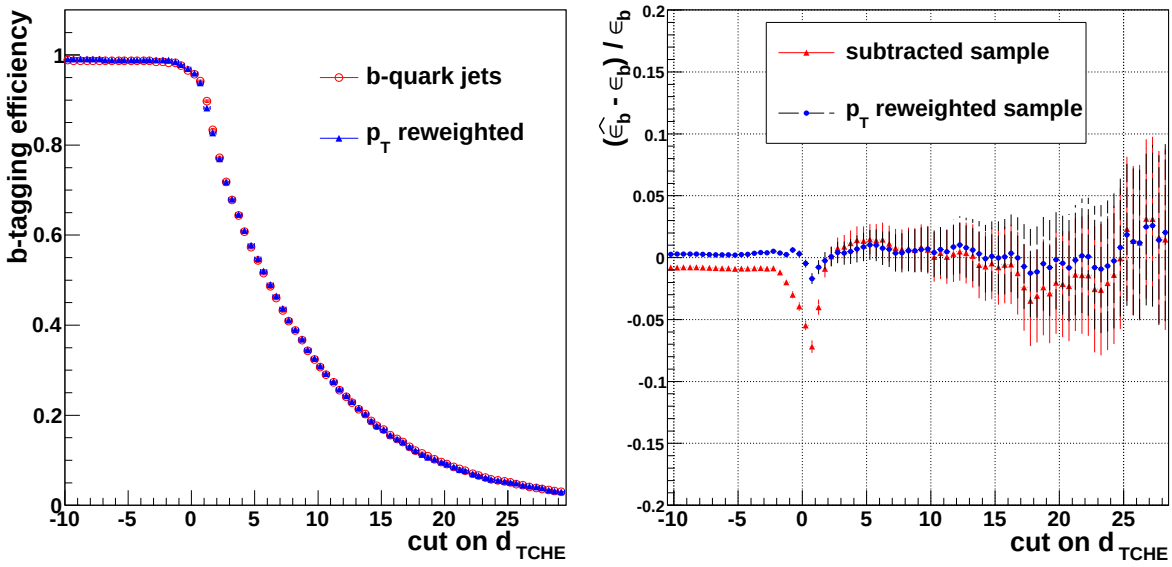


Figure 5.29: The mean value of the d_{TCHE} in the bins of m_{ej} for non-b-quark jets in the b-candidate jet sample for the semi-electron $t\bar{t}$ events. The vertical dashed lines are the boundaries by which the b-depleted and the b-dominated subsamples are constructed.



(a) (b)

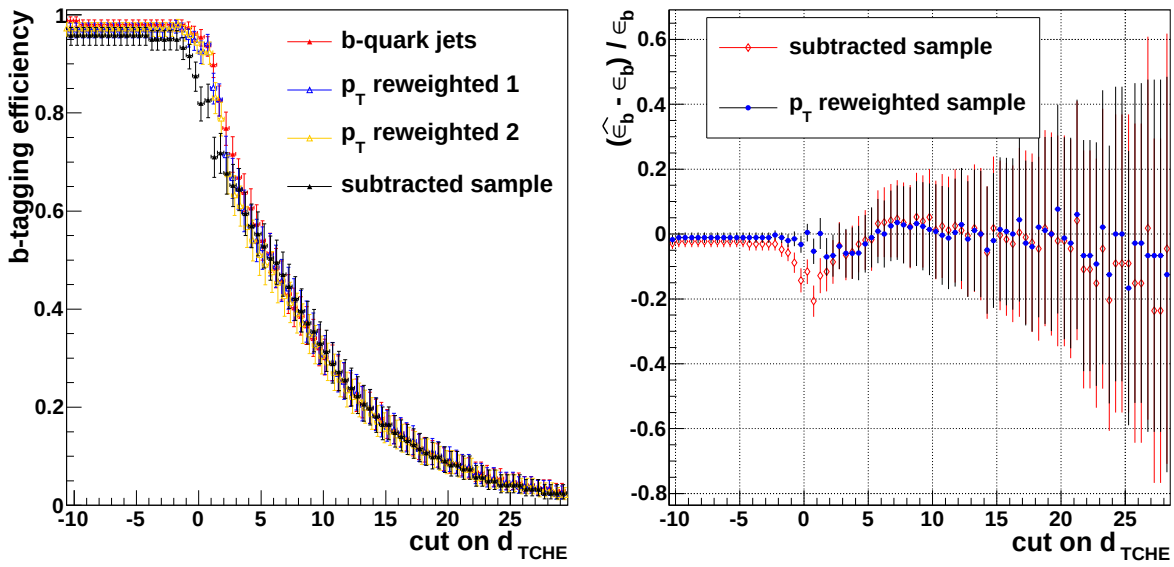
Figure 5.30: The $\widehat{\epsilon_b}$ after the p_T reweighting, compared to ϵ_b versus the cuts on the b-discriminant, (a). The relative error on $\widehat{\epsilon_b}$ before and after the p_T reweighting, (b).

5.3.3 Evaluation of $\widehat{\epsilon_b}$ including backgrounds

Including backgrounds at $100\,pb^{-1}$ of integrated luminosity, the scale factor F changes from $F = 1.607$ to $F = 1.695$. The efficiency derived by the method, $\widehat{\epsilon_b}$, shows a discrepancy with ϵ_b in the lower b-tag values, similar to the case of the "signal only" analysis. This discrepancy is also resolved by the p_T reweighting. To reweigh the jets, different functions have been fitted to the p_T ratio distribution and the final efficiency result has been found to be almost independent from the fit functions.
Figure 5.31 (a) is the b-tagging efficiency as a function of a cut on d_{TCHE} for the b-quark jets (ϵ_b) and for the jets in the subtracted sample before and after the p_T reweighting. Two fit functions among all are picked to illustrate the method improvement after p_T reweighting. Regarding the agreement between the different weight functions, only one of them is presented in Figure 5.31 (b). In this figure the relative difference between ϵ_b and $\widehat{\epsilon_b}$ is shown for the default and the reweighted b-depleted jet sample.

The effect of the W+jets background

The discrepancy that resulted in the reconsideration of the m_{ej}-d_{TCHE} correlation happens at low b-tag values and is sensitive to the non-b-quark jet contribution. The W+jets background events on the other hand, introduce a considerable amount of additional non-b-quark content to b-candidate jet sample. Therefore at first look, it makes sense to find a way to reduce this background contamination. This is the reason behind the effort made in Section 5.1.4, i.e. trying different variables to discriminate



(a) (b)

Figure 5.31: The $\widehat{\epsilon_b}$ as a function of a cut on the b-discriminator value compared to ϵ_b within the b-dominated jet sample for the signal and background processes together at $100\,pb^{-1}$ integrated luminosity, (a). The relative error on $\widehat{\epsilon_b}$ in (a) for different b-discriminator cut values, (b).

between the W+jets and the $t\bar{t}$ events to reject this background effectively.

As shown in Figure 5.32, the cross section of the W+jets background is changed to see the effect of these events on the dip. Two extreme scenarios, "no W+jets" and "doubled W+jets" hypotheses, together with the expected W+jet in real data are demonstrated. The statistical uncertainties at $100\,pb^{-1}$ are not so small and have not been drown to make the changes visible. The dip evolves according to the non-b-quark jets contamination from the W+jets events and it is indeed smaller at lower W+jets cross section. Here, the p_T reweighting is not yet applied.

The same study is done with the p_T reweighting. For each scenario, the p_T weight functions are extracted and applied on the jets in the b-depleted sample. The b-tagging efficiencies are calculated and the difference between the estimated, $\widehat{\epsilon_b}$, and true, ϵ_b, efficiencies for each case is computed as a function of a cut on the b-discriminator.

The evolution of the dip together with the $\widehat{\epsilon_b}$ are illustrated in Figure 5.33. The changes in the dip are not as dramatic as the previous case where the b-tagging efficiencies are also in a fair agreement within the uncertainties. The plot for the b-tagging efficiency is zoomed to the region of interest to show the changes better. It seems that the p_T reweighting method can handle, to a good extent, the non-b-quark jet contamination regardless of the physics processes.

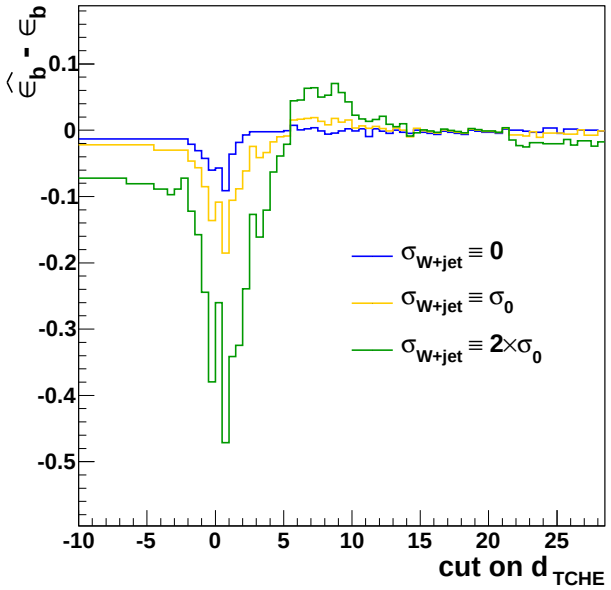
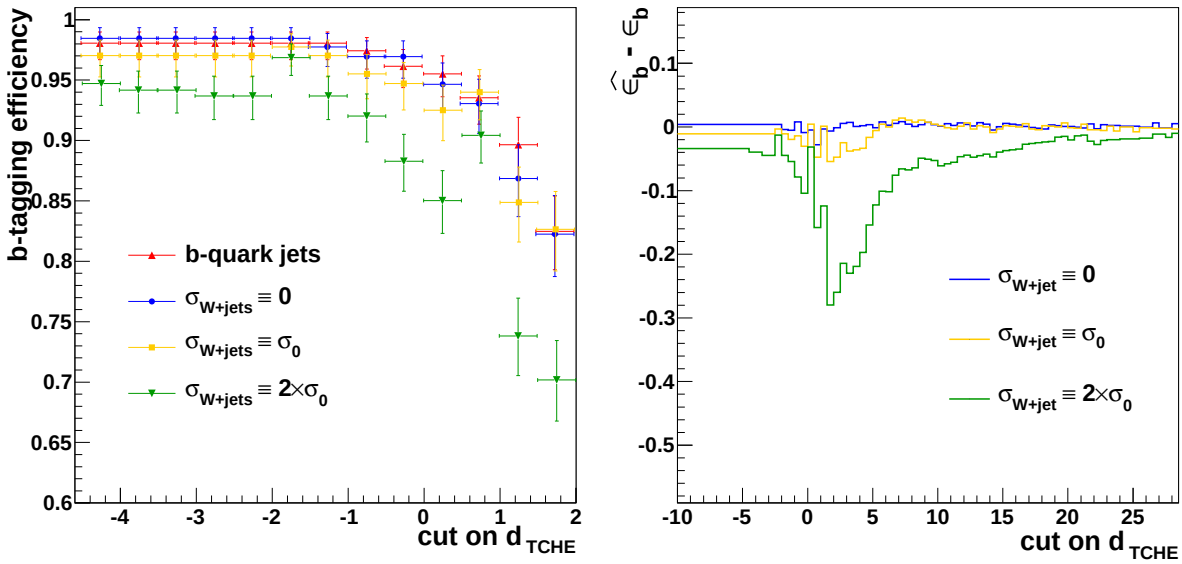


Figure 5.32: The difference between the estimated efficiency, $\widehat{\epsilon_b}$, and true efficiency, ϵ_b, for different W+jets background scenarios at $100\,pb^{-1}$ integrated luminosity. The quantity σ_0 is the nominal cross section of the W+jets process given in Table 3.2.



(a) (b)

Figure 5.33: The estimated efficiency $\widehat{\epsilon_b}$, (a), and its differenece with the true effiency, (b), for different W+jets background scenarios after reweighting the jets in the Right region according to their p_T. The quantity σ_0 is the nominal cross section of the W+jets process given in Table 3.2.

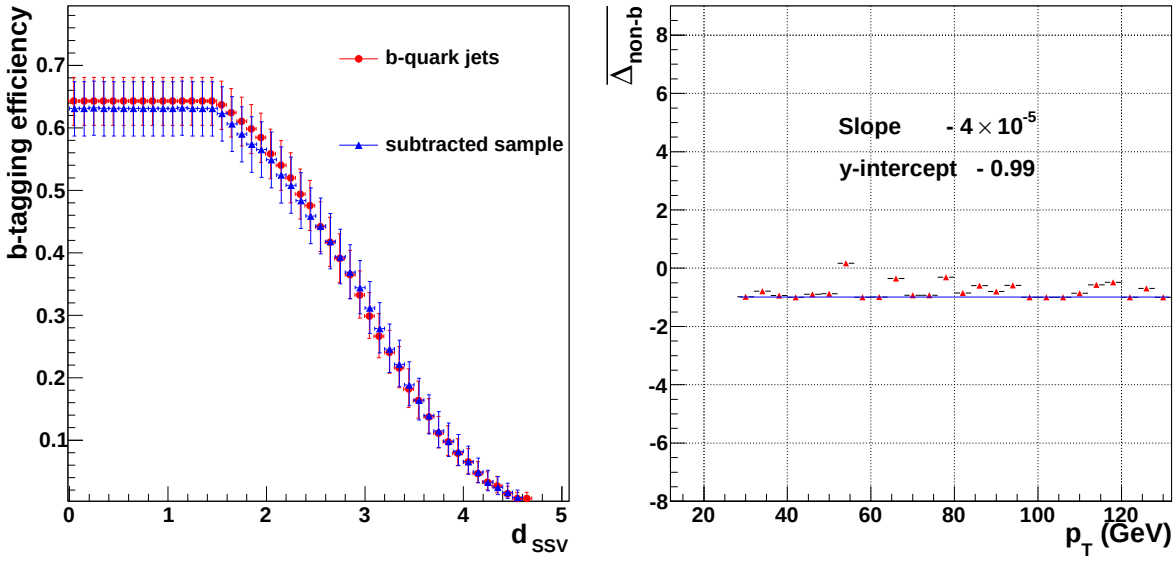
5.3.4 Other b-tagging algorithms

The method that is just developed to measure the b-tagging efficiency, is based on the kinematic properties of the semi-electron final state of $t\bar{t}$ events. Hence it can be generalized to other b-tagging algorithms. The only part which may need to be adapted accordingly, is the p_T reweighting. Depending whether the b-discriminator is strongly correlated to the jet p_T, the p_T reweighting algorithm may or may not be needed.
In this subsection, the generalization of the method to other b-tagging algorithms is presented. It is aimed to study at least one b-tagging algorithm from each of the b-jet identification categories introduced in Section 4.4. The results are obtained for the mixture of the signal and background processes at $100\,pb^{-1}$.

- **Simple secondary vertex**
  The efficiency for the b-quark jets and the jets in the subtracted sample are compared in Figure 5.34 (a) as a function of a cut on the d_{SSV} variable [10]. The good agreement between the $\widehat{\epsilon_b}$ and the true b-tag efficiency implies the success in the estimation of the Δ^L_{non-b} with the Δ^R_{all} in Equation 5.3. It means that the b-discriminant of the simple secondary vertex algorithm does not have a strong correlation with the m_{ej} variable.
  The correlation between the m_{ej} quantity and the b-discriminator is a consequence of the correlation between the b-discriminant and the jet p_T. Figure 5.34 (b) illustrates the mean value of this b-discriminator in the bins of the jet p_T for the non-b-quark jets in the b-candidate jet sample. As it can be seen also in the slope of the fitted line, there is almost no correlation for the d_{SSV} to the p_T of the jet.

- **Combined secondary vertex**
  As it has been explained in Section 4.4.2, more extensive than the simple secondary vertex method, the combined secondary vertex algorithm uses additional jet properties to identify the b-quark jets. The b-discriminant of this algorithm has some correlation with the jet p_T and consequently with the m_{ej} variable. Hence, the efficiency estimation method needs the p_T reweighting to resolve such correlation.
  Figure 5.35 (a) is the result of the method as a function of a cut on d_{CSV} before and after the p_T reweighting together with the efficiency for b-quark jets. While before the p_T reweighting, the method returns even non-physical values at lower d_{CSV}'s, the estimated efficiency $\widehat{\epsilon_b}$ shows a good agreement with the true efficiency ϵ_b after the reweighting.
  The relative difference between the true efficiency and the outcome of the method after the p_T reweighting is illustrated in Figure 5.35 (b). The values are compatible with zero within the uncertainties.

[10] The high efficiency version of the simple secondary vertex algorithm is used. See Section 4.4.2 for more details.

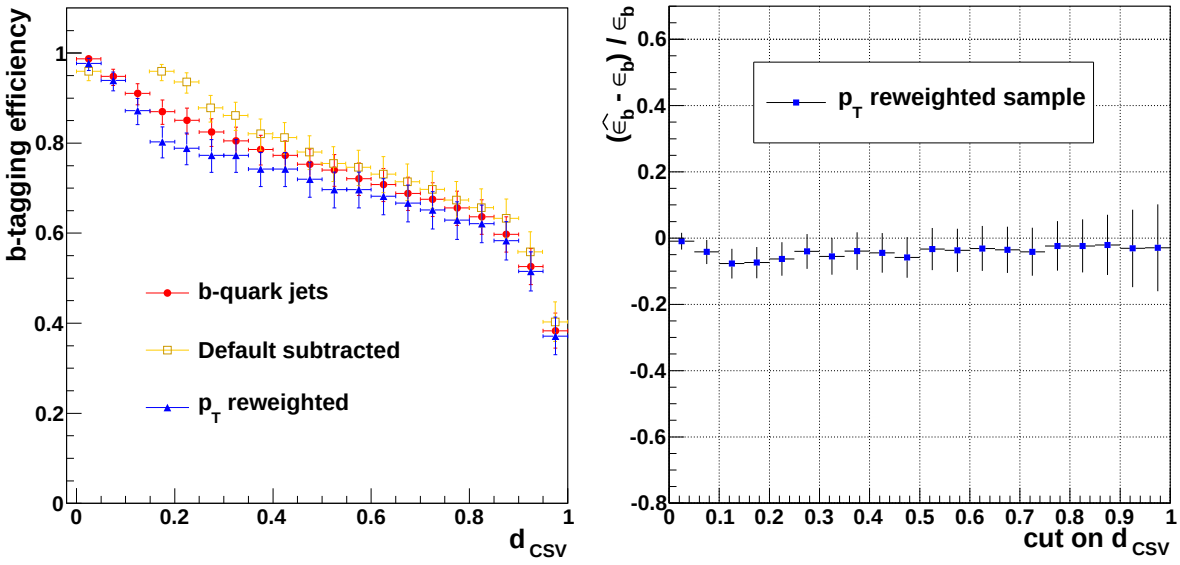


(a) (b)

Figure 5.34: The b-tagging efficiency of the simple secondary vertex algorithm, calculated by the method and compared to the efficiency of the b-quark jets in the same jet sample, (a). The mean value of the b-discriminant as a function of the jet p_T for the non-b-quark jets in the b-candidate jet sample, (b).

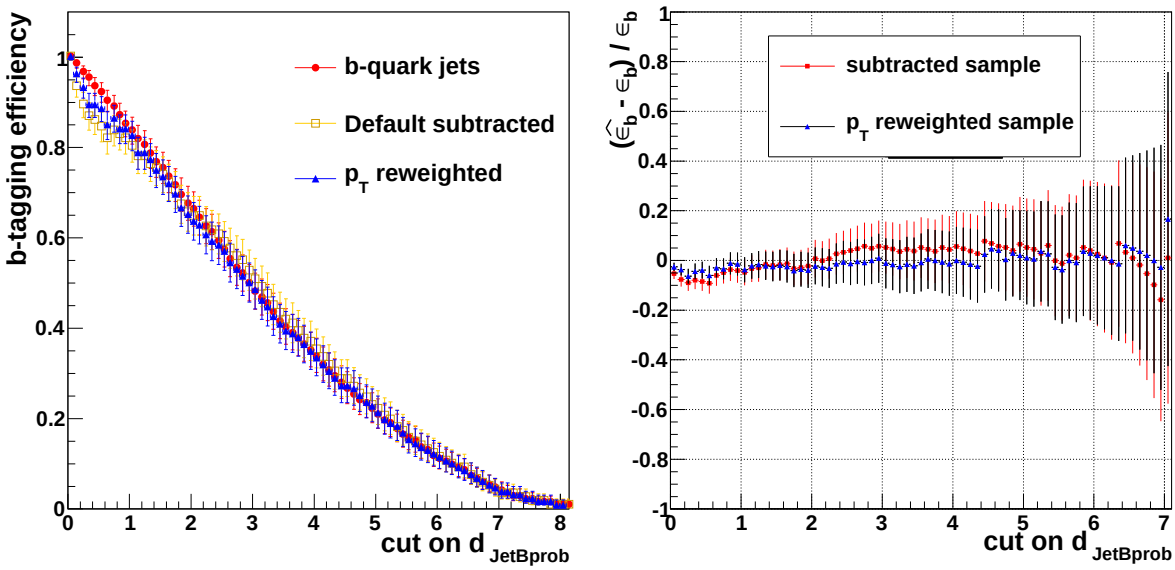
- **Jet B Probability**
  Like the Track Counting High Efficiency, the Jet B Probability algorithm belongs to the category of the "impact parameter" based b-tagging algorithms (see Section 4.4.1). Hence after the p_T reweighting, the method is improved at lower $d_{JetBProb}$ values as it can be seen in Figure 5.36 (a).
  In Figure 5.36 (b), the relative difference between the $\widehat{\epsilon_b}$ and the true b-tagging efficiency is closer to zero after the p_T reweighting.

- **Soft muon**
  From the soft lepton category, the performance of the method is presented for the soft muon b-tagging algorithm (see Section 4.4.3). The subtraction of the scaled b-discriminator distribution of the unweighted jets in the Right region from Δ^L_{all}, leads to non-physical values at lower $d_{soft\,\mu}$'s. As demonstrated in Figure 5.37 (a), the p_T reweighting significantly improves the results. The relative difference between the result of the method after the p_T reweighting and the true efficiency, ϵ_b, is shown in Figure 5.37 (b). Considering the uncertainties, no bias is observed.



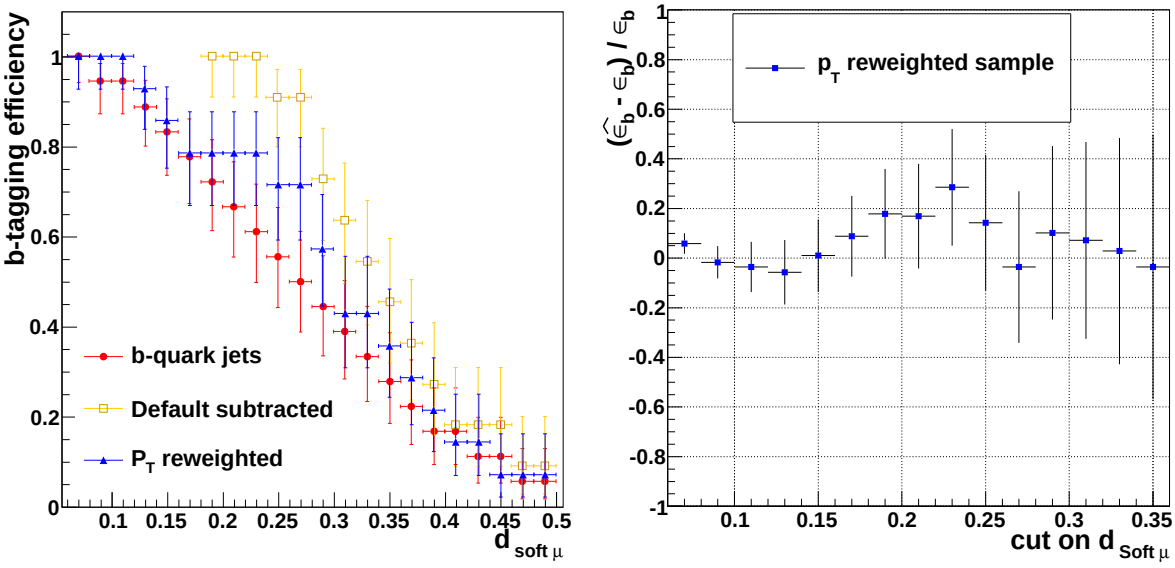
(a) (b)

Figure 5.35: The b-tagging efficiency of the combined secondary vertex algorithm, calculated by the method and compared to the true efficiency before and after the p_T reweighting are illuatrated in (a) where the non-physical efficiencies before the p_T reweighting are not shown. The relative error on the method with respect to the true efficiency after the p_T reweighting is shown in (b).



(a) (b)

Figure 5.36: The efficiency of the jet B probability b-tagging algorithm, calculated by the method and compared to the true efficiency, (a), together with the relative error on the method with respect to the true efficiency, (b), both before and after the p_T reweighting.



(a) (b)

Figure 5.37: The $\widehat{\epsilon_b}$ for the soft muon b-tagging algorithm compared to the ϵ_b before and after the p_T reweighting are illustrated in (a) where the non-physical efficiencies before the p_T reweighting are not shown. The relative error on the method with respect to the true efficiency after the p_T reweighting is shown in (b).

5.3.5 The fully data-driven approach

The presented method for the b-tagging efficiency estimation has relied partialy on the generator level information from the simulated samples so far, i.e. in the evaluation of the scale factor, F. The scale factor is to represent the ratio of the number of non-b-quark jets in the b-dominated and b-depleted samples or, in another words, in the two different m_{ej} regions.
These jets in the semi-electron $t\bar{t}$ events are expected to come from the W boson or the radiation and are wrongly associated to the leptonic b-candidate. They do not necessarily have a strong kinematic correlation with the electron of the leptonic side of the event.
Therefore if one can prepare a jet sample, mostly dominated by non-b-quark jets, and make the invariant mass of the jets in the sample with the electron, one may expect to see a similar m_{el} distribution [11] as it is seen from non-b-quark jets in the b-candidate jet sample.
This new jet sample dominated by non-b-quark jets, the *control jet sample*, is made up of the jets associated to the hadronically decayed W boson by the χ^2_{min} jet association. The b-purity of the control sample is about 18% confirming the dominant contribution by the non-b-quark jets. With the same m_{el} boundary values as in the b-candidate jet sample, the control sample can be divided in two pieces, Left and Right and the scale factor can be defined as

$$F = \frac{N_C^L}{N_C^R}, \tag{5.8}$$

where N_C^X is the number of jets in the X region of the control jet sample.
To work in a situation similar to the real data, from now on all the numbers and plots are for the signal together with the background processes at $100\,pb^{-1}$ except when it is stated differently.
The data driven scale factor for this condition is $F = 2.913$. This is about two times larger than the $F = 1.695$ derived from simulated samples. Such increment results in an overestimation of $F \cdot \Delta^R_{all}$ and makes the right hand side of Equation 5.7 negative for many d_{TCHE} bins. Hence non physical values for the efficiency will be obtained.

The enlargement of the scale factor indicates the presence of more jets in the Left region then the right interval. This is illustrated in Figure 5.38 where the shapes of the lepton-jet invariant mass for the non-b-quark jets in the b-candidate jet sample and for all jets in the control sample are compared.
The shape difference can be due to the presence of b-quark jets in the control jet sample since in making the hadronic top combination, there is no way in data to check if the W boson constituents are really non-b-quark jets or the jets from the B-mesons. As a consequence, for the physics processes including a leptonically decayed top quark, if the leptonic b-quark jet is associated to the W boson, the lepton-jet invariant mass will move towards the Left region resulting in larger values for the F.
Another reason behind such a discrepancy can be the different jet kinematics within two jet sample. Because of the low b-purity in the control sample, the kinematic source

[11] "l" denotes "light"-candidate jets to be distinguished from b-candidate jets. The c-candidate jets here are labeled as "light" as well.

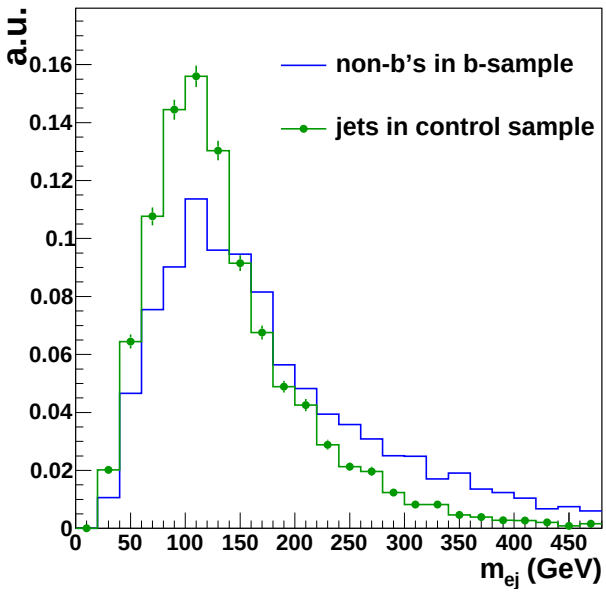


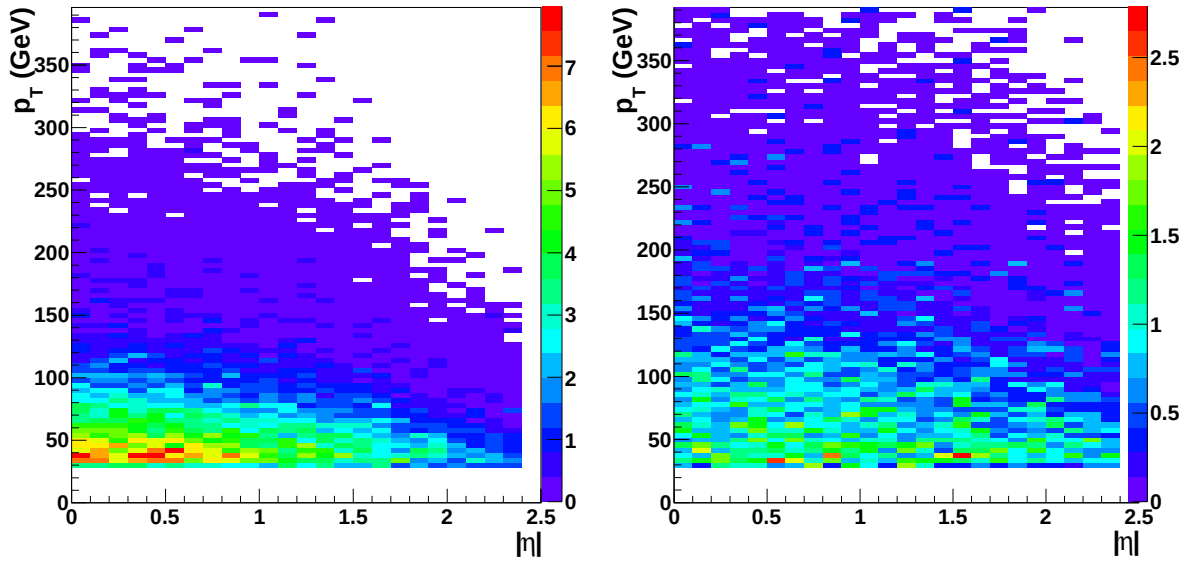
Figure 5.38: The distribution of the electron-jet invariant mass for the jets in the control sample compared to the non-b-quark jets in the signal sample.

of inconsistency seems to have more influence. Both hypotheses are investigated in the following:

- **Kinematic solution: Reweighting the m_{el} distribution**
  If the jets in the b-candidate jet sample and in the control sample were randomly picked, one would expected to see a similar lepton-jet invariant mass shapes or equivalently similar kinematic distributions for the two jet samples. However, these jets have had such a distinguishable kinematic properties that they have fulfilled the χ^2_{min} criterion with the W boson and the top quark mass constraints. They have in fact been pushed to different jet samples because of their kinematic differences. Figure 5.39 shows the two dimensional η-p_T distributions for the jets in the b-candidate and control jet samples. The jets in the control sample are softer and more central.

  The idea of kinematic reweighting is exploited once more here to reweight the jets in the control sample regarding to their η and p_T. The two dimensional η-p_T distributions of all jets in the b-candidate jet sample is divided to the same distribution of the jets in the control samples to produce the weights in η-p_T bins. As demonstrated in Figure 5.40, the lepton-jet invariant mass shapes in the b-candidate and the control jet sample look more similar after the $(\eta; p_T)$ reweighting. The scale factor calculated from the reweighted distribution is $F = 1.669$ having a relative difference with the simulation driven F of 1.5%.
  It is notable that the η-p_T weights can be obtained in data, making the method fully independent from simulation.

- **Anti-tagging the m_{el} distribution**
  Although the b-quark jet contamination in the control jet sample is less than 20%, one may still think about the anti-tagging of the jets in the control sample



(a) (b)

Figure 5.39: The two dimensional $p_T - \eta$ distribution for the jets in the control (a) and the leptonic b-candidate (b) jet sample.

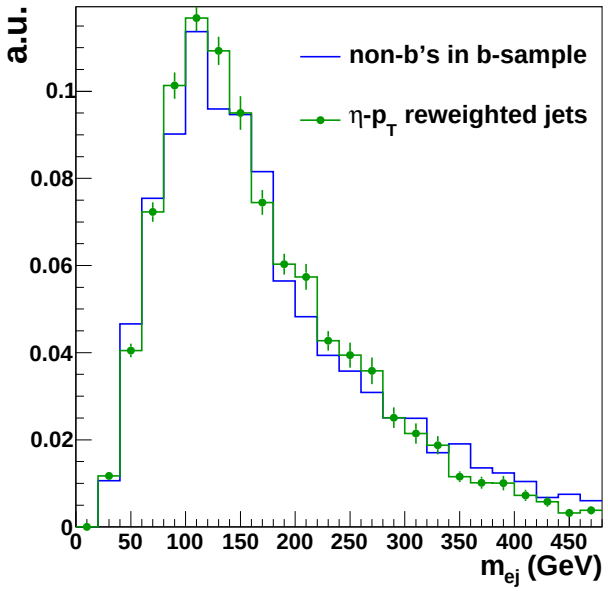
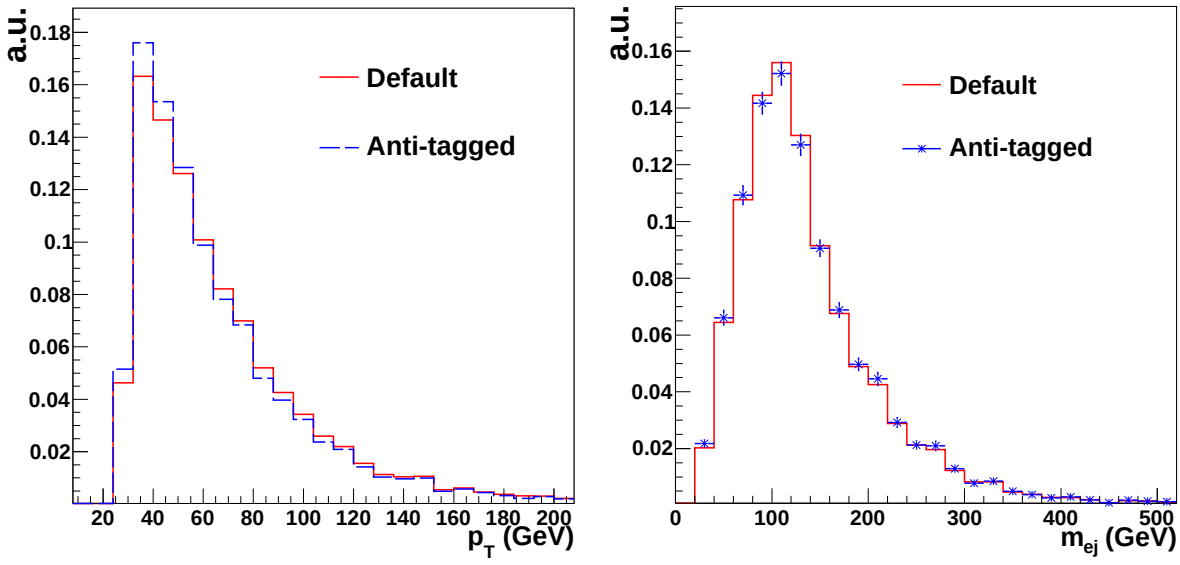


Figure 5.40: The distribution of the electron-jet invariant mass for the jets in the control sample compared to the non-b-quark jets in the signal sample. The jets in the control sample are reweighted by the two dimensional $p_T - \eta$ weight function.

in addition to the kinematic reweighting to reject the b-quark jets even more. Asking for a d_{TCHE} less than some value gives the b-quark jets less chance to enter the control sample and to populate the peak area in the m_{el} distribution which leads to an increment of the scale factor.
The requirement of $d_{TCHE} < 3$ results in a smaller scale factor as expected, $F = 2.805$, however it is still large enough to give non-physical values for the final efficiency results. The anti-tagged jets in the control sample are further reweighted according to their η and p_T as already explained.
The weight factors are obtained dividing the η-p_T distributions of all jets in the b-candidate jet sample to the same distribution of the jets in the "anti-tagged" control samples. The scale factor changes to $F = 1.562$ with 8% of relative difference to the simulation driven F.

Due to the fact that the b-quark jets carrying more energy, the anti-tagging



(a) (b)

Figure 5.41: The p_T (a) and m_{el} distribution for the jets in the anti-tagged control jet sample.

procedure rejects mainly the high p_T jets in the control sample (Figure 5.41 (a)). This however does not influence the m_{el} distribution that much (Figure 5.41 (b)). Therefore the scale factor remains almost unchanged, $\frac{\Delta F}{F} \sim 3\%$, when applying the anti-tag cut.

Loosing the jets mainly from the tail of the p_T distribution, the η-p_T reweighting gives larger weights to the jets with higher p_T in the control sample to match to the jet p_T distribution of the b-candidate sample. Since jets with higher p_T give in general larger m_{el} values, the weighted entries in the high m_{el} range are enhanced. This results in the relatively smaller value for the scale factor. Figure 5.42 demonstrates the m_{el} shape for the reweighted jets in the default and the anti-tagged control sample. It can be seen that for the anti-tagged jets,

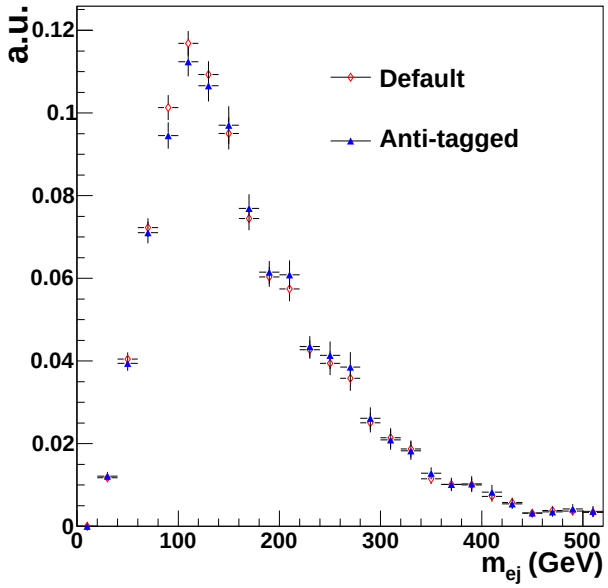


Figure 5.42: The lepton-jet invariant mass distribution for the jets in the $(\eta; p_T)$ reweighted control jet sample. The reweighted default jet sample is compared to the reweighted anti-tagged one.

the shape has shifted toward the higher m_{el} values and this leads to the smaller value for the scale factor.

The results of the data driven F study is summarized in Table 5.7 where the relative error with respect to the F derived from simulation is also presented.

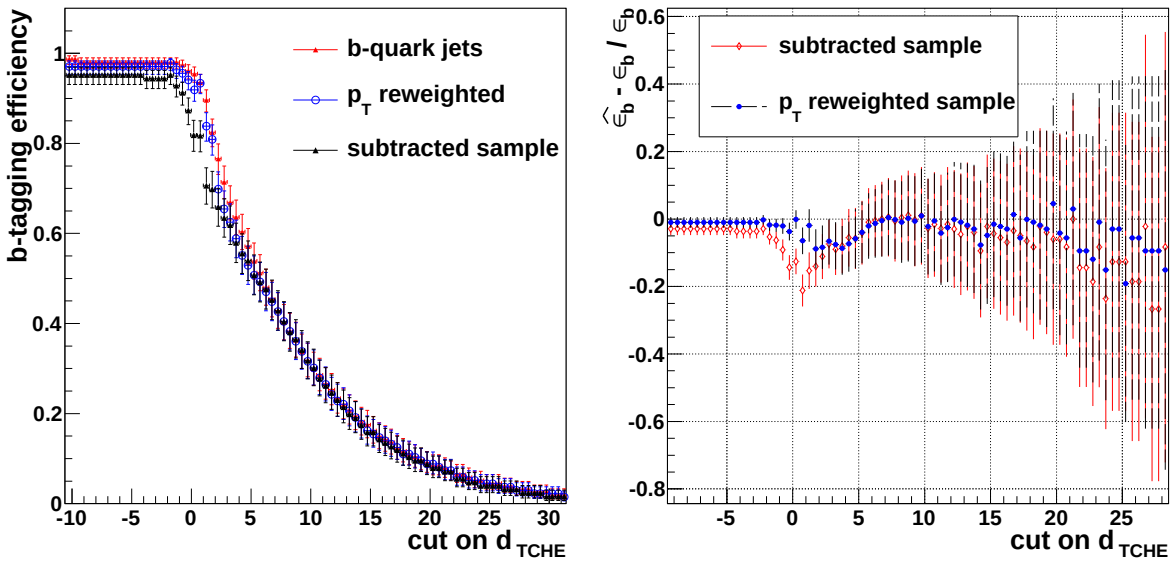
	default	anti-tagged (I)	$p_T - \eta$ reweighted (II)	(I) & (II)
$F_{data\ driven}$	2.913	2.805	1.669	1.562
$\lvert\Delta F\rvert / F_{simulation}$	72%	65%	1.5%	7.8%

Table 5.7: The data driven scale factor together with the relative bias with respect to the F from simulation. The effect of anti-tagging and the $(\eta; p_T)$ reweighting of the control sample on the $\widehat{F}$ are presented both individually and on top of each other.

Because of the good performance of the $p_T - \eta$ reweighting, only this method is used in the estimation of the data driven scale factor. The b-tagging efficiency is then calculated in a fully data driven approach. The true b-tagging efficiency, ϵ_b, together

with the outcome of the method, $\widehat{\epsilon_b}$, are illustrated in Figure 5.43 (a) where the $\widehat{\epsilon_b}$ is shown before and after the p_T reweighting.

As it can be seen in Figure 5.43 (b) the method has almost no bias within the given



(a) (b)

Figure 5.43: The $\widehat{\epsilon_b}$ calculated using the data driven scale factor, as a function of a cut on the b-discriminator value. The scale factor is extracted from the $p_T - \eta$ reweighted control sample. The efficiency is illustrated before and after p_T reweighting of the jets in the b-depleted sample, (a). The relative error on $\widehat{\epsilon_b}$ in (a) for different b-discriminator cut values both before and after the p_T reweighting, (b).

statistics, after the p_T reweighting. The residual biases for three different working points are listed in Table 5.8 where the statistical errors are also quoted. The loose, medium and tight working points are corresponding to the b-discriminator cuts which result in the $\epsilon_b \approx 75\%$, 50% and 25% respectively.

5.4 Statistical properties of the estimators

Using the simulated subsamples with the same statistics as expected in the data for a given integrated luminosity, is one of the ways to test the statistical reliability of the estimators and their uncertainties calculated in the analysis. If the estimation is correctly performed, repeating the measurement on the independent experiments would result in a similar output. Having N experiments in which the measured quantity X_i has a mean value of μ and an expected variance of σ^2, the sample average is defined as

$$X_N = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{5.9}$$

	ϵ_b^{exp}	$\widehat{\epsilon_b}$ (no p_T-rew)	$\widehat{\epsilon_b}$ (p_T-rew)
loose	0.766 ± 0.044	0.659 ± 0.045	0.698 ± 0.048
medium	0.480 ± 0.040	0.476 ± 0.042	0.471 ± 0.046
tight	0.253 ± 0.036	0.246 ± 0.039	0.243 ± 0.039

Table 5.8: The expected, ϵ_b^{exp}, and the estimated b-tagging efficiency, $\widehat{\epsilon_b}$, using the fully data driven approach before and after the p_T reweighting, for the loose, medium and tight working points. Comparing the first and the last columns, there is almost no bias on the estimated efficiencies within the given statistics.

According to the Central Limit Theorem, X_i is expected to have an approximately normal distribution with a mean around the expectation value μ (estimated by $\hat{\mu} = X_N$) and a variance of σ^2. The distribution of $\frac{X_i - X_N}{\sigma}$, known as *pull distribution* is therefore expected to be a normal distribution centered at zero with the width of unity.
While the mis-estimation of X_i's changes the mean value of the pull, the overestimation (underestimation) of the variance in each experiment leads to a width less (greater) than unity. Hence the statistical properties of the estimators can be studied by looking at the pull distributions.
The simulated subsamples, pseudo-experiments, are reflecting the experiments in which for each of them the data driven scale factor, $\widehat{F}$, and the b-tagging efficiency $\widehat{\epsilon_b}$ is estimated. The pull for the data driven scale factor is the distribution of

$$\frac{F_N - \widehat{F}_i}{\widehat{\delta F}_i}, \tag{5.10}$$

where F_N is the averaged scale factor over all pseudo-experiments. The estimated scale factor in pseudo-experiment i together with its uncertainty are denoted as $\widehat{F}_i$ and $\widehat{\delta F}_i$, respectively. The pseudo-experiments contain the expected population for $100\,pb^{-1}$.

Figure 5.44 illustrates the pull distribution for the data driven scale factor for 500 pseudo-experiments. A Gaussian function is fitted to the distribution for which the width is equal to ~ 1.21, meaning that the statistical error on the $\widehat{F}$ is underestimated by $\sim 21\%$.
Due to the sensitivity of the b-tagging efficiency to the scale factor (Equation 5.7), it makes sense to check the mean value of the $\widehat{F}$ before looking at the pull of the $\widehat{\epsilon_b}$. Figure 5.45 (a) is the distribution of $\widehat{F}$. It can be seen that the mean value of the distribution, $F_{mean} \approx 2$, is much larger than the $\widehat{F} \approx 1.7$ derived from the full set of the simulated samples. The consequence of such an overestimation is the appearance

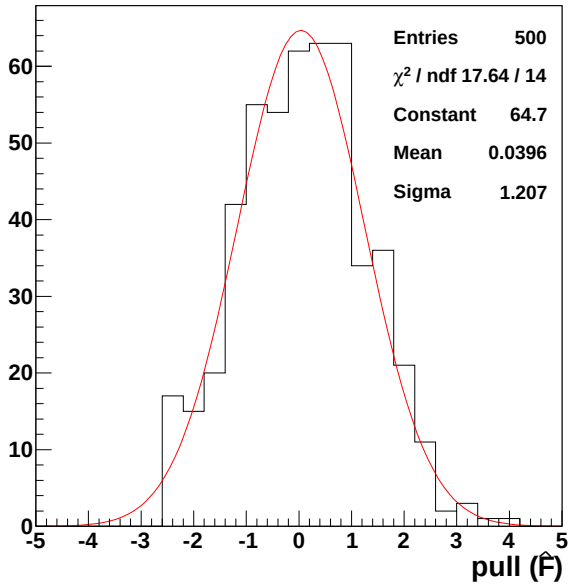


Figure 5.44: The pull distribution of the data driven scale factor at $100\,pb^{-1}$ integrated luminosity.

of non-physical values for the b-tagging efficiency, illustrated in Figure 5.45 (b).

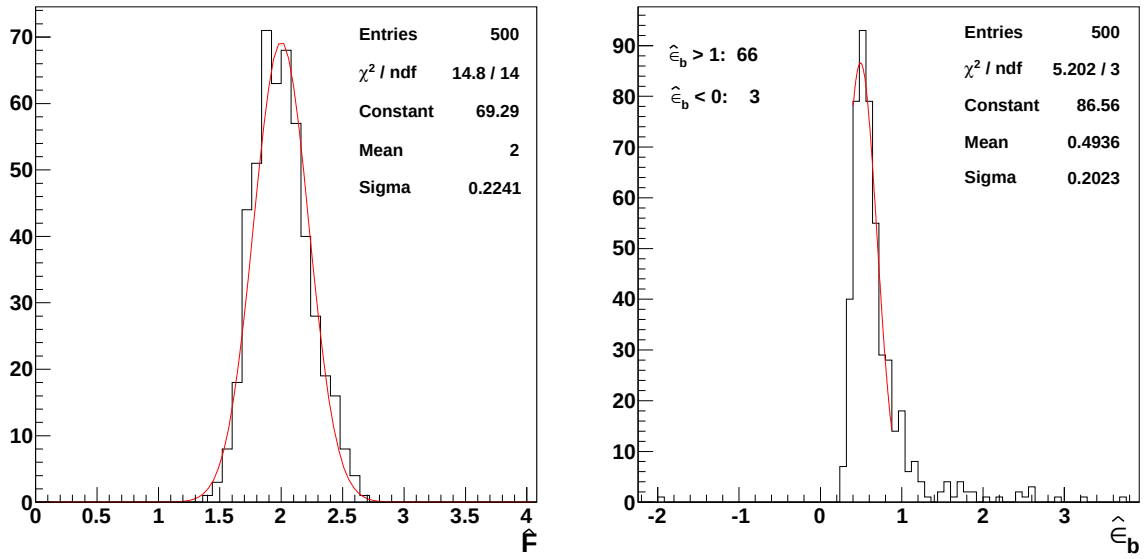
The fully data driven approach is followed in each pseudo-experiment, means the $(\eta; p_T)$ weights for the $\widehat{F}$ calculation are computed within the pseudo-experiment. The $(\eta; p_T)$ weights may however be nonsense due to the insufficient amount of statistics. Instead, these weights can be substituted by the values extracted from the full statistic of the simulated samples which are statistically more reliable. The distributions of the $\widehat{F}$ and the $\widehat{\epsilon_b}$ after the $(\eta; p_T)$ weight substitution are shown in Figure 5.46. A better mean value, $F_{mean} \approx 1.65$, is obtained for the scale factor and the fraction of non-physical efficiencies is less than 1%.

The pull for the b-tagging efficiency is similarly defined as

$$\frac{\epsilon_{b,N} - \widehat{\epsilon_b^i}}{\widehat{\delta\epsilon_b^i}}, \tag{5.11}$$

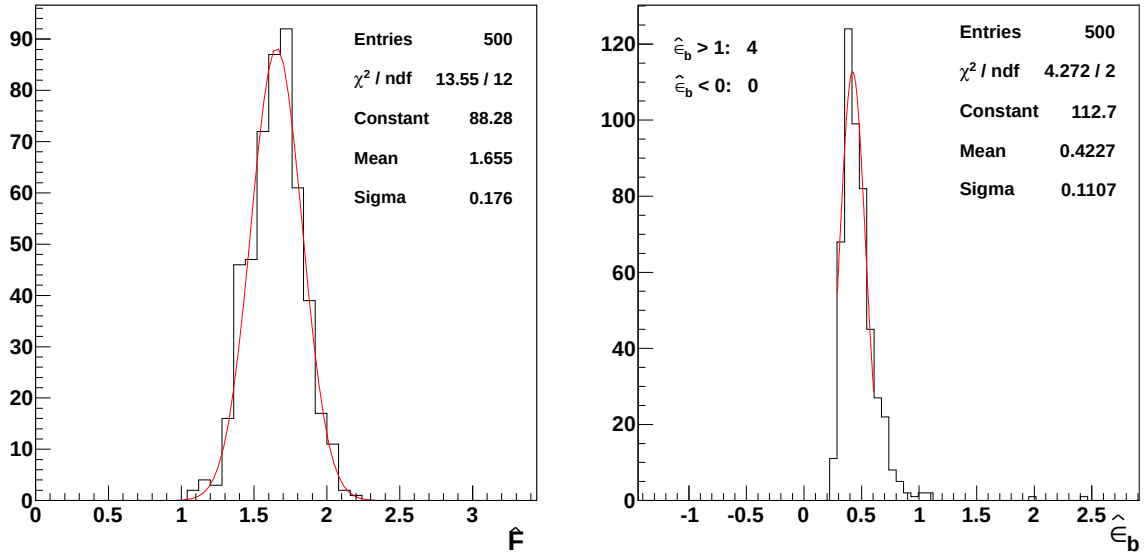
where $\epsilon_{b,N}$ is the b-tagging efficiency averaged over all pseudo-experiments and $\widehat{\epsilon_b^i}$ together with $\widehat{\delta\epsilon_b^i}$ are the efficiency and the uncertainty obtained from the i'th pseudo-experiment.

Figure 5.47 (a) shows the pull distribution of the $\widehat{\epsilon_b}$ where all values are computed within the pseudo-experiments. For the $(\eta; p_T)$ weights obtained from the full simulated sample, the pull distribution of the b-tagging efficiency at the medium working point is shown in Figure 5.47 (b). Although the shape of the pull is still different from a symmetric Gaussian, a better behavior is observed in the latter while in the former, neither the mean value nor the width is meaningful.



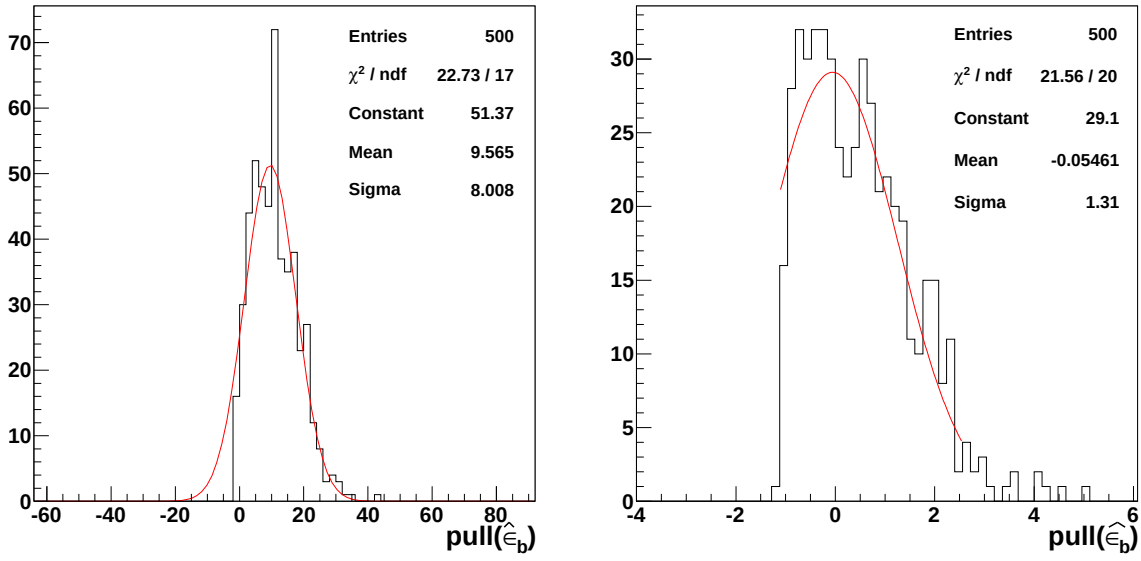
(a) (b)

Figure 5.45: The distribution of the scale factor, (a), and the b-tagging efficiency, (b), from ~ 500 pseudo-experiments at $100\,pb^{-1}$ integrated luminosity. The $(\eta; p_T)$ weights for the $\widehat{F}$ calculation are computed within the pseudo-experiments.



(a) (b)

Figure 5.46: The distribution of the scale factor, (a), and the b-tagging efficiency, (b), from 500 pseudo-experiments at $100\,pb^{-1}$ integrated luminosity. The $(\eta; p_T)$ weights for the $\widehat{F}$ calculation are computed using the full statistic of the simulated samples.



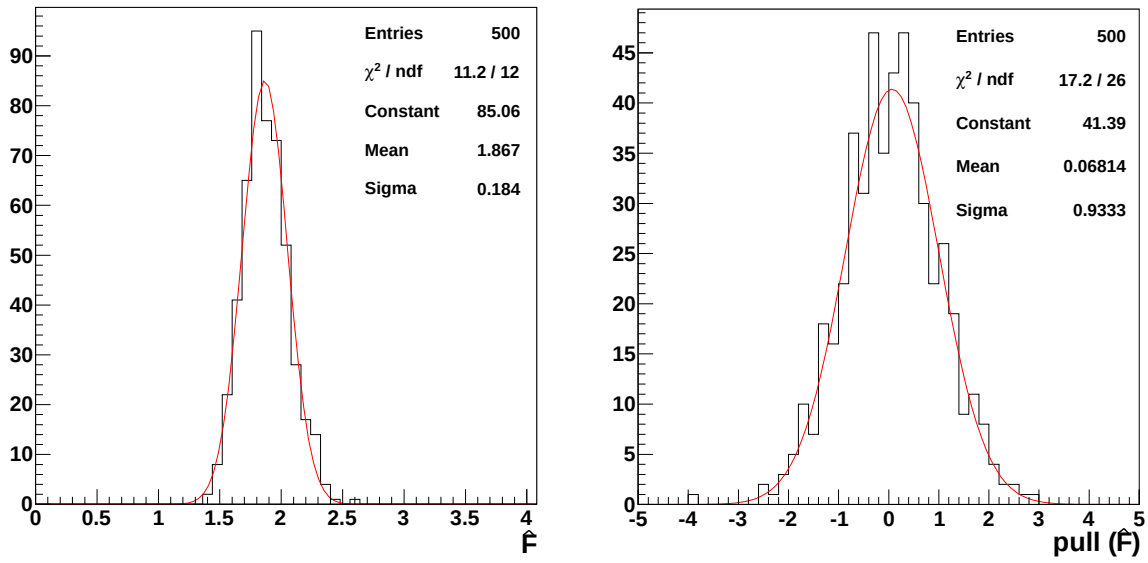
(a) (b)

Figure 5.47: The pull distribution of the b-tagging efficiency with the $(\eta; p_T)$ weights calculated within the pseudo-experiments, (a), and using the full statistic of the simulated samples, (b).

5.4.1 The statistical effect of anti-tagging

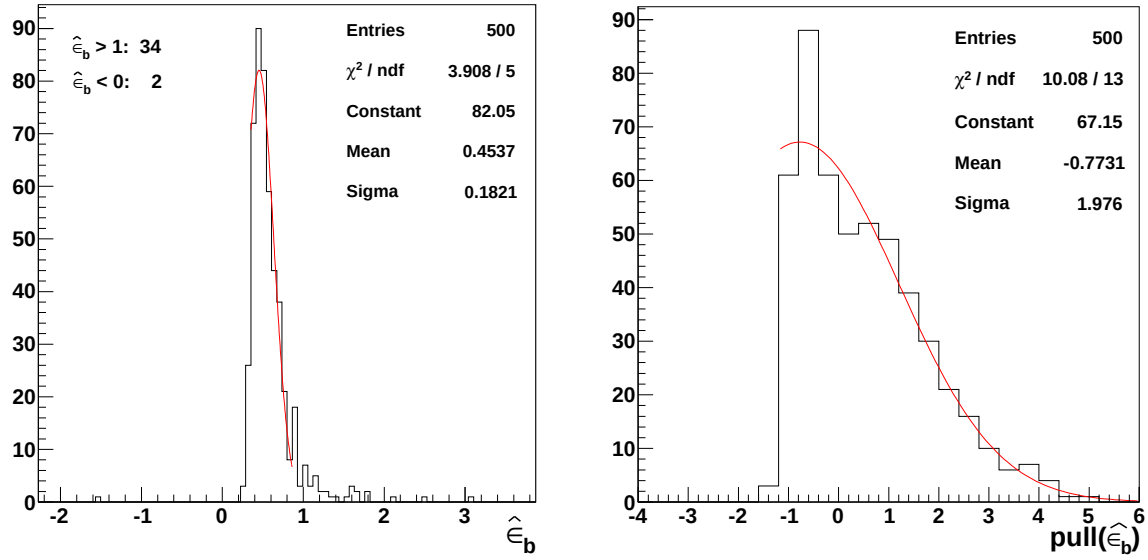
Anti-tagging the jets in the control sample has been discussed in Section 5.3.5 as a possible way for a better $\widehat{F}$ estimation. It has been shown that the anti-tagging does not give a good estimate for $\widehat{F}$ without the $(\eta; p_T)$ reweighting. An underestimation has been observed for the $\widehat{F}$ extracted from the anti-tagged and reweighted control jet sample. Because of this underestimation, the $\widehat{F}$ from the reweighted control sample without anti-tagging has been considered in the analysis.
The impact of anti-tagging on the statistical behavior of the $\widehat{F}$ is also investigated. Since the anti-tagging rejects more jets in the tail of the p_T distribution, the jets with nonsense weights can be avoided. This is specially useful when the amount of the statistics is not enough within the pseudo-experiment.
About 500 pseudo-experiments are performed for which the control jet samples contain the jets with $d_{TCHE} < 3$. The $(\eta; p_T)$ weights are computed within the pseudo-experiments. As it can be seen in Figure 5.48 (a), a better mean value is achieved for the $\widehat{F}$. The width of the pull distribution for $\widehat{F}$, Figure 5.48 (b), shows an overestimation of $\sim 4\%$ of the statistical uncertainties. In Figure 5.49 (a), it is shown that the efficiency has still non-physical values, although it is much less than the case for which the $\widehat{F}_i$'s were extracted from the default control samples. The pull distribution for the efficiency in Figure 5.49 (b), reflects the asymmetry observed in the efficiency distribution. It is tried to fit a Gaussian to the pull distribution. The resulting mean and width are -0.7 and 1.9, respectively.

It is worth investigating whether the use of $(\eta; p_T)$ weights extracted from the full



(a) (b)

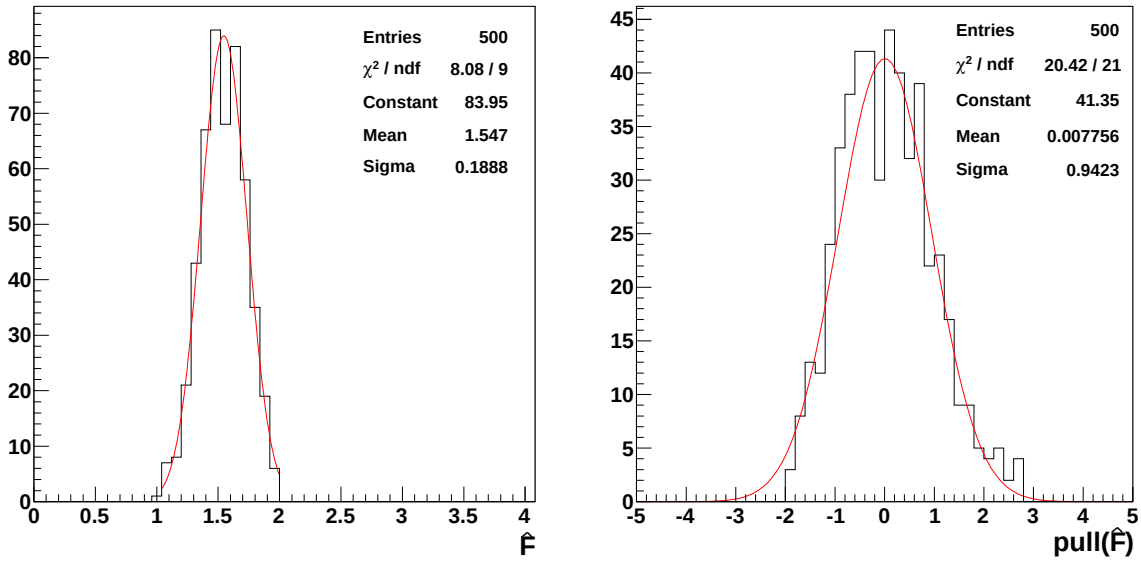
Figure 5.48: The scale factor, (a), and its pull distribution, (b), extracted from the anti-tagged control jet sample. The values are obtained from 500 pseudo-experiments at $100\,pb^{-1}$ integrated luminosity. The $(\eta; p_T)$ weights are computed within the pseudo-experiments.



(a) (b)

Figure 5.49: The b-tagging efficiency, (a), and its pull distribution, (b), calculated using the $\widehat{F}$ extracted from the anti-tagged control jet sample. The values are obtained from 500 pseudo-experiments at $100\,pb^{-1}$ integrated luminosity. The $(\eta; p_T)$ weights are computed within the pseudo-experiments.

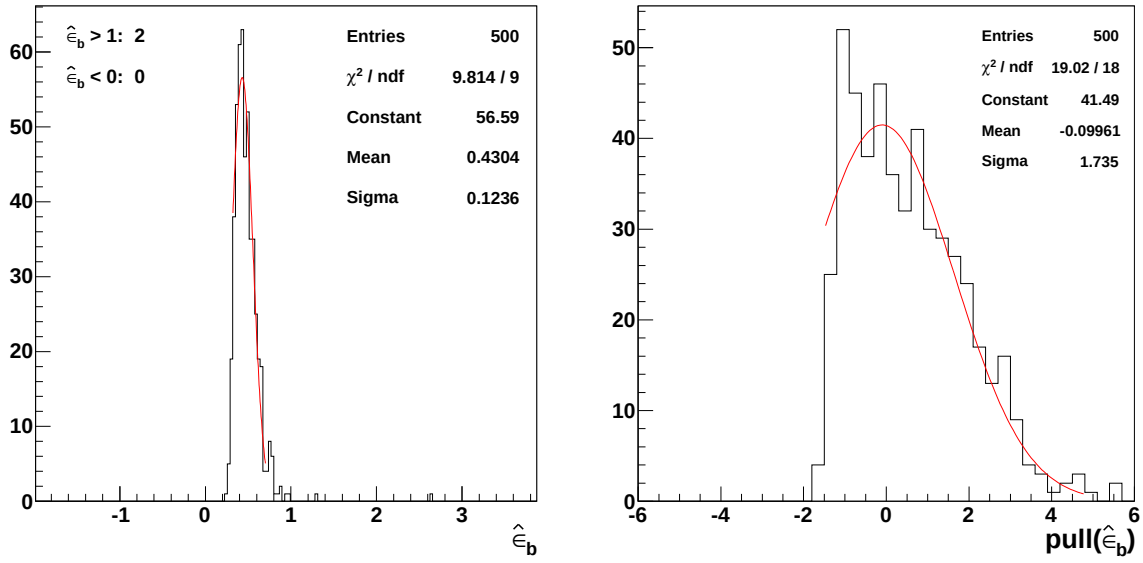
statistics of the simulated samples leads to an improvment in the statistical properties of the estimators for the case that the control jet sample is anti-tagged. The



(a) (b)

Figure 5.50: The scale factor, (a), and its pull distribution, (b), extracted from the anti-tagged control jet sample. The values are obtained from 500 pseudo-experiments at $100\,pb^{-1}$ integrated luminosity. The $(\eta; p_T)$ weights are obtained using the full statistics of the simulated samples.

data driven scale factor evealuated in 500 pseudo-experiments is illustrated in Figure 5.50 (a) where its pull distribution is shown in Figure 5.50 (b). The mean value of the estimator $\widehat{F} = 1.547$ should be compared with the value quoted in Table 5.7, $\widehat{F} = 1.562$. Considering the relative difference of $\sim 20\%$ which was obtained when the $(\eta; p_T)$ weights were calculated within the pseudo-experiments (see Figure 5.48 (a)), the estimation of the scale factor is improved. There is however no significant change in the width of the pull distribution, comparing to Figure 5.48 (b), where it is notable that the fit probability (χ^2/ndf) is improved. For the same set of pseudo-experiments, the estimated b-tagging efficiency and its pull distribution are given in Figure 5.51. It can be seen that the non-physical b-tagging efficiencies are almost disappeared. The shape of the pull distribution is still different from being Gaussian. The Gaussian fit is however performed with a better fit probability compared to Figure 5.49. The mean value and the width of the fitted Gaussian to the pull distribution of estimator $\widehat{\epsilon_b}$, are also improved.
Similar to the case of the default control jet sample (i.e. with no anti-tag requirement), one can conclude that with an insufficient amount of statistics, the computation of $(\eta; p_T)$ weights within the pseudo-experiments can spoil the statistical properties of the estimators.



(a) (b)

Figure 5.51: The b-tagging efficiency, (a), and its pull distribution, (b), calculated using the $\widehat{F}$ extracted from the anti-tagged control jet sample. The values are obtained from 500 pseudo-experiments at $100\,pb^{-1}$ integrated luminosity. The $(\eta; p_T)$ weights are obtained using the full statistics of the simulated samples.

5.4.2 Sampling distributions with higher statistics

From what has been previously discussed, it is difficult to draw a conclusion about the statistical properties of the estimators at $100\,pb^{-1}$. Since the observed statistical instability can be the consequence of the limited statistics at $100\,pb^{-1}$, another round of pseudo-experiments is performed for an integrated luminosity of $300\,pb^{-1}$ where all calculations are done internally following the fully data driven approach.

Figure 5.52 compares the $\widehat{\epsilon_b}$ pull distributions in $L = 100\,pb^{-1}$ and $L = 300\,pb^{-1}$ where the very wide pull distribution at $L = 100\,pb^{-1}$ has narrowed at higher integrated luminosity. The distributions of $\widehat{F}$ and its pull are shown in Figure 5.53 while Figure 5.54 illustrates $\widehat{\epsilon_b}$ and its pull distribution.

A better behavior is observed for the estimators at higher integrated luminosities means that the more accumulated data the better statistical stability of the method. The quantity χ^2/ndf, which is an indicator of the goodness of the fit is close to one reflecting a good fit probabilty for $300\,pb^{-1}$ integrated luminosity while it was worse at $L = 100\,pb^{-1}$. However one needs to be careful in looking at the width of the pull distributions since the available statistic for the W+jet sample is limited and the pseudo-experiments are correlated.

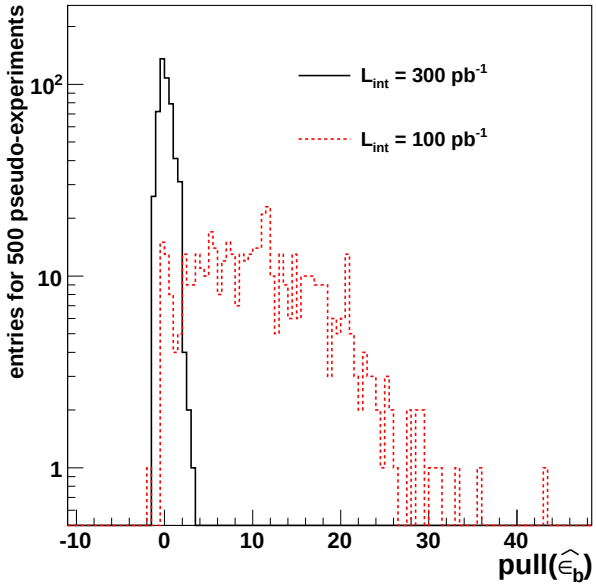
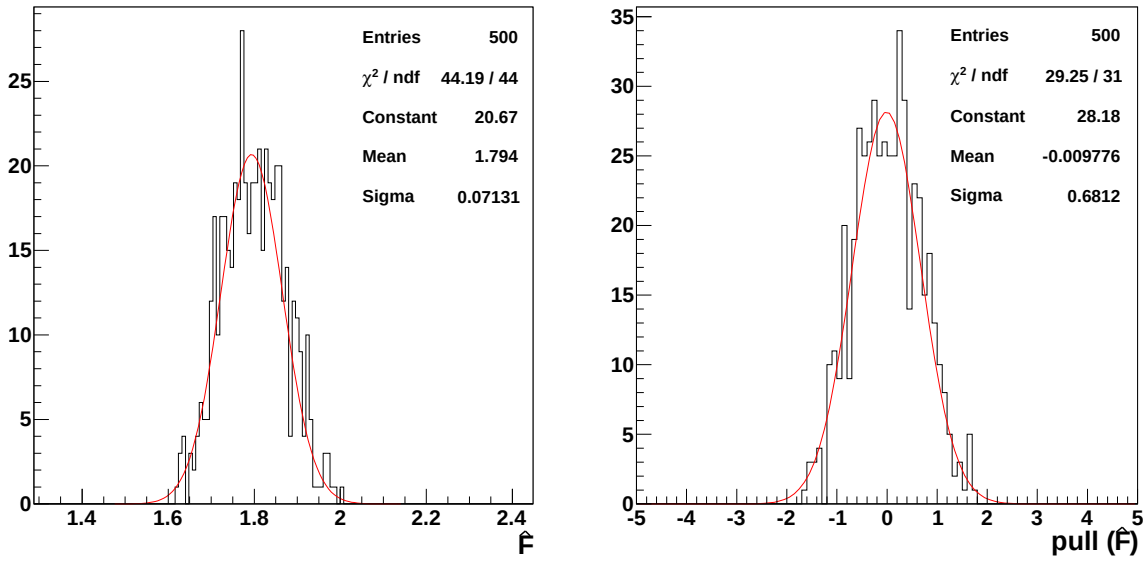
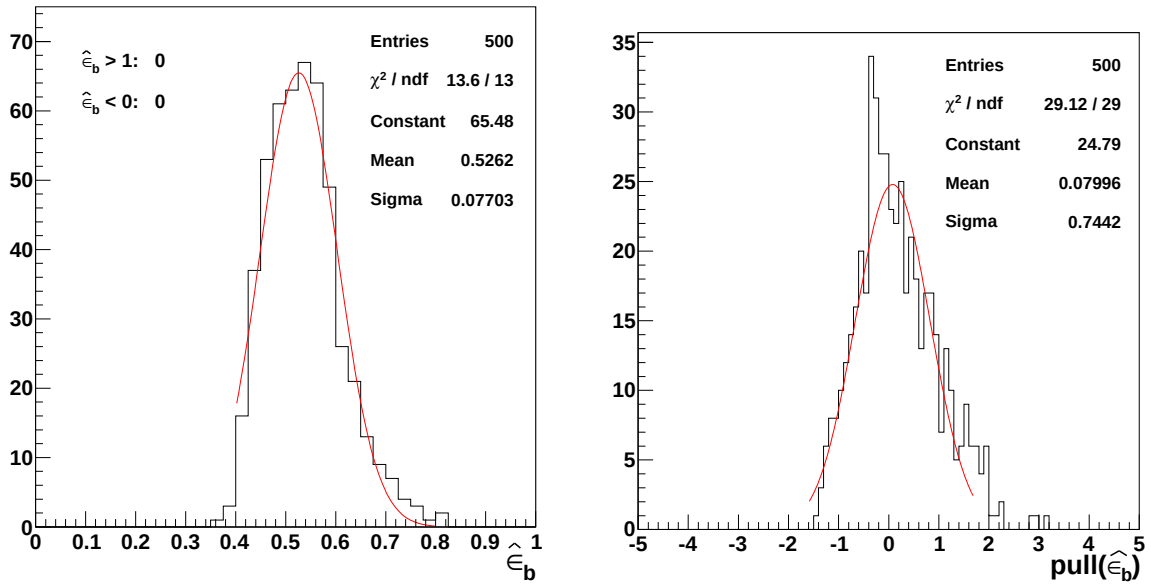


Figure 5.52: The b-tagging efficiency pull distribution $L = 100\,pb^{-1}$ and $L = 300\,pb^{-1}$ integrated luminosities. For each integrated luminosity, 500 pseudo-experiments are performed.



(a) (b)

Figure 5.53: The scale factor, (a), and its pull distribution, (b) obtained from ~ 500 pseudo-experiments at $300\,pb^{-1}$ integrated luminosity.



(a) (b)

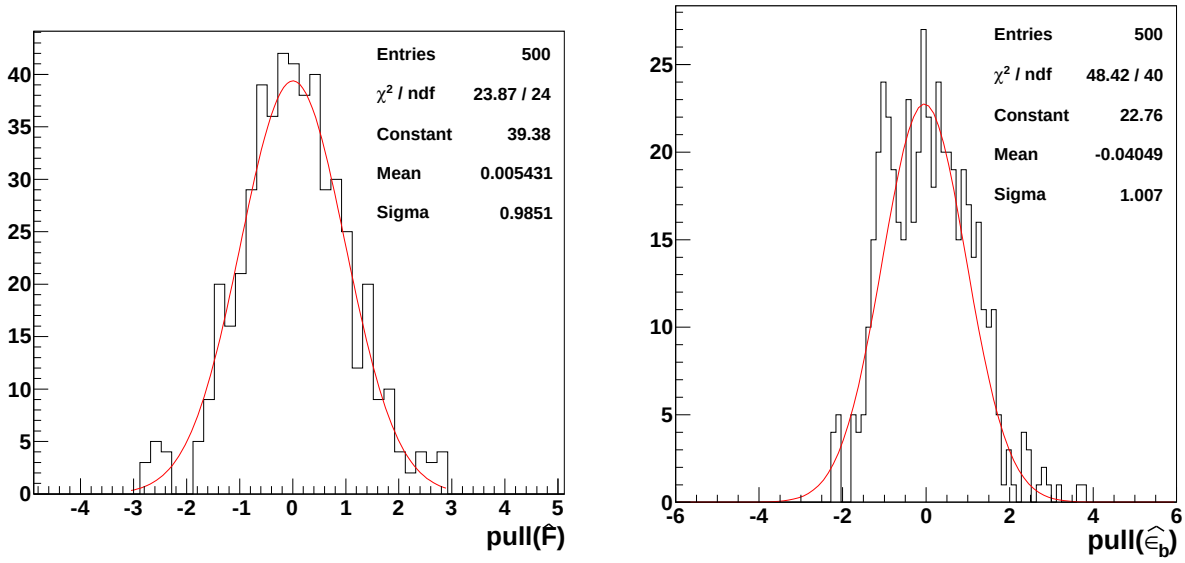
Figure 5.54: The b-tagging efficiency, (a), and its pull distribution, (b) obtained from ~ 500 pseudo-experiments at $300\,pb^{-1}$ integrated luminosity.

5.4.3 Sampling distributions for the $t\bar{t}$ event sample

The non-physical efficiencies occur at lower integrated luminosities and the pseudo-experiments become correlated at higher integrated luminosities. Hence to investigate the statistical properties of the estimators, the final check is to make the pseudo-experiments only out of the signal event sample at $L = 300\,pb^{-1}$. The fully data driven approach is followed i.e. the $(\eta; p_T)$ weights are calculated within the pseudo-experiments.

Figure 5.55 (a) shows the pull distribution for the data driven scale factor extracted from the $t\bar{t}$ events. The width of the Gaussian fit is a bit less than unity indicating about 1.5% uncertainty overestimation. This can be due to the conservative approach taken for the uncertainty calculation of $\widehat{F}$ after the $(\eta; p_T)$ reweighting.
The pull distribution for the b-tagging efficiency at the medium working point is illustrated in Figure 5.55 (b). Despite of the $\widehat{\delta F}$ overestimation, the uncertainty on the b-tagging efficiency seems to be well estimated. Comparing these results with those with background contributions at $L = 300\,pb^{-1}$, one can conclude that the statistical properties of the estimator is affected by the background contaminations.

The analysis described in Section 5.3, is performed on the $t\bar{t}$ signal together with the background processes. Therefore, the realistic correction on the statistical uncertainty is determined by those pull distributions that include backgrounds. Moreover, no anti-tagging requirement is applied on the control jet sample in the analysis. Hence, the pseudo-experiments with no anti-tagging request are the candidates to



(a) (b)

Figure 5.55: The pull distribution for the data driven scale factor, (a), and the b-tagging efficiency, (b), obtained from ~ 500 pseudo-experiments at $300\,pb^{-1}$ integrated luminosity within the $t\bar{t}$ signal events.

deduce the statistical properties of estimated scale factor and the b-tagging efficiency. On the other hand, unstabilties were obsereved at $L = 100\,pb^{-1}$ which were improved at the higher integrated luminosity, $L = 300\,pb^{-1}$. If the method was statistically stable at $L = 100\,pb^{-1}$, the width of the pull distributions would be expected to remain the same at higher integrated luminosities. Therefore, the width of the pull distributions at $L = 300\,pb^{-1}$ is taken as a more realistic choice to correct the statistical uncertainty on the scale factor by a factor of 0.68 and on the b-tagging efficiency by a factor of 0.74. Such corrections are applied in the final calculation of uncertainties in Section 5.5.7.

5.5 The evaluation of the systematic uncertainties

The statistical uncertainty has to be added to the systematic uncertainties to reflect the available knowledge about the reported measurements. The sensitivity of the method to its parameters together with the possible intrinsic bias on the estimated quantities are the first group of uncertainty sources investigated in the following.
The reconstruction issues are another cause for the systematic uncertainties. The energy mis-calibration of the jets is the most important systematic source from this group which can affect the final results of the analysis.
The non $t\bar{t}$ physics processes as well as the contamination from other $t\bar{t}$ final states are another sources of systematics. The main lack of knowledge in this case is the total cross section of the background processes which are uncertain from the theoretical and

experimental point of view.
The other group of the uncertainties relates to the modeling of the simulated samples. A special set of parameters is used for the event modeling at generator level (see Section 3.3.1). The variation of each parameter would lead to the variation of the final results, hence introducing a systematic uncertainty. In addition, one can take into account the systematic uncertainties arising from the differences between the event generators.
The systematic uncertainties influence the performance of the method, i.e. the ability to estimate the true values for the data driven scale factor and the b-tagging efficiency. Since both the estimated and the true value of the estimators fluctuate according to a given systematic source, the variation in the value of $(e_{obs} - e_{exp})/e_{exp}$ is investigated where e stands for either the scale factor or the b-tagging efficiency.
For the b-tagging efficiency, the relative difference is calculated at the medium b-tagging working point (see Section 4.4.4) with the efficiency of about $\sim 50\%$.
The evaluated uncertainties from the mentioned sources of variations are combined together and with the statistical uncertainty to give a more realistic picture about the performance of the method.

5.5.1 The intrinsic bias and the robustness of the method

To check the possible bias on the method, the events containing the physics objects truly coming from the semi-electron $t\bar{t}$ decay are needed. Thus, only events with the four leading jets matched to the quarks from the $t\bar{t}$ semi-electron final state are considered. The matching could have been extended to the electron as well to prepare a "fully matched" event sample. However, regarding the high efficiency of the electron reconstruction this request is not really necessary.
The b-tagging efficiency for the medium working point is obtained using the fully data driven approach and is compared to the efficiency of the b-quark jets. Summarized in Table 5.9, the $\Delta\epsilon/\epsilon_b$ is found less than $(0.8 \pm 4)\%$ and the relative bias on the F is $\sim (1 \pm 2)\%$. Due to the limited statistics after the jet matching, a large statistical uncertainty accompanies the negligible bias on the b-tagging efficiency. This uncertainty is instead quoted in the final combination of uncertainties to cover for the lack of knowledge arising from the statistical limitation.
An important parameter of the method is the m_{ej} value of the boundaries which divide the b-candidate jet sample into the b-depleted (Right) and b-dominated (Left) subsamples. The variation of the boundaries has been performed considering the fact that the lowest m_{ej} value, corresponding to the far left boundary, cannot be lower than some threshold because of the p_T cut on the electron and the b-jet candidate. The far right boundary which is basically in the tail of the distribution has to be in a meaningful range regarding the available statistics in the tail. Besides, the intercept between the Left and the Right region is in general not too far from the m_{ej} value at which the distribution falls sharply.
The boundaries have been varied independently by at most 10%. The method is found to be stable with respect to the changes in this parameter. The final relative fluctuation in both scale factor and the efficiency have been less than $< 0.5\%$. This is therefore not included in the final systematic calculation.

F_{exp}	$\widehat{F}$	$\Delta F/F_{exp}$
3.876 ± 0.093	3.919 ± 0.058	0.011 ± 0.028
ϵ_b^{exp}	$\widehat{\epsilon_b}$	$\Delta\epsilon/\epsilon_b$
0.518 ± 0.014	0.514 ± 0.017	$(-7.72 \pm 42.4) \times 10^{-3}$

Table 5.9: The estimated intrinsic bias of the method on the scale factor and the b-tagging efficiency at the medium working point. Although the residual biases are small, due to the limited statistics the statistical uncertainties on the values are conservatively considered.

5.5.2 The influence of the jet energy mis-calibration

The Anti-κ_T jet reconstruction algorithm has been used in this thesis to reconstruct the jets. The energy of the jets has undergone both the relative (η) and the absolute (p_T) energy corrections (see Section 4.3). However, the reconstructed jet energy can still be mis-estimated. This possible fluctuation is considered as a uniform scaling up/down for the jet energy, the Jet Energy Scale variation.
This is accounted as a source of systematic uncertainty in the analysis since it can affect the method from different aspects. The m_{ej} distribution in both the b-candidate and the control jet samples would change and it would result in the variation of the scale factor as well as the b-tagging efficiency itself.
Both the scale factor and the b-tagging efficiency rely on the construction of the b-dominated and b-depleted jet samples. These samples are made using the boundaries on the m_{ej} distribution. The boundaries are chosen based on the expected m_{ej} shape and their values can fluctuate with the energy variation of the jets. Hence, the method seems to be able to handle the jet energy scale within itself in a consistent way.
What is investigated here is the effect of the jet energy scale uncertainty under the assumption of the fixed boundaries. To account for an uncertainty arising from a relative $\pm\alpha$ variation in the jet energy, the energy of the jets is up-/down-scaled as follows:

$$p_{jet}^{\pm\alpha} = (1 \pm \alpha)(\vec{p_{jet}}, E_{jet}), \tag{5.12}$$

where α is taken $\pm 10\%$ for this analysis. The scale factor and the b-tagging efficiency are re-evaluated over all signal and background processes for each scaling scenario. Needless to say that the constraints in Equation 5.1 have also been re-estimated. The b-tagging efficiency as well as the data driven scale factor extracted from the energy scaled jet samples are summarized in Table 5.10 where the variation of the estimators

with respect to the nominal sample is also presented.
The uncertainty on the b-tagging efficiency seems larger for the positive variation of the jet energy. To examine the JES effect on the data driven scale factor, the $\Delta F/F$ is extracted for both scaling scenarios and compared to the value calculated in the nominal jet sample. Events with the down-scaled jet energy, lead to a larger fluctuation in the value of $F_{data\,driven}$. Due to the limited size of the samples, the JES systematics are of the size of the statistical uncertainties for both the $\widehat{\epsilon_b}$ and the $\widehat{F}$ estimators. However, since the samples with the scaled jet energies are highly correlated with each other and with the nominal sample, this statistical uncertainty is not included in the evaluation of jet energy scale systematic uncertainty.
On the $\widehat{\epsilon_b}$ value, a relative systematic uncertainty of 4% is quoted to reflect the largest

	F_{exp}	$\widehat{F}$	$(\Delta F/F_{exp})$ $-(\Delta F/F_{exp})_{nom}$
nominal	1.695 ± 0.049	1.669 ± 0.043	—
$\alpha = -10\%$	1.650 ± 0.059	1.673 ± 0.047	-0.029
$\alpha = +10\%$	1.652 ± 0.040	1.612 ± 0.038	0.009
	ϵ_b^{exp}	$\widehat{\epsilon_b}$	$(\Delta\epsilon/\epsilon_b)$ $-(\Delta\epsilon/\epsilon)_{nom}$
nominal	0.48 ± 0.04	0.47 ± 0.05	—
$\alpha = -10\%$	0.49 ± 0.04	0.50 ± 0.05	0.04
$\alpha = +10\%$	0.49 ± 0.04	0.47 ± 0.04	0.02

Table 5.10: The data driven scale factor together with the b-tagging efficiency for different jet energy scenarios. The last column contains the variation of the estimators with respect to the nominal situation.

effect out of the -10% and $+10\%$ change in the JES. It should be noted however that the JES is currently much better calibrated than 10%. Hence it can be expected that this realtive systematic uncertainty can be reduced to the $1-2\%$ level.

5.5.3 The uncertainty on the backgrounds cross section

The variation on the cross section of the signal and background processes can change the composition of the final selected sample, i.e. the S/B ratio. The modification in the S/B can be studied either by changing the rate of the signal events or by varying the background cross sections. To account for the deviation of different background

contributions, the latter is followed. This choice is also more realistic since the signal cross section is better estimated/measured in particular comparing to the QCD multi-jets events.
Regarding the event selection in Table 5.3 two major backgrounds are the QCD multi-jets and the W+jets processes. One may add the other $t\bar{t}$ final states to the list. The basic idea is to variate the total background cross section and estimate the b-tagging efficiency together with the data driven scale factor. For the case of the QCD background, first the estimators are evaluated including the QCD background. This background contamination has not been included in the whole analysis for the $\widehat{\epsilon_b}$ estimation because of the limited size of the sample and the large event weights as explained in Section 5.2.1. The inclusion of the QCD multi-jets, changes the data driven scale factor by less than 1%. The b-tagging efficiency at the medium working point is stable within the available statistics.
The cross section of the QCD multi-jets is then enhanced by 100% and the method is redone. Due to the limited size of the event samples, the uncertainties are statistically limited. However, because of the existing correlation similar to the JES case, this statistical uncertainty is not quoted for the final systematic calculation.
The effect of the W+jet background on the final b-tagging efficiency has been studied in detail in Section 5.3.3 and it has been shown that the p_T reweighting method is able to handle this contamination. However for the sake of completeness, the uncertainty arising from this background is examined by 30% variation [188] of the σ_{W+jets} and it results in an uncertainty of 1% for $\widehat{\epsilon_b} = 50\%$. The relative change on the data driven scale factor with respect to the nominal backgrounds composition is 7% which means that the estimator $\widehat{F}$ is more sensitive to the W+jets contamination. This sensitivity however does not affect the estimation of the b-tagging efficiency too much.
The effect of the other $t\bar{t}$ final states is found to be negligible. Table 5.11 contains the uncertainties on F and ϵ_b resulting from the altered cross section of the W+jets and QCD multi-jets backgrounds. A relative uncertainty of 2% is finally quoted as systematic uncertainty due to the fluctuation in the background cross sections.

5.5.4 The model dependent fluctuations

Many parameters used to model the proton-proton collisions need to be tuned with experimental data. For some of these parameters, it is crucial to investigate the effect of their fluctuations on the physics estimators. As explained in Section 3.3.1, three sets of samples are generated to study separately the influence of variations in

- the ISR/FSR content of the event,
- the factorization scale $Q^2 = \mu_F^2$,
- the energy threshold of the matrix element and praton shower matching,

where in each set the increasing and decreasing parametrization is provided. The estimators resulting from these uncorrelated samples will be compared with the outcome of a sample produced with the nominal parametrization. An extra sample is simulated as detailed in Section 3.4.1 to study the effect of pileup.

	F_{exp}	$\widehat{F}$	$(\Delta F/F_{exp})$ $-(\Delta F/F_{exp})^*$
$\sigma_{qcd} \equiv \sigma^*_{qcd}$	1.696 ± 0.058	1.655 ± 0.043	—
$\sigma_{qcd} \equiv 0$	1.695 ± 0.049	1.669 ± 0.043	-0.009
$\sigma_{qcd} \equiv 2 \times \sigma^*_{qcd}$	1.696 ± 0.075	1.645 ± 0.035	0.006
$\sigma_{W+jets} \equiv \sigma^*_{W+jets}$	1.695 ± 0.049	1.669 ± 0.043	—
$\sigma_{W+jets} \equiv 0.7 \times \sigma^*_{W+jets}$	1.678 ± 0.051	1.569 ± 0.041	0.05
$\sigma_{W+jets} \equiv 1.3 \times \sigma^*_{W+jets}$	1.709 ± 0.048	1.556 ± 0.045	0.07
	ϵ_b^{exp}	$\widehat{\epsilon_b}$	$(\Delta\epsilon/\epsilon_b)$ $-(\Delta\epsilon/\epsilon)^*$
$\sigma_{qcd} \equiv \sigma^*_{qcd}$	0.48 ± 0.04	0.48 ± 0.04	—
$\sigma_{qcd} \equiv 0$	0.48 ± 0.04	0.47 ± 0.05	0.02
$\sigma_{qcd} \equiv 2 \times \sigma^*_{qcd}$	0.48 ± 0.04	0.48 ± 0.04	0
$\sigma_{W+jets} \equiv \sigma^*_{W+jets}$	0.48 ± 0.04	0.47 ± 0.05	—
$\sigma_{W+jets} \equiv 0.7 \times \sigma^*_{W+jets}$	0.48 ± 0.04	0.47 ± 0.04	0
$\sigma_{W+jets} \equiv 1.3 \times \sigma^*_{W+jets}$	0.48 ± 0.04	0.46 ± 0.04	-0.02

Table 5.11: The effect of the possible mis-estimation of the major background processes on the data driven scale factor and the b-tagging efficiency at the medium working point. The fluctuation of the estimators are shown in the last column. The notation σ^* stands for the nominal cross section estimation presented in Table 3.2.

Initial and Final State Radiation

It was discussed in Section 5.2.1 that the presence of the jets originated from radiations, the initial state radiation in particular, results in the large tails for the χ^2_{min} distribution. These wrong combinations can influence the data driven scale factor directly and consequently the m_{ej} distribution.
Since no cut is applied on the χ^2_{min} value, the wrong combinations have also been allowed to participate in the efficiency measurement. Therefore an investigation of the ISR/FSR systematic uncertainties is necessary.
The $t\bar{t}$ samples with an increased and decreased radiation content together with the sample generated with nominal parameters are used to study the effect of the ISR and FSR (see Section 3.3.1).
As it can be seen in Table 5.12, the expected and the observed values for the data driven scale factor, F_{exp} and $\widehat{F}$ respectively, are individually compatible between the samples with different radiation contents. Regarding the available statistics, the b-tagging efficiencies at the medium working point show no difference either. It seems that the subtraction of the scaled b-depleted jet sample from the b-dominated one performs well in erasing the jets from radiation.
A relative uncertainty of 1.4%, the larger estimate, is taken as the systematic uncer-

	F_{exp}	$\widehat{F}$	$(\Delta F/F_{exp})$ $-(\Delta F/F_{exp})_{nom}$
nominal	1.642 ± 0.029	1.815 ± 0.022	—
more ISR/FSR	1.684 ± 0.030	1.810 ± 0.023	0.030
less ISR/FSR	1.660 ± 0.030	1.807 ± 0.023	0.017
	ϵ_b^{exp}	$\widehat{\epsilon_b}$	$(\Delta\epsilon/\epsilon_b)$ $-(\Delta\epsilon/\epsilon)_{nom}$
nominal	0.483 ± 0.005	0.534 ± 0.005	—
more ISR/FSR	0.493 ± 0.005	0.54 ± 0.006	-0.010
less ISR/FSR	0.477 ± 0.005	0.534 ± 0.005	0.014

Table 5.12: The systematic uncertainties due to the variation of the radiation content.

tainty due to the amount of radiation in the event.

Matching threshold between the matrix element and the parton shower

The variation of matching threshold between the matrix element and the jets produced by parton showering can affect the method presented here in two ways. First, since the events are kept only if all the jets are matched to the partons in matrix element, some events may be lost by tightening the matching threshold and vice versa. On the other hand, the same events may contribute to the analysis with different jet multiplicities. In some cases this can change the configuration of the four leading jets in the event. Separated for the higher and lower matching threshold effects, the resulting estimators together with the uncertainties are listed in Table 5.13. For a higher matching threshold, both the estimated and the expected scale factors are systematically smaller where the b-tagging efficiency on the other hand is increased. The lower matching threshold is less effective. The larger relative uncertainty on the $\widehat{\epsilon_b}$ estimator, 1.8%, is considered as the systematic uncertainty coming from the variation of the matrix element and the parton shower matching threshold.

	F_{exp}	$\widehat{F}$	$(\Delta F/F_{exp}) - (\Delta F/F_{exp})_{nom}$
nominal	1.642 ± 0.030	1.815 ± 0.022	—
higher threshold	1.509 ± 0.031	1.715 ± 0.025	-0.031
lower threshold	1.691 ± 0.038	1.841 ± 0.029	0.017
	ϵ_b^{exp}	$\widehat{\epsilon_b}$	$(\Delta\epsilon/\epsilon_b) - (\Delta\epsilon/\epsilon)_{nom}$
nominal	0.483 ± 0.005	0.534 ± 0.005	—
higher threshold	0.488 ± 0.005	0.548 ± 0.006	-0.017
lower threshold	0.485 ± 0.006	0.527 ± 0.006	-0.018

Table 5.13: The uncertainty arising from the variation of the matching threshold between the matrix element partons and the jets from the parton showers. The last column shows the residual variations of the estimators.

Variation of the factorization scale

As defined in Section 3.2, the factorization scale, μ_F^2, factorizes the short-distance physics from the non-perurbative long-distance interaction. The samples used for the analysis in this chapter are generated by `MadGraph` for which the factorization scale is defined by Equation 3.12. The factorization scale is therefore changing on an event

by event basis depending on the momenum of the generated partons.
Events generated with higher μ_F^2 are accompanied by more radiations and would end up in relatively more energetic final states. Therefore, not only the wrong combinations are more probable but also the m_{ej} distributions in the b-candidate and the control jet sample can be changed. The scale factor and the b-tagging efficiency derived from the samples with altered factorization scale in addition to the resulting uncertainties are presented in Table 5.14. The $\widehat{\epsilon_b}$ efficiency increases for the scaled down μ_F^2 while scaling up the μ_F^2 factor results in some efficiency loss. For the final calculation of systematic uncertainties, the relative uncertainty of 1.2% which is introduced by the down scaled μ_F^2 is taken into account. A relative bias of ~ 0.05 is introduced to the data driven scale factor for the reduced μ_F^2 value.
This uncertainty is not completely uncorrelated to the ISR/FSR from the radiation point of view. However, it is difficult to estimate the amount of correlation. Thus they are added in quadrature keeping in mind the possible changes due to the correlation term.

	F_{exp}	$\widehat{F}$	$(\Delta F/F_{exp}) - (\Delta F/F_{exp})_{nom}$
nominal	1.642 ± 0.030	1.815 ± 0.022	—
μ_F^2 Up	1.730 ± 0.042	1.897 ± 0.031	0.009
μ_F^2 Down	1.464 ± 0.029	1.695 ± 0.025	-0.052
	ϵ_b^{exp}	$\widehat{\epsilon_b}$	$(\Delta\epsilon/\epsilon_b) - (\Delta\epsilon/\epsilon)_{nom}$
nominal	0.483 ± 0.005	0.534 ± 0.005	—
μ_F^2 Up	0.481 ± 0.006	0.527 ± 0.006	-0.010
μ_F^2 Down	0.491 ± 0.006	0.537 ± 0.007	-0.012

Table 5.14: The effect of the variation of the factorization scale on the F and the b-tagging efficiency.

The effect of pileup

Extra proton-proton collisions in the same bunch crossing with their associated tracks would influence the performance of the b-tagging algorithms. Hence the study of the pileup effects is relevant for this analysis. For the 2010 data taking, the number of

pileup vertices added to the simulated events has been on average equal to one. This leads to a relative uncertainty of 1.3% on the estimated b-tagging efficiency which is quoated in the final calculation of the systematic uncertainty. The data driven scale factor remains stable within the statistical uncertainties. The details on the variation of the estimators due to the pileup effect can be found in Table 5.15

	F_{exp}	$\widehat{F}$	$(\Delta F/F_{exp})$ $-(\Delta F/F_{exp})_{nom}$
nominal	1.642 ± 0.030	1.815 ± 0.022	—
pileup	1.671 ± 0.030	1.788 ± 0.022	0.035
	ϵ_b^{exp}	$\widehat{\epsilon_b}$	$(\Delta\epsilon/\epsilon_b)$ $-(\Delta\epsilon/\epsilon)_{nom}$
nominal	0.483 ± 0.005	0.534 ± 0.005	—
pileup	0.475 ± 0.005	0.519 ± 0.005	-0.013

Table 5.15: The systematic uncertainty on the b-tagging efficiency and the data driven scale factor, arising from pileup.

5.5.5 The variations imposed by different event generators

The simulated samples used in this analysis are generated by MadGraph in which the matrix elements are calculated at leading order. The calculation of the higher order corrections is however implemented in MC@NLO as addressed in Section 3.3. To investigate the effect of the higher order corrections, the method has been applied on a $t\bar{t}$ sample generated by MC@NLO. The results are compared with the method outcome on the nominal MadGraph generated $t\bar{t}$ events. Since both event samples have equally been simulated, one can extract the effect of the higher order corrections coming from the differences at generator level [12].
As presented in Table 5.16, the estimated b-tagging efficiency at the medium working point does not show a big difference while the data driven scale factor is systematically about 1% smaller in the MC@NLO generated event sample. The statistcal uncertainties on the evaluated fluctuations are also quoted in the last column of Table 5.16 to demonstrate the statistical limitation on the relative systematic uncertainties. To be

[12] The two generators differ in the input m_t by the relatively small value of $\sim 1.8\,\text{GeV}$. This can also affect the physics estimators, although this effect is expected to be small.

accounted in the final combination of the systematic uncertainties, a conservative choice is made by taking the statistical uncertainty of 1.8% instead of the relative systematic uncertainty on the $\widehat{\epsilon_b}$ estimator.

	F_{exp}	$\widehat{F}$	$(\Delta F/F_{exp})$ $-(\Delta F/F_{exp})_{MadGraph}$
MadGraph	1.607 ± 0.027	1.953 ± 0.017	—
MC@NLO	1.443 ± 0.028	1.736 ± 0.026	-0.012 ± 0.037
	ϵ_b^{exp}	$\widehat{\epsilon_b}$	$(\Delta\epsilon/\epsilon_b)$ $-(\Delta\epsilon/\epsilon)_{MadGraph}$
MadGraph	0.485 ± 0.004	0.490 ± 0.005	—
MC@NLO	0.536 ± 0.005	0.540 ± 0.005	-0.002 ± 0.018

Table 5.16: The effect of the higher order corrections in the matrix elements at generator level on the method estimators. The higher order calculations are implemented in MC@NLO while for MadGraph, only the leading orders are considered.

5.5.6 Other sources for systematics

The systematic uncertainties that are discussed are not claimed to cover all possible systematic effects but include those which are expected to have a higher effect on the final results. There are also systematic sources to which the method is not sensitive.

Integrated Luminosity : The method is only sensitive to the relative rate of the signal and background processes, hence it is stable with a variation of the integrated luminosity.

The parton distribution function : The uncertainties on the parton distribution functions (see Section 3.2.2) in the $t\bar{t}$ event sample have negligible effect, too. The variation in p.d.f changes the interaction probability for two partons carrying the momentum fractions of x_1 and x_2. This change ultimately appears in a weight for the event while it does not touch the event topology. In fact within an event, the jets that end up in the b-candidate sample, the b-enriched or the b-depleted, together with the jets in the control sample would take the same weight. Therefore, in a single event no difference is expected.

On the other hand, the change in the relative weights between two events is similar to a variation of the background cross sections. The interesting distributions may slightly change but regarding the data-driven character of the method and the tools like the p_T reweighting which proves to be promising even in worse situations, this fluctuations can be covered.
In another view, since in the method everything is performed consistently using the information from data, no strong dependence on the p.d.f variations at generator level is expected. In the previous works [189] this uncertainty has been found to be negligible even for not fully data driven approaches.

5.5.7 Combined uncertainty on the b-tagging efficiency

The goal of this subsection is to combine the systematic uncertainties on the b-tagging efficiency evaluated in the previous subsections together with the statistical uncertainty. Starting from the intrinsic bias on the method, a conservative choice is made by taking the statistical uncertainty instead of the relative bias. The variation due to the down-scaled jet energy which is larger, is taken as the systematic imposed by the jet energy scale. Neither for the jet energy scale nor for the background cross sections, the statistical uncertainty is considered since the samples are highly correlated. To be conservative, within the model dependent uncertainties the larger variation is always taken into account. The difference between the event generators, mainly due to the calculation of the higher order corrections, is finally added to the total systematic uncertainty where the statistical uncertainty on the variation is conservatively taken instead.
Apart from the intrinsic bias and the uncertainty imposed by different event generators for which the statistical uncertainty is taken, the jet energy scale seems to be source of the dominant systematic uncertainty. Considering the recent improvements in the JES calibration, this uncertainty is expected to be reduced. Thereafter, the fluctuation in the background cross sections introduces the largest relative systematic uncertainty which is 2% at the medium working point.
It is notable that the matrix element and parton shower matching threshold imposes the largest systematic uncertainty within the model dependent sources.
The same procedure is followed for the loose ($\epsilon_b \approx 75\%$) and the tight ($\epsilon_b \approx 25\%$) working points to estimate the influence of the systematic sources on different b-tag cuts. An overview of the systematics for the three working points is provided in Table 5.17 where the relative statistical uncertainties are also included.

5.6 First look at the data collected in 2010

A method to estimate the b-tagging efficiency within the semi-electron final state of $t\bar{t}$ event has been developed. This section is devoted to the first application of the method in the electron channel on the LHC data collected by the CMS experiment in 2010 at 7 TeV center of mass energy.
The full set of 2010 data, $36 \pm 4\, pb^{-1}$ integrated luminosity, categorized into two primary

	loose $\epsilon_b^{exp} = 0.75$	medium $\epsilon_b^{exp} = 0.50$	tight $\epsilon_b^{exp} = 0.25$
statistical ($100\,pb^{-1}$)	5.1%	7.2%	11%
intrinsic bias	5.8%	4.2%	6%
JES	3%	4%	1%
background cross section	2%	2%	1%
ISR/FSR	1.5%	1.4%	$\leq 1\%$
factorization scale	1.4%	1.2%	$\leq 1\%$
ME-PS matching	1.9%	1.8%	1.8%
pileup	1.3%	1.3%	1%
event generators	1.2%	1.8%	4.1%
total systematic	7.5%	7%	8.6%
combined	9%	10%	14%

Table 5.17: Overview of the relative systematic uncertainties arising from different sources together with the relative statistical uncertainty for the loose, medium and tight b-tagging working points. The relative statistical uncertainty is correted for the overestimation observed in the pull distribution (see Section 5.4).

datasets (see Section 2.3.2) of

- `/EG/Run2010A-Nov4ReReco-v1/`
- `/Electron/Run2010B-Nov4ReReco-v1/` ,

are taken for the analysis. On the simulation side, the so-called Fall10 samples generated by `MadGraph` are used (see Section 3.4 for more explanation). To reduce the size of the samples both in data and simulation, events are asked for the presence of at least one electron with $p_T > 30\,GeV$.

5.6.1 Selection of the "top-like" events

The same criteria as in [25] are used to select events with a signature similar to the semi-electron final state of $t\bar{t}$.
After being filtered based on the "GOOD"ness of runs [13], the data events within the primary datasets are further triggered according to the electron trigger [14] paths listed in Table 5.18. The variety in the electron triggers is the consequence of the electron trigger evolution during the 2010 data taking. It can be seen in the table that tighter criteria have been used for more recent runs. The instantaneous luminosity has been increasing steeply and to select as many signal candidate events as possible, new trigger paths had to be defined for data. As a results, many of the triggers in Table 5.18 are not available in the simulated samples. Therefore no trigger selection is applied on the simulation side although the final event yield is corrected for the trigger efficiency which is about $\sim 98\%$, (see Section 4.2).

run range	electron trigger path name
< 140041	`HLT_Ele10_LW_L1R`
$140041 - 143962$	`HLT_Ele15_SW_L1R`
$143963 - 146427$	`HLT_Ele15_SW_CaloEleId_L1R`
$146428 - 147116$	`HLT_Ele17_SW_CaloEleId_L1R`
$147117 - 148818$	`HLT_Ele17_SW_TightEleId_L1R`
$148819 - 149180$	`HLT_Ele22_SW_TighterEleId_L1R_v2`
> 149180	`HLT_Ele22_SW_TighterEleId_L1R_v3`

Table 5.18: The list of the run ranges and the corresponding electron triggers used for the $t\bar{t}$ analyses in 2010 data taking.

The first vertex in the primary vertex collection, ordered by ndof defined in Equation 2.5, is checked to be positioned in a circle with radius $\rho < 2\,\text{cm}$ around the beam line. The $|z| < 24\,\text{cm}$, $ndof > 4$ and `!isFake` are further requirements on the primary vertex.

[13] The procedure of the run certification and the usage of GOOD runs at analysis level via the JSON files are described in Section 2.3.4.

[14] The basic definition of the electron trigger is given in Section 2.2.4.

Requirements on the electron candidate

The event has to contain exactly one electron with the following criteria:

- p_T >30 GeV and $|\eta| < 2.5$, with the supercluster out of the EE-EB gap
- $|z_{pv} - z_e| < 1\,\text{cm}$
- identified as and electron according to the electron identification working point with the efficiency of 70% (see Table 4.1).
- relative isolation < 0.1 (see Equation 4.1).
- rejecting electrons from conversion by asking for
  - $d_0(b.s.) < 200\,\mu m$ where b.s. denotes for the beam spot
  - no layer without hit in the pixel detector
  - no partner track in a cone of size $R = 0.3$. The partner track veto is applied based on the $|\Delta\cot(\Theta)| < 0.02$ and $|\text{Dist}| < 0.02$ requirements.

Loose muon and Z boson veto

Events containing a muon candidate with $p_T > 10\,\text{GeV}$, $|\eta| < 2.5$ and relative isolation less than 0.2 are rejected if the muon candidate is also labeled as "GlobalMuon" [15].
The Z boson veto here means the presence of a loose electron candidate that gives an invariant mass of $76\,\text{GeV} < M_{ee} < 106\,\text{GeV}$ with the prompt electron candidate. It is defined as an eletron candidate with $p_T > 20\,\text{GeV}$, $|\eta| < 2.5$ (supercluster out of the EE-EB gap) and the relative isolation < 1, fulfilling the identification criteria of WP95, (see Table 4.1).

The jet selection criteria

The calorimeter jets reconstructed with Anti-κ_T algorithm are used in the following. In the simulated samples, the jets are corrected with the relative η and absolute p_T calibrations explained in Section 4.3.2 while in data, additional treatments seem necessary. Based on the studies [174] on the 2010 collision data at $\sqrt{s} = 7\,\text{TeV}$, there are small $\eta-$dependent differences in the comparison between the data and simulation while the absolute p_T dependent energy scale seems to be modeled very well in the simulation.
Therefore a small residual energy correction is applied after the simulation-based Level 2 and Level 3 calibrations which take care of the bulk of the energy response. The jets in data are additionally corrected for the Level 1 offset.
The corrected jet collections both in data and simulation are first cleaned from the electrons where the electron is a candidate fulfilling the electron selection requirements other than the conversion rejection. The corrected jets can participate in the rest of the analysis if they meet the following criteria:

[15] A muon candidate is labeled as "GlobalMuon" if it fulfills the global muon reconstruction conditions as explained in Section 2.2.3.

- $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$
- $f_{em} > 0.01$, $N90_{hits} > 1$ and $f_{HPD} < 0.98$

The expected event yields for a dataset corresponding to $\sim 36\,pb^{-1}$ integrated luminosity are estimated based on the theoretical cross sections, given in Table 3.2. The evaluated number of events for each simulated physics process as well as in data are summarized in Table 5.19. The pre-selection part includes the trigger selection in data together with the request for at least one electron with $p_T > 30\,\text{GeV}$ and the primary vertex selection both in data and simulation.
In addition to the trigger efficiency, the electron selection scale factors calculated by the Tag&Probe method, listed in Table 4.2, are also applied on the number of simulated events. Another correction factor comes from the W boson branching ratio, $Br(W \rightarrow l\nu_l)$, for which the LO approximation, $Br = 0.111$, is implemented in `MadGraph`. Events containing the $W \rightarrow l\nu_l$ decay [16] are reweighted to the measured value of $Br = 0.108$ [3].
The final row in the last two columns shows that the expected total number of events in simulation is different from the data event yield. This discrepancy however is covered by an uncertainty of 11% on the integrated luminosity. Besides, the uncertainties arising from the theoretical predictions (Table3.2) can also be added to the total uncertainty on the simulated event yields. It should be noted that the cross section of the QCD multi-jet processes are not known very well for which an approximation is quoted in Table 3.2. A small uncertainty is also introduced by the uncertainties on the measured scale factors as in Section 4.2.1.
The attempt for the $t\bar{t}$ cross section measurement in the semi-electron final state has lead to some correction factors for the estimation of signal and different background contributions, [25]. These correction factors are applied on the simulated samples used in the following for the b-tagging efficiency estimation. The p_T distribution of the prompt electron candidate and the four leading jets within the selected events in data and simulation after applying the correction factors on the simulated event yields are illustrated in Figure 5.56. Considering the available amount of statistics, the simulation is well describing the data. It should be noted that the overflow entries are not contained in the last bin of histograms.

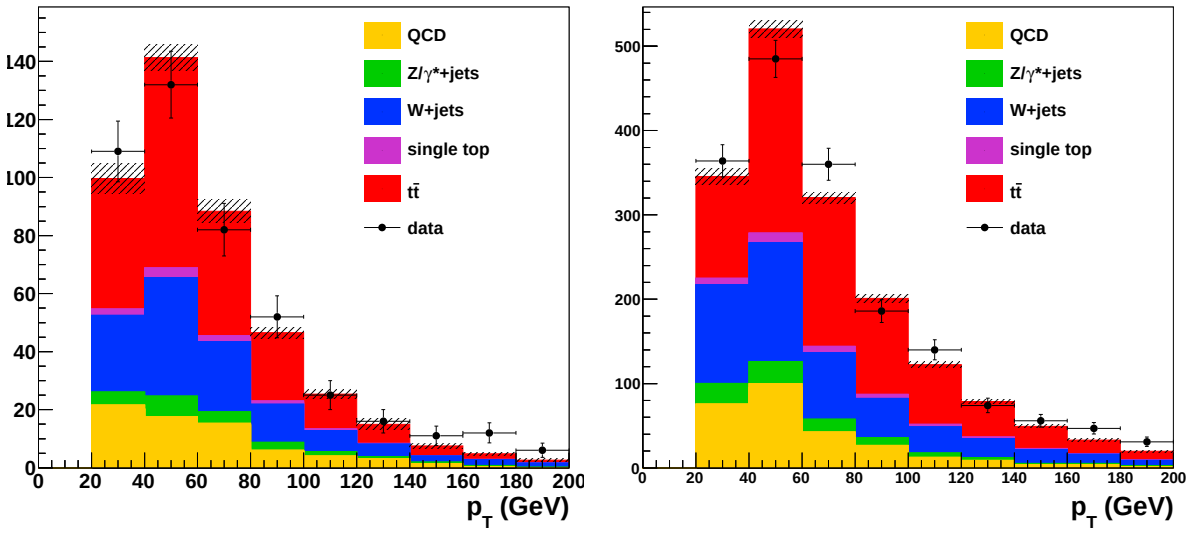
Figure 5.57 shows the number of selected jets before and after the four jet request while in Figure 5.58 the distribution of the TCHE b-tag discriminator as well as the number of jets with $d_{TCHE} > 4$ are presented. The lower jet bins are populated by the QCD multi-jets and the W+jets as expected. Within the given statistics, the data and simulation are in fair agreement over a wide range of the jet multiplicity. A good agreement between the data and simulation is also observed in the TCHE b-discriminator.

The distribution of the $M3$ variable which is a simplistic estimator of the top quark mass is shown in Figure 5.59 where the distribution of the scalar sum of the p_T's of the four leading jets in the event, H_T, is also illustrated. The data and simulation shows reasonable similarity in both distributions. The bin contents are compatible within three standard deviation even for the bins with the largest discrepancies. It needs to

[16] According to the description of the simulated samples in Section 3.4 the W+Jets and Z+jets refer to the leptonically decaying W and Z bosons only.

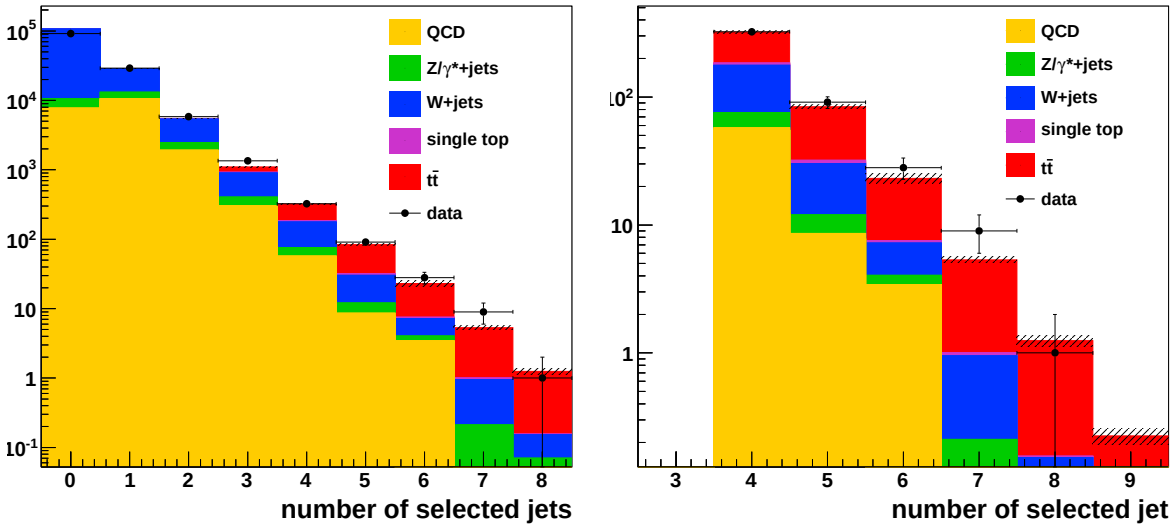
	$t\bar{t}$	single-top	W+jets	Z+jets	QCD	sum MC	data
pre-selection	1530 ± 2.49	268 ± 0.57	143010 ± 101	22306 ± 30	$(607 \pm 3.16) \times 10^4$	$(623 \pm 3.16) \times 10^4$	2424928
exactly one prompt electron	567 ± 1.52	146 ± 0.43	107543 ± 87.5	11546 ± 21.6	62606 ± 296	182407 ± 309	159648
loose μ veto	502 ± 1.43	141 ± 0.42	107528 ± 87.5	11496 ± 21.5	62588 ± 296	182256 ± 309	159469
Z boson veto	495 ± 1.42	139 ± 0.42	107506 ± 87.5	7543 ± 17.4	62539 ± 296	178223 ± 309	155515
conversion rejection	466 ± 1.38	131 ± 0.41	100967 ± 84.8	7009 ± 16.8	22343 ± 191	130916 ± 210	116826
at least 4 qualified jets	186 ± 0.87	9.8 ± 0.1	107 ± 2.8	26 ± 1	75 ± 8	404 ± 9	453

Table 5.19: The cut flow table for the simulated physics processes and the $36\,pb^{-1}$ data collected in 2010 by the CMS experiment. The numbers for the simulated samples are normalized to the same integrated luminosity where the statistical uncertainties are also quoted. The correction factors from the trigger and the electron efficiency are applied on the number of simulated events. For events containing $W \to l\nu_l$ decay, the branching ratio is corrected for the measured value.



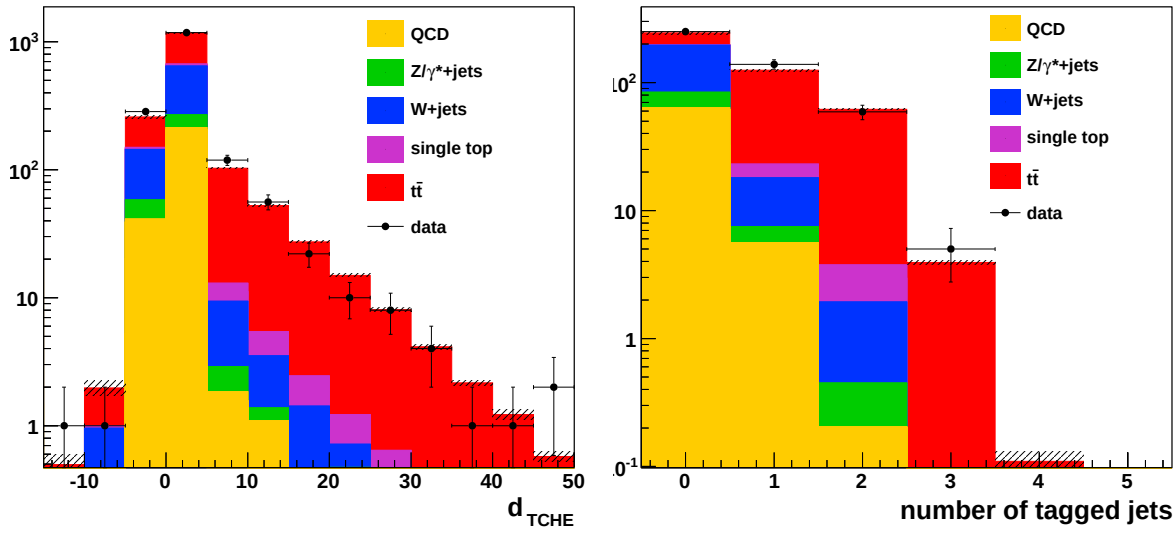
(a) (b)

Figure 5.56: The p_T distribution of the prompt electron, (a), and the four leading jets, (b), for different processes at $L \approx 36\,pb^{-1}$. The simulation is normalized using the correction factors extracted in [25]. The uncertainties on the simulation are purely statistical. The histograms do not contain the overflow entries in the last bin.



(a) (b)

Figure 5.57: The number of selected jets before, (a), and after, (b), the four jet request for different processes at $L \approx 36\,pb^{-1}$. The simulation is normalized using the correction factors extracted in [25]. The uncertainties on the simulation are purly statistical. The histograms do not contain the overflow entries in the last bin.



(a) (b)

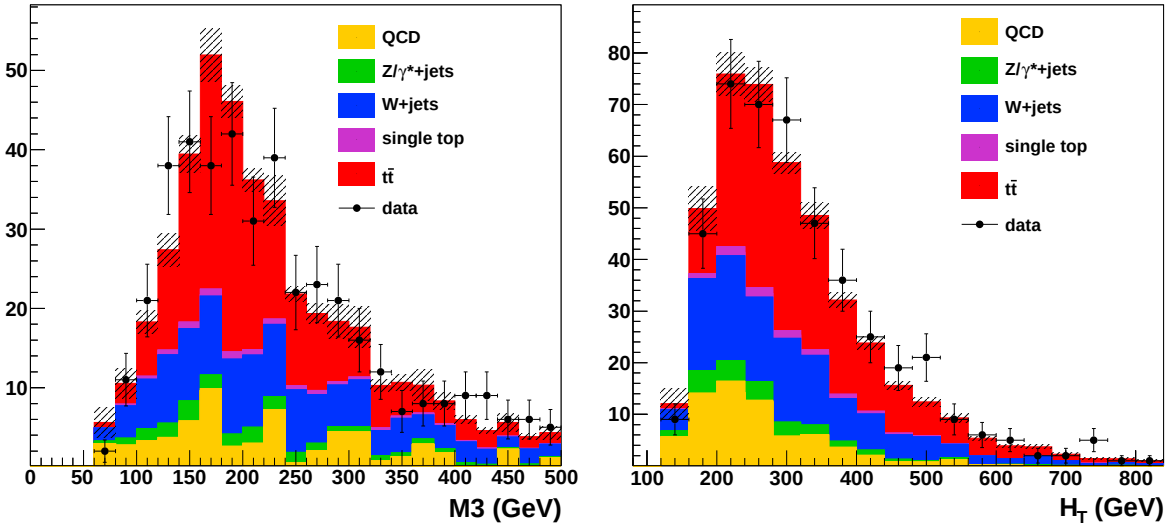
Figure 5.58: The distribution of the track counting high efficiency b-discriminator, (a), together with the number of tagged jets with $d_{TCHE} > 4$ requirement, (b), illustrated for different processes at $L \approx 36\,pb^{-1}$. The simulation is normalized using the correction factors extracted in [25]. The uncertainties on the simulation are purely statistical. The histograms do not contain the overflow entries in the last bin.

be emphasized that the systematic uncertainties on the simulation are not shown on the plots.

5.6.2 Measurement of the ϵ_b in top-like events

The candidates for hadronically decayed top quarks in the selescted events are reconstructed with the three jet combinations minimizing the χ^2 defined in Equation 5.1 where the constraints are re-adjusted for the current set of simulated $t\bar{t}$ events. Figure 5.60 shows the distribution of χ^2_{min} for different simulated processes contributing to the analysis as well as for the data events. The b-candidate jet sample is constructed with the remaining jet out of the four leadings where the control jet sample is made up of the jets contributing in the reconstruction of the hadronically decayed W boson. While the former is aimed for the b-tagging efficiency measurement, the latter is dedicated for the calculation of the scale factor, F. Figure 5.61 illustrates the jet-electron invariant mass distribution for the b-candidate and the control jet sample. The data and simulation agree well regarding the available amount of statistics.
The same boundaries as in Section 5.3 are chosen to divide the b-candidate jet sample into two subsamples, the b-dominated and the b-depleted. In the control jet sample, the same boundaries are used to extract the data driven scale factor, F.

For the sake of a better estimation of the data driven scale factor, the jets in the



(a) (b)

Figure 5.59: The distribution of the M3, (a), together with the scalar sum of the p_T's of the four leading jets, (b), illustrated for different processes at $L \approx 36\,pb^{-1}$. The simulation is normalized using the correction factors extracted in [25]. The uncertainties on the simulation are purly statistical. The histograms do not contain the overflow entries in the last bin.

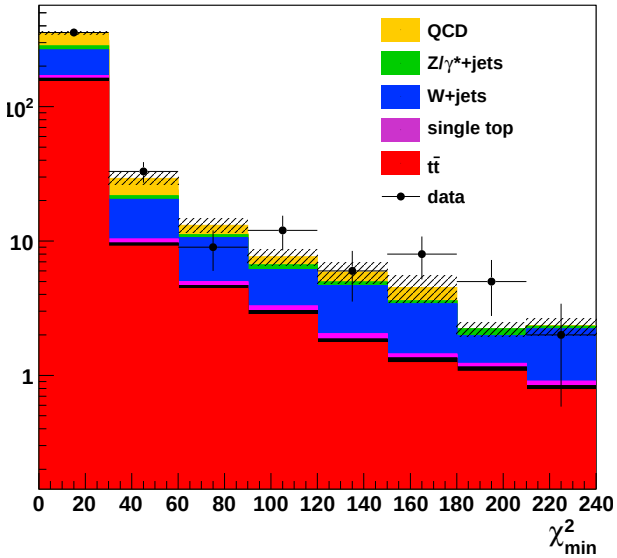
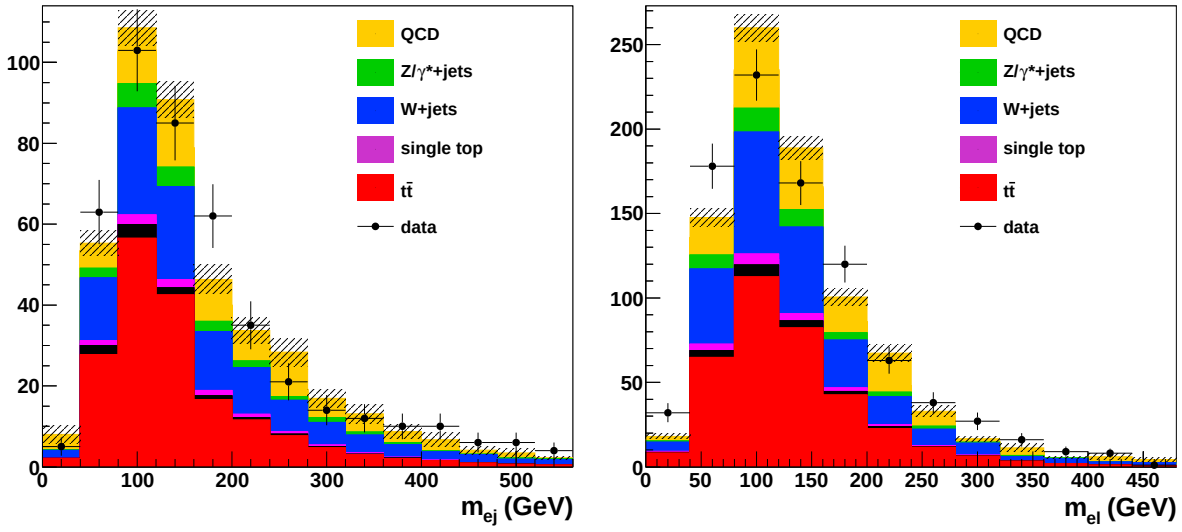


Figure 5.60: The χ^2_{min} distribution for different processes at $L \approx 36\,pb^{-1}$. The simulation is normalized using the correction factors extracted in [25]. The uncertainties on the simulation are purly statistical. The histograms do not contain the overflow entries in the last bin.



(a) (b)

Figure 5.61: The distribution of the jet-electron invariant mass in the b-candidate, (a), and the control, (b), jet sample illustrated for different processes at $L \approx 36\,pb^{-1}$. The simulation is normalized using the correction factors extracted in [25]. The uncertainties on the simulation are purely statistical. The histograms do not contain the overflow entries in the last bin.

control sample are reweighted according to the value of their η and p_T. Dividing the two dimensional $(\eta; p_T)$ distribution of the jets in the b-candidate sample to the one in the control jet sample, the $(\eta; p_T)$ weights are computed. Figure 5.62 illustrates the two dimensional $(\eta; p_T)$ distributions for the data in the b-candidate and the control jet sample. The amount of statistics is too low to make a robust comment on these distributions. They are expected to resemble the scatter plots in Figure 5.39 for more accumulated data.

In Table 5.20, the scale factor extracted from simulated events using the control sam-

F_{sim}	$\widehat{F_{sim}}$	$\frac{(\widehat{F_{sim}} - F_{sim})}{F_{sim}}$	$\widehat{F_{data}}$	$\frac{(\widehat{F_{data}} - \widehat{F_{sim}})}{\widehat{F_{sim}}}$
1.76 ± 0.05	1.73 ± 0.04	-0.0198 ± 0.036	1.99 ± 0.19	0.149 ± 0.113

Table 5.20: The scale factor computed using the generator level information, F_{sim}, compared to the data driven $\widehat{F_{sim}}$ in simulation normalized to an integrated luminosity of $L \approx 36\,pb^{-1}$. The normalization corresponds the accumulated data in 2010, for which the scale factor, $\widehat{F_{data}}$, is extracted from the control jet sample.

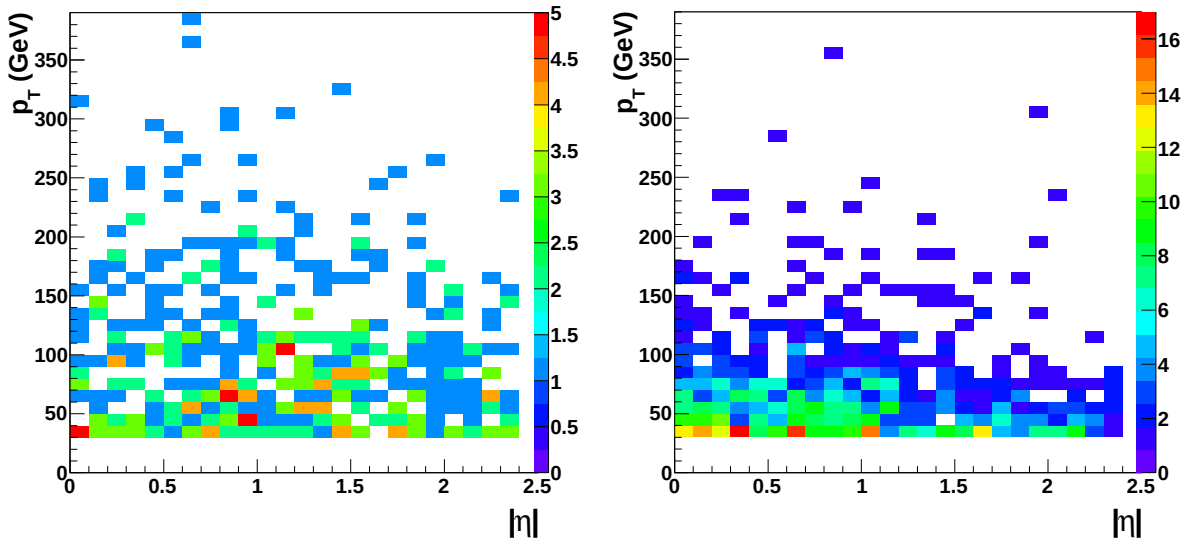


Figure 5.62: The $(\eta; p_T)$ distribution of the jets in the b-candidate, (a), and the control, (b), jet sample illustrated for the $L \approx 36\, pb^{-1}$ of collision data.

ple, $\widehat{F_{sim}}$, is compared with the true F_{sim} from non-b-quark jets in the b-candidate jet sample. The table includes the value for the scale factor calculated in data, $\widehat{F_{data}}$. A large uncertainty exists for the data results due to the small amount of statistics.
The subtraction of the scaled b-tag distribution in the b-depleted jet sample, $F.\Delta^R_{all}$, from the same distribution in the b-dominated jet sample, Δ^L_{all}, as described by Equation 5.7 leads to non-physical values for the b-tagging efficiency especially for the low d_{TCHE} values.
The reason basically is the overestimation of the non-b-quark jet content in the b-dominated jet sample that results in negative entries for the subtracted b-tag histogram. The overestimation is cured by reweighting the jets in the b-depleted sample to match the p_T distribution of the b-dominated jet sample (see Section 5.3.2). Figure 5.63 shows the subtracted b-discriminator distribution in data before and after the p_T reweighting. The bins are adjusted to represent the b-tagging working points. This distribution, $\Delta^L_{all} - F.\Delta^R_{all} = \hat{\Delta}_b$, is an estimator for the distribution of the b-tagging discriminator for true b-quark jets.

The b-tagging efficiencies for the loose, medium, and tight working points, measured after the p_T reweighting are summarized in Table 5.21 where the corresponding estimation in the simulated samples together with the expected ϵ_b^{sim} is also presented.

The huge statistical uncertainties on the measured values are expected as explained by the study carried out in Section 5.4. The method is statistically unstable for the low integrated luminosities while more robust results are expected with an increased amount of the accumulated data.
The systematic uncertainties are much smaller than the statistical uncertainty. Hence, they are not mentioned for this measurement on the 2010 data.

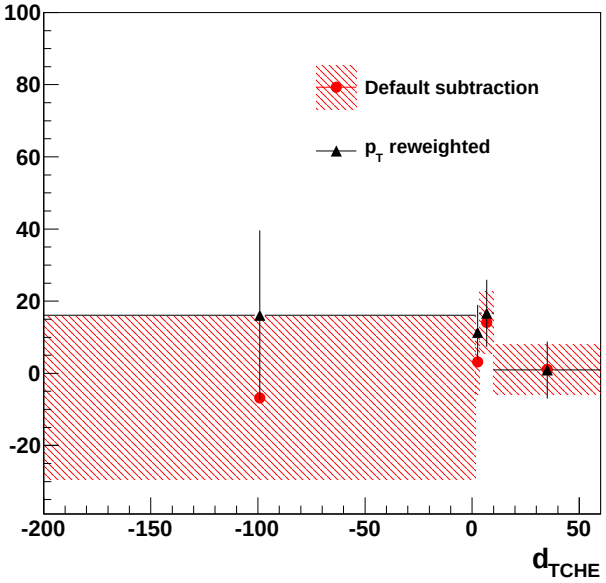


Figure 5.63: The b-discriminator distribution for the subtracted jet sample in the data set of $L \approx 36\,pb^{-1}$ integrated luminosity before and after the p_T reweighting.

working point		ϵ_b^{sim}	$\widehat{\epsilon_b^{sim}}$	$\frac{\left(\widehat{\epsilon_b^{sim}}-\epsilon_b^{sim}\right)}{\epsilon_b^{sim}}$	$\widehat{\epsilon_b^{data}}$	$\frac{\left(\widehat{\epsilon_b^{data}}-\widehat{\epsilon_b^{sim}}\right)}{\widehat{\epsilon_b^{sim}}}$
loose	$d = 1.7$	0.79 ± 0.07	0.69 ± 0.073	-0.126 ± 0.12	0.73 ± 0.36	0.058 ± 0.53
medium	$d = 3.3$	0.64 ± 0.058	0.55 ± 0.06	-0.141 ± 0.122	0.42 ± 0.26	-0.236 ± 0.48
tight	$d = 10.2$	0.29 ± 0.031	0.24 ± 0.033	-0.172 ± 0.144	0.20 ± 0.13	-0.167 ± 0.553

Table 5.21: The efficiency of the TCHE b-tagging algorithm computed using the generator level information, ϵ_b^{sim}, compared to the data driven estimator $\widehat{\epsilon_b^{sim}}$ in simulation normalized to an integrated luminosity of $L \approx 36\,pb^{-1}$. The normalization corresponds the accumulated data in 2010, for which the b-tagging efficiency, $\widehat{\epsilon_b^{data}}$, is measured. The results are presented for the loose, medium and the tight b-tagging working points.

Chapter 6

Conclusion and towards a $t\bar{t}$ cross section measurement

The top quark characterized as the heaviest quark in the Standard Model of particle physics has been discovered in 1995 by the DØ and CDF experiments at the Tevatron collider in Fermilab. The experiments at the Tevatron have succeeded to measure the top quark cross section and its mass with a very good precision. The branching ratio to different possible final states has also been investigated resulting in the probability of $\sim 99\%$ for the $t \to bW$ decay mode.
At the beginning of the LHC era, the top quark has been observed by the CMS and ATLAS experiments within the $\sim 36\,pb^{-1}$ of accumulated data at 7 TeV center of mass energy and its mass and cross section have been re-measured where greater precisions are expected for more integrated luminosities.
The measured cross section at CMS for the semi-lepton final state without the b-jet identification requirement has been [25]

$$\sigma_{t\bar{t}} = 173^{+39}_{-32}\,(\text{stat} + \text{syst}) \pm 19\,(\text{lumi})\ pb, \tag{6.1}$$

where the value of

$$\sigma_{t\bar{t}} = 150 \pm 9\,(\text{stat.}) \pm 17\,(\text{syst.}) \pm 6\,(\text{lumi})\ pb, \tag{6.2}$$

is obtained for the measurement with the use of the b-jet identification [163]. The semi-lepton channels contain the lepton identification to search for the top quark event candidates. Hence the efficiency of such identifications are accounted for in the reported cross section values.

The experimental signature of the top quark which is mostly produced in pair at the LHC contain most of the physics objects reconstructed by the CMS detector. Hence regarding the high $t\bar{t}$ production rate at the LHC, the $t\bar{t}$ events are quiet useful for calibration and commissioning purposes.
In particular, the rich source of b-jets provided by the top quark events can be exploited for estimating the efficiency of the b-jet identification algorithms in a data

driven way. The importance of such measurement becomes clearer by considering the use of the b-jet identification not only for the physics analyses in the Standard Model but also in the searches for new physics.
In Section 6.1, a brief review is given for the measurement of the electron identification and isolation scale factors using the Tag&Probe method. Section 6.2 is an overview of the method developed in this thesis for the b-tagging efficiency estimation in the semi-electron $t\bar{t}$ final state. The result of the first look at the data is presented, the combination with the semi-μ final state is studied and the performance of the method at higher integrated luminosities is discussed.
Possible extensions to the method towards a $t\bar{t}$ cross section measurement are investigated in Section 6.3.

6.1 Electron isolation and identification scale factors

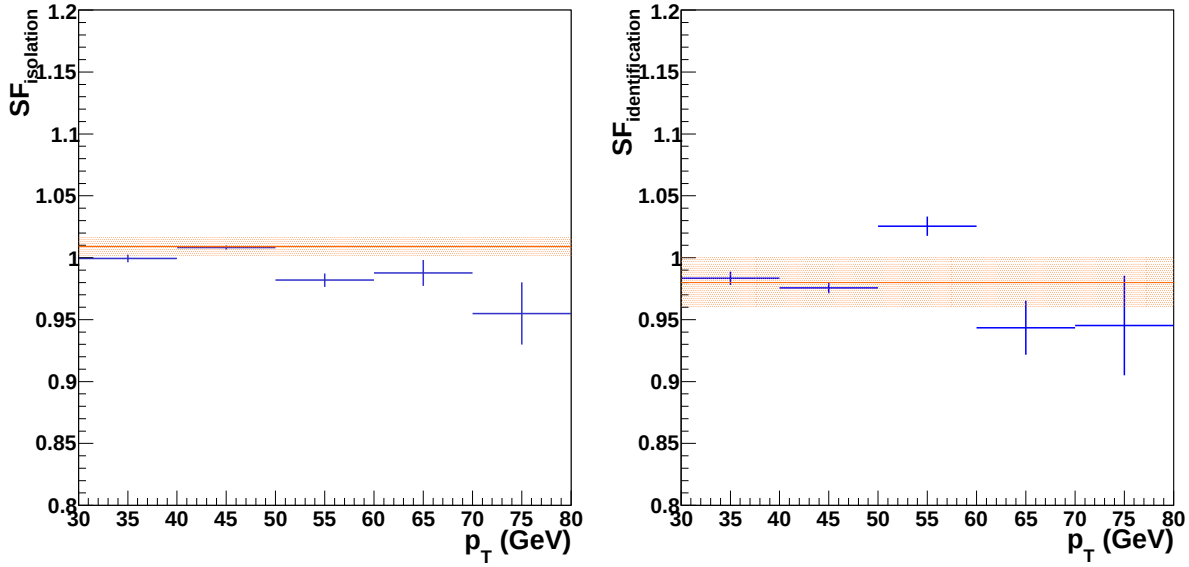
The electron candidates are reconstructed by matching the supercluster of energy deposits in the ECAL to a track in the tracking system. It has been briefly reviewed in Chapter 2 how the energy spread in the ϕ direction in ECAL resulting from the bremsstrahlung energy losses is accounted for in the reconstruction of the superclusters. In Chapter 4, the dedicated tracking algorithm for the electron track reconstruction which considers the multiple scattering in tracker material as well as the bremsstrahlung effect has been presented where it has been explained how the supercluster and track matching is optimally performed. Special treatments are also carried out to estimate the electron momentum resulting in a well reconstructed electron candidate suitable for the physics analyses.
For the $t\bar{t}$ cross section measurement in the semi-electron final state, the presence of an isolated high p_T electron meeting the qualification criteria described in Chapter 4, is a crucial requirement.
The electron selection has an efficiency which needs to be accounted for in the final computation of the cross section. The efficiency which is factorized into trigger [1], reconstruction, acceptance, isolation and identification components, can be deduced from the generator level information but a more consistent approach is the data driven efficiency measurement using the Tag&Probe method in $Z \rightarrow ee$ events described in Section 4.2.
Considering the possible differences in the electron properties between the $Z \rightarrow ee$ and the $t\bar{t}$ events, the scale factors, SF $= \frac{\epsilon_e^{data}}{\epsilon_e^{MC}}$, are obtained from the Tag&Probe method in $Z \rightarrow ee$ events and applied on the simulation-driven electron efficiency in the $t\bar{t}$ events.
For the $t\bar{t}$ cross section measurement in 2010, various cross checks mainly different from the background subtraction point of view are performed to evaluate the electron selection scale factors [164].
As detailed in Section 4.2.2, making a side band subtraction under the Z boson mass

[1] Needed where an electron candidate is looked for at High Level Trigger selection.

peak which is an alternative to reject the background electron pairs, has resulted in

$$SF_{id} = 0.98 \pm 0.02\,(\text{syst.} + \text{stat.}),$$
$$SF_{iso} = 1.009 \pm 0.007\,(\text{syst.} + \text{stat.}). \tag{6.3}$$

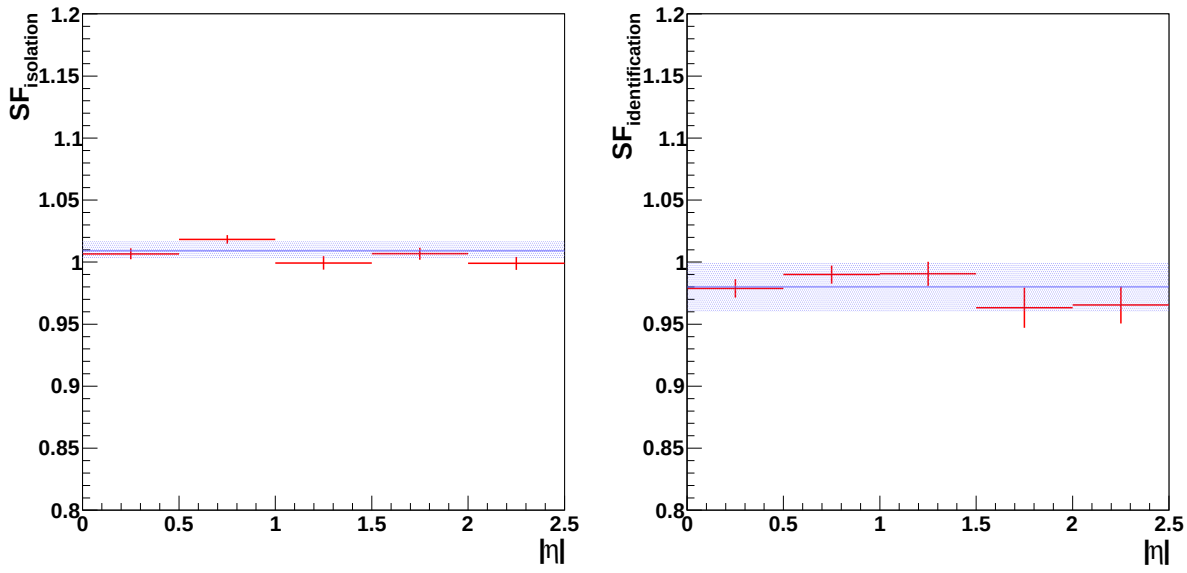
The evaluated scale factors give already a good estimate of the electron efficiency in the $t\bar{t}$ events. However due to the dissimilar characteristic of the $t\bar{t}$ and $Z \rightarrow ee$ events, the scale factors may be unequal between them. This can be covered by the additional systematic uncertainty obtained from the electron efficiency difference between the $t\bar{t}$ and $Z \rightarrow ee$ simulated samples. This inequality is found to be negligible for the identification efficiency while for the isolation efficiency, a difference of $\Delta\epsilon_{iso} \approx 6\%$ is obsereved. This information is used in [25, 163].
In Section 4.2.2 it was also shown that the electron isolation and identification efficiencies change as a function of p_T and $|\eta|$ both in data and simulation. As far as the changes in simulation follow those in data, the scale factors are expected to be flat with respect to the kinematic variables, p_T and $|\eta|$, while a difference in this behavior between the data and simulation would lead to p_T ($|\eta|$) dependent scale factors.



(a) (b)

Figure 6.1: The electron isolation, (a), and identification, (b), scale factors as a function of p_T. The scale factors are obtained within $36\,pb^{-1}$ of the data collected in 2010.

Figure 6.1 illustrates the isolation and identification scale factors as a function of the p_T of the electron where the scale factors in different electron pseudo-rapidities are shown in Figure 6.2. The averaged scale factors together with their uncertainties as stated in Equation 6.3, are also indicated on each plot. Apart from the statistical fluctuations, the p_T ($|\eta|$) dependent scale factors are in a good agreement with the averaged scale factors within the given uncertainties. This observation can be validated with more accomulated data.



(a) (b)

Figure 6.2: The electron isolation, (a), and identification, (b), scale factors as a function of $|\eta|$. The scale factors are obtained within $36\,pb^{-1}$ of the data collected in 2010.

6.2 Estimation of the b-tagging efficiency

The reconstruction of jets with the Anti-κ_T algorithm is introduced in Chapter 4 where the different steps of the jet energy correction are reviewed. For the jets participating in the analysis presented in this thesis, the absolute (p_T) and the relative (η) energy calibration was applied.

An event sample enriched by top-quark events was constructed by looking for an isolated high p_T electron together with four well defined jets. The jets fulfilled the energy requirement as well as some identification criteria as described in the first section of Chapter 5. For $100\,pb^{-1}$ integrated luminosity and at 7 TeV center of mass energy, the expected event yield for the semi-electron final state of $t\bar{t}$ is determined to be 373 while a similar number of events survived from the background processes. The QCD multi-jets, W(Z)+jets and single top as well as the other $t\bar{t}$ final states are the considered backgrounds. The event yield for backgrounds was shown to be dominated by the W+jets process.

Within the selected event sample, the hadronically decayed top-quark was reconstructed using a minimum χ^2 requirement constrained by the mass of the top-quark and the W boson as explained in Section 5.2. The remaining jets from the four leading jets in the events form a b-candidate jet sample with a b-quark jet purity of $\sim 30\%$. Due to the existing kinematic correlation between the electron and the leptonic b-jet in the $t\bar{t}$ event, the special shape of the electron-jet invariant mass was exploited to divide the b-candidate jet sample into the b-enriched and the b-depleted subsample with b-purities of 39% and 11% respectively.

The b-dominated jet sample was further purified by subtracting the non-b-quark jets

as detailed in Section 5.3.1. The shape of the non-b-quark jets was estimated in the b-depleted jet sample and scaled to match the expected non-b-quark jet distribution in the b-dominated jet sample. The efficiency of the Track Counting High Efficiency b-jet identification algorithm was estimated in the subtracted sample where the method was validated for other b-tagging algorithms introduced in Section 4.4. A discrepancy at low positive b-discriminator values was observed and was thereafter resolved by reweighting the jets in the b-depleted sample according to their p_T.
The method became fully data driven by extracting the scale factor from a control jet sample that was constructed by the jets associated to the hadronically decayed W boson. It is explained in Section 5.3.5 how the jets in the control sample were reweighted in order to account for the kinematic dissimilarities between the control and the b-candidate jet samples.
The study of the statistical properties of the estimators in Section 5.4 showed that the method needs more accumulated data than $100\,pb^{-1}$ to be stable.
The influence of different systematic sources on the method was studied in Section 5.5 where the conservative choices were made in particular for the intrinsic bias on the method and the fluctuations due to different event generators. The estimated b-tagging efficiency for the loose, medium and tight working points of the track counting high efficiency b-tag algorithm together with the statistical and systematic uncertainties were found to be

$$\begin{aligned}
\widehat{\epsilon_b}(\text{loose}) &= 0.698 \pm 0.036\,(\text{stat.}) \pm 0.033\,(\text{syst.}),\\
\widehat{\epsilon_b}(\text{medium}) &= 0.471 \pm 0.034\,(\text{stat.}) \pm 0.025\,(\text{syst.}),\\
\widehat{\epsilon_b}(\text{tight}) &= 0.243 \pm 0.029\,(\text{stat.}) \pm 0.008\,(\text{syst.}),
\end{aligned} \tag{6.4}$$

at $100\,pb^{-1}$ integrated luminosity. The uncertainties in all working points are dominated by statistics. The systematic uncertainties from the intrinsic bias on the method and from the event generators were removed from the list because what was quoted for this uncertainties were the statistical uncertainty on the obtained values and this can be overcome with larger simulated samples.
Within the given statistics, the estimated b-tagging efficiencies showed no bias with respect to the expected values, presented in Table 5.8.

6.2.1 The measurement with the 2010 data collected by CMS

To run the method over the whole dataset collected in 2010 equivalent to $\sim 36\,pb^{-1}$ of integrated luminosity, events with one isolated high p_T electrons and four energetic jets were selected. Beside the Level 2 and Level 3 energy corrections, the jets were calibrated with respect to the pile-up and possible electronic noise. Different distributions were checked for the data-simulation comparison where a fair similarity were observed. The key distributions including the minimum χ^2 and electron-jet invariant mass in the b-candidate and the control jet samples in simulation showed a good agreement with data within the available amount of the statistics.
Following the fully data driven approach, the b-tagging efficiency was measured at the loose, medium and the tight working points of the track counting high efficiency

b-tagging algorithm,

$$\begin{aligned} \widehat{\epsilon_b}(\text{loose}) &= 0.73 \pm 0.36\,(\text{stat.}), \\ \widehat{\epsilon_b}(\text{medium}) &= 0.42 \pm 0.26\,(\text{stat.}), \\ \widehat{\epsilon_b}(\text{tight}) &= 0.20 \pm 0.13\,(\text{stat.}). \end{aligned} \tag{6.5}$$

Large statistical uncertainties are observed on the measured efficiencies because of the limited size of the data sample and systematic uncertainties are negligible at this point.

6.2.2 The potential of the method for higher integrated luminosities

The LHC machine is currently working with a great performance. The increment of the instantaneous luminosity will afford a considerable amount of accumulated data, giving the prospect of better statistical uncertainties in physics analyses.
Figure 6.3 illustrates the evolution of the statistical uncertainty on the b-tagging efficiency at the medium working point of the track counting high efficiency b-jet identification algorithm with respect to the integrated luminosity at 7 TeV center of mass energy. To cover the wide range of integrated luminosities from $36\,pb^{-1}$ to $10\,fb^{-1}$ which sounds feasible in the following years of the LHC operation, the horizontal axis is shown on the logarithmic scale.
To obtain the statistical uncertainties at other integrated luminosities, the uncertainty

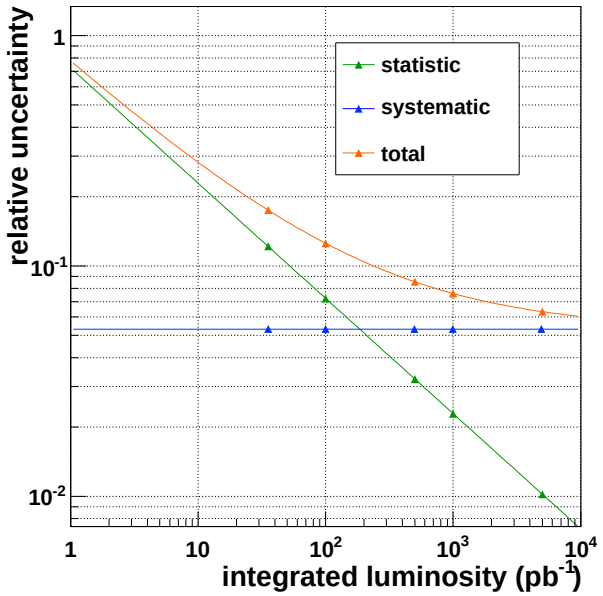


Figure 6.3: The relative uncertainty on the $\widehat{\epsilon_b}$ as a function of integrated luminosity. The total uncertainty together with the statistical and systematic uncertainties are presented.

at $L = 100\,pb^{-1}$ is rescaled,

$$\delta_{L\,pb^{-1}} = \delta_{100\,pb^{-1}}\sqrt{\frac{L}{100}},$$

where $\delta_{L\,pb^{-1}}$ is the statistical uncertainty at the desired integrated luminosity, L, stated in pb^{-1} unit. It can be seen that the falling statistical uncertainty will reach the limit of the systematics uncertainty at $L = 200\,pb^{-1}$. The systematic uncertainty has contributions from all sources investigated in Section 5.5 exccept from the intrinsic bias on the method and from different event generators. The reason is that for these two sources, the statistical uncertainties were quoted which will disappear with larger simulated samples. The total systematic uncertainty is assumed to remain constant which is a conservative assumption. The uncertainty due to the jet energy scale will be decreased with the better reconstruction algorithms and the more robust understanding about the detector. The parameters for modeling the pp collisions will also be tuned more precisely. Moreover, the background cross sections will be measured more accurately. The total relative uncertainty is shown in the same figure which is ultimately limited by the systematic uncertainty for $L > 200\,pb^{-1}$.

6.2.3 Combination with the semi-μ final state

The b-tagging efficiency estimation which was performed in the semi-electron final state can be combined with the results in the semi-muon final state of $t\bar{t}$ at any integrated luminosity. While the systematic uncertainties are about the same, combining the two channels will result in a better statistical uncertainty which is useful for low integrated luminosities.

The method to estimate the b-tagging efficiency in the semi-muon final state has been developed for $1\,fb^{-1}$ at 10 TeV center of mass energy [186]. Assuming the same event selection efficiency at $\sqrt{s} = 7$ TeV, the event yields in [186] are recalculated for the signal and background cross sections at 7 TeV center of mass energy and are scaled to $100\,pb^{-1}$. Table 6.1 summarizes the semi-muon channel event yields for $100\,pb^{-1}$ integrated luminosity at 7 TeV compared with 10 TeV center of mass energy. Although the selection efficiencies are assumed to be the same in the two center-of-mass energy scenarios, it can be seen that the signal over background ratio is slightly better at the higher center of mass energy, $\sqrt{s} = 10$ TeV. This is due to the fact that the $t\bar{t}$ cross section increases with the center of mass energy faster than the cross section of background processes (in particular W+jets) [2]. Because of the slightly larger background contamination, a slightly different estimation of the b-tagging efficiency is also expected at $\sqrt{s} = 7$ TeV comparing to the results at $\sqrt{s} = 10$ TeV in [186].

However, similar to what was discussed for the semi-electron channel in Section 5.5, the uncertainty due to the background contamination on the $\widehat{\epsilon_b}$ at the medium working point is expected to be small for the semi-muon channel as well [3].

[2] The $t\bar{t}$ and W+jets cross sections at $\sqrt{s} = 10$ TeV are expected to be $414\,pb^{-1}$ and $45.6 \times 10^3\,pb^{-1}$, respectively [186]. These values can be compared to what is presented in Table 3.2 for the two processes: $\sigma_{t\bar{t}} = 157\,pb^{-1}$ and $\sigma_{W+jets} = 31314\,pb^{-1}$.

[3] About 1.9% relative uncertainty is reported in [186] from the background contributions at the medium working point of the track counting high efficiency algorithm.

event yield @ $100\,\mathrm{pb}^{-1}$	$t\bar{t}$ (semi-μ)	$t\bar{t}$ (others)	single-top	W+jets	Z+jets
$\sqrt{s} = 10\,\mathrm{TeV}$	1025	232.3	66.3	472.2	73.6
$\sqrt{s} = 7\,\mathrm{TeV}$	389.9	88.4	30.2	324.3	53.4

Table 6.1: The estimated semi-muon channel event yield for $100\,pb^{-1}$ at 10 TeV and 7 TeV center of mass energy. The event selection efficiencies are assumed to be the same as in two different energy scales. The expected event yields at 7 TeV center of mass energy can be compared with those for the semi-electron final state in Table 5.3.

Therefore, in the semi-mu channel at 7 TeV center of mass energy the same efficiency value, $\widehat{\epsilon_b} = 0.499$, as at $\sqrt{s} = 10$ TeV can be taken for the medium working point of the track counting high efficiency, TCHE, b-tagging algorithm. Although this value in [186] is computed not only at higher center of mass energy but also at higher integrated luminosity, $L = 1\,fb^{-1}$, it can be considered as valid for the integrated luminosity of $L = 100\,pb^{-1}$ since technically the change in the integrated luminosity does not result in a different efficiency value as far as the center of mass energy is kept the same.
The amount of statistics in each channel is important in combining the semi-electron and semi-muon results. From a comparison between Table 6.1 (semi-muon channel) and Table 5.3 (semi-electron analysis), it can be deduced that the event yields at 7 TeV center of mass energy are about the same for the signal and background processes in both channels. Hence, the final results in both channels would have about the same sensitivity. As a result, the uncertainty on the combined estimation is expected to be smaller than the single estimations by a factor of $1/\sqrt{2}$. Such approximation is valid under the assumption of a negligible overlap between the semi-muon and the semi-electron selected samples and leads to an absolute statistical uncertainty of

$$\delta^{\mathrm{m.w.p}}_{\mathrm{combined}} \approx \frac{1}{\sqrt{2}}\delta^{\mathrm{m.w.p}}_{\mathrm{e+jets}} = \frac{0.034}{\sqrt{2}} = 0.024$$

at an integrated luminosity of $100\,pb^{-1}$. The notation m.w.p stands for the medium working point of the TCHE b-tagging algorithm. Figure 6.4 illustrates the evolution of the combined statistical uncertainty at higher integrated luminosities. The systematic uncertainty which is conservatively taken to be constant is the same as for the semi-electron channel. The statistical uncertainty seems to reach the limit of systematic uncertainty at $L = 100\,pb^{-1}$ which is attained earlier than the limit for the single electron+jets analysis shown in Figure 6.3.

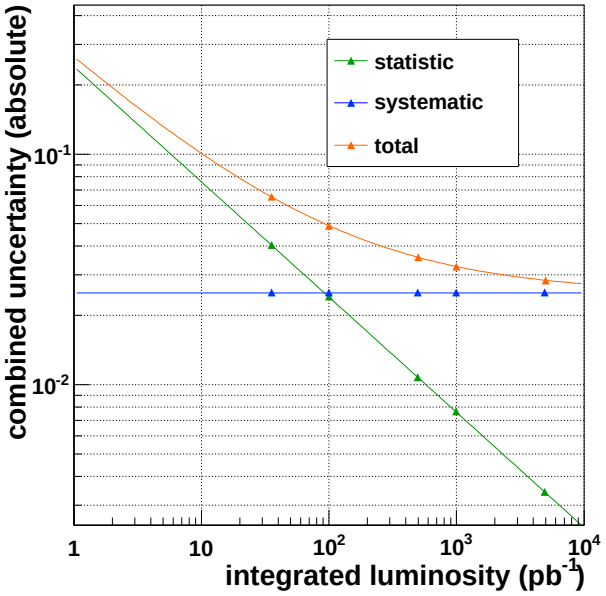


Figure 6.4: The combined uncertainty on the estimated b-tagging efficiency at the medium working point of TCHE algorithm as a function of the integrated luminosity. The total uncertainty together with the statistical and systematic uncertainties are presented.

6.3 The potential extensions from the ϵ_b estimation to a $\sigma_{t\bar{t}}$ measurement

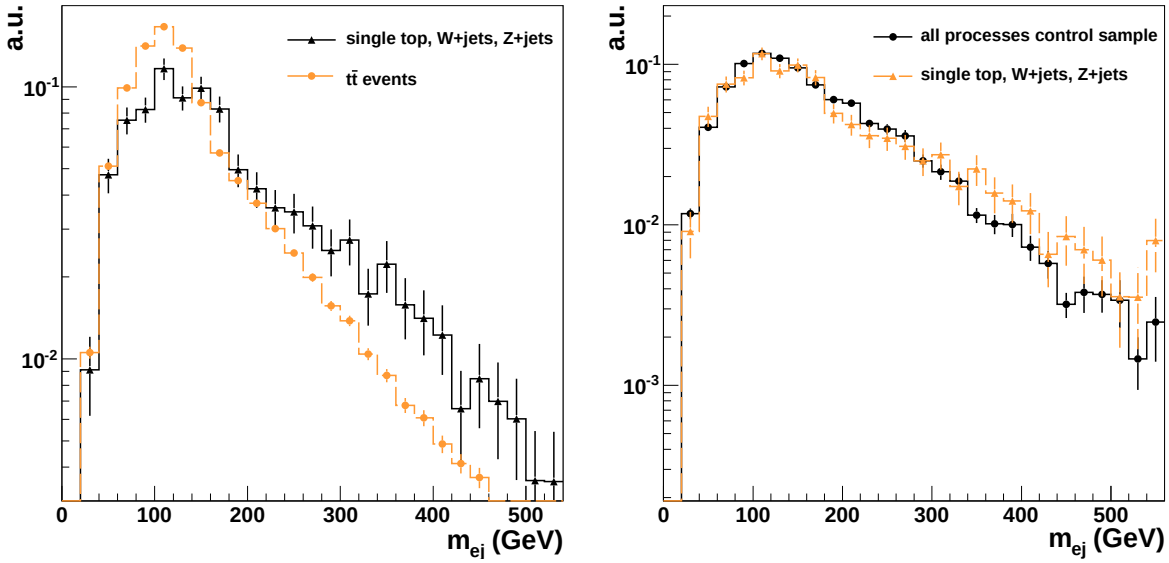
Concerning the desire for a more precise $t\bar{t}$ cross section measurement in terms of the uncertainties and the interest to perform the measurement as independent as possible from the simulation, the method presented in this thesis for the b-tagging efficiency measurement can be extended in different ways.
Introducing two distinct jet samples based on the distinguishing characteristics of the jets in the $t\bar{t}$ event, the method can provide a data driven template for the background processes.
Moreover, the $t\bar{t}$ cross section can be measured with the b-tagging efficiency, giving the prospect of a reduced total uncertainty on the measured cross section.

6.3.1 A data-driven template for the background contributions

In the $\sigma_{t\bar{t}}$ measurement performed in [25], the invariant mass of the 3-jets vectorial sum with the highest p_T, the so-called $M3$ variable, plays the role of the template discriminating between the $t\bar{t}$ and other processes containing the vector bosons. The $t\bar{t}$ cross section has been measured by a simultaneous $M3$ template fit for signal and backgrounds [4].

[4] The $M3$ variable is used simultaneously in a template fit with the missing transverse energy to estimate the contribution of QCD. multi-jets.

For the analysis presented in this thesis, the m_{ej} variable in the b-candidate jet sample, introduced in Section 5.3, has a distinct shape for the $t\bar{t}$ events with respect to the other physics processes. Figure 6.5 (a) compares the shape of m_{ej} for the jets from



(a) (b)

Figure 6.5: The jet-electron invariant mass distribution for the $t\bar{t}$ and the background processes in the b-candidate jet sample, (a), and for the background processes in the b-candidate jet sample comparing to all processes in the control jet sample, (b). The background sample contains W+jets, Z+jets and single top processes.

the $t\bar{t}$ events and the jets from other processes in the b-candidate jet sample where the non-$t\bar{t}$ processes include the Z+jets, W+jets and single top. A broader shape is observed for the backgrounds.

Although a simultaneous template fit on the m_{ej} distribution for the signal and background process where the templates are taken from simulation would result in the estimation of the number of $t\bar{t}$ events, one can take the background template from the data itself. Since the control jet sample is dominated by non-b-quark jets, the m_{el} distribution [5] in this jet sample can provide the data driven template for the background processes in the b-candidate jet sample.

Figure 6.5 (b) illustrates the shapes of the electron-jet invariant mass for the jets from background processes in the b-candidate jet sample and the jets in the control sample. The shapes are similar enough for the m_{ej} template of the background processes in the b-candidate jet sample to be approximatied by the m_{el} shape in the control jet sample. To investigate the performance of such a template fit, 500 pseudo-experiments corresponding to $100\,pb^{-1}$ of integrated luminosity are made for which a simultaneous

[5] The notation l in m_{el} stands for the "light" jet candidates. Here the jet flavors other than b are considered as light and this notation is to emphasize that the control jet sample is dominated by non-b-quark jets.

template fit is performed on the signal and background contributions. While the template for the signal is taken using the generator level information, for the background the fit has been done once with the data driven template and once more with the template derived from simulation.

Figure 6.6 shows the N^c_{bkg}/N^b_{bkg} computed within the pseudo-experiments where $N^{c(b)}_{bkg}$

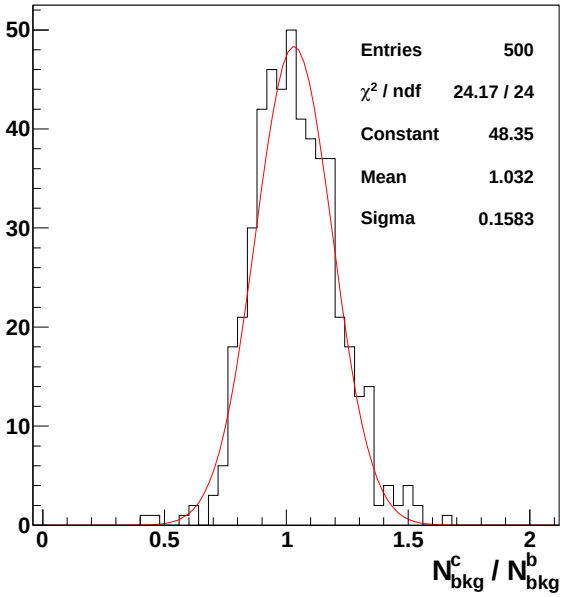


Figure 6.6: The ratio between the estimated number of background events from the data driven template and the template made using the generator level information. The data driven template is obtained from the jets in the control jet sample.

is the estimated number of background events for which the template is taken from the control (b-candidate) jet sample. The histogram is fitted with a Gaussian and a mean value of $N^c_{bkg}/N^b_{bkg} = 1.032$ is obtained. It confirms that the data driven template is indeed a good approximate for the m_{ej} shape in the background processes.

A preliminary estimate for the $t\bar{t}$ cross section is obtained by dividing the evaluated number of $t\bar{t}$ events from the template fit, $N_{t\bar{t}}$, to the integrated luminosity and the total selection efficiency,

$$\sigma_{t\bar{t}} = \frac{N_{t\bar{t}}}{L \cdot \epsilon_{selection}}. \tag{6.6}$$

Within the same pseudo-experiments, a simultaneous fit is performed for the signal ($t\bar{t}$) and background processes where the signal template is taken from simulation and the template for background processes is obtained from the control jet sample as explained. This leads to an estimation of $N_{t\bar{t}}$ and N^c_{bkg} per pseudo-experiment. Regarding the event selection results in Tabel 5.3, the total $t\bar{t}$ selection efficiency is $\sim 2.7\%$. Therefore the $t\bar{t}$ cross section in each pseudo-experiment can be evaluated following Equation 6.6. The mean value of the Gaussian which is fitted to computed cross sections is taken as the estimated $\sigma_{t\bar{t}}$ at $100\,pb^{-1}$ integrated luminosity where the width is considered as the statistical uncertainty on the cross section estimation,

$$\widehat{\sigma}_{t\bar{t}} = 145.3 \pm 51.4\ pb.$$

6.3.2 The prospect for the simultaneous $(\sigma_{\mathbf{t\bar{t}}}; \epsilon_{\mathbf{b}})$ measurement

As discussed in Section 5.1.4, asking for the presence of at least one b-jet candidate is a powerful requirement to make a purer sample of signal events, hence it gives the opportunity for a better cross section measurement. The measured value is however accompanied by a larger systematic uncertainty due to the b-tagging efficiency. Measurements which simultaneously fit for the b-tagging efficiency and the cross section are performed to overcome this additional systematic uncertainty.
One way to reduce such uncertainty is to benefit from the possible correlation between b-tagging efficiency and the cross section measurements when they are performed within the same event sample. The method to estimate the b-tagging efficiency in the top-quark events gives the opportunity for the simultaneous $\sigma_{t\bar{t}}; \epsilon_b$ measurement.
To estimate the $t\bar{t}$ cross section after the b-tagging, the jets in the b-candidate jet sample are asked to fulfill the desired b-tagging criterion. Figure 6.7 shows the evolution of the m_{ej} distribution in the b-candidate jet sample before b-tagging and after the loose, medium and tight b-tagging requests. A smaller background contribution is obtained by tightening the b-tag cut, as expected.

The jets in the control sample are supposed to provide the expected electron-jet invariant mass shape of non-b-quark jets for the computation of the data driven scale factor, F. Hence, these jets are not asked to fullfil any b-tagging requirement. On the other hand, the template for the background processes in the b-candidate jet sample which is b-tagged, cannot be estimated from a control sample which is not required for any b-tag selection. Thus to estimate the $\sigma_{t\bar{t}}$ with the use of b-tagging, the data driven background template introduced in Section 6.3.1 is substituted with the one from simulation.
The $t\bar{t}$ cross section can be deduced as follows from the number of events obtained by the b-tagged template fit, $N_{t\bar{t}}^{tagged}$:

$$\sigma_{t\bar{t}}^{b} = \frac{N_{t\bar{t}}^{tagged}}{L \cdot \epsilon_{selection} \cdot \epsilon_{b-tag}}. \tag{6.7}$$

where ϵ_{b-tag} is the efficiency of the b-jet selection in the b-candidate jet sample. One has to be careful in the ϵ_{b-tag} calculation since non-b-quark jets can also be mis-tagged. Considering the mis-tag rate one can write the ϵ_{b-tag} as

$$\epsilon_{b-tag} = \alpha \cdot \widehat{\epsilon_b} + (1 - \alpha) \cdot \overline{\epsilon_b}, \tag{6.8}$$

where $\alpha \sim 0.47$ is the fraction of the b-quark jets in the un-tagged b-candidate jet sample. Within the same jet sample, $\widehat{\epsilon_b}$ is the estimated b-tagging efficiency and $\overline{\epsilon_b}$ is the mis-tag rate, the b-tagging efficiency for the non-b-quark jets in the un-tagged b-candidate jet sample.
The mis-tag rate here is obtained using the generator level information in the simulation and is found to be $\sim 20\%$. For a data driven $\overline{\epsilon_b}$ estimation in the $t\bar{t}$ events, a lot of statistics is needed.
To study the potential correlation between $\sigma_{t\bar{t}}$ and $\widehat{\epsilon_b}$ evaluated within the top-quark events, first, the estimator $\widehat{\epsilon_b}$ for the TCHE b-tagging algorithm at the loose working point is estimated within the same set of pseudo-experiments used for the cross section

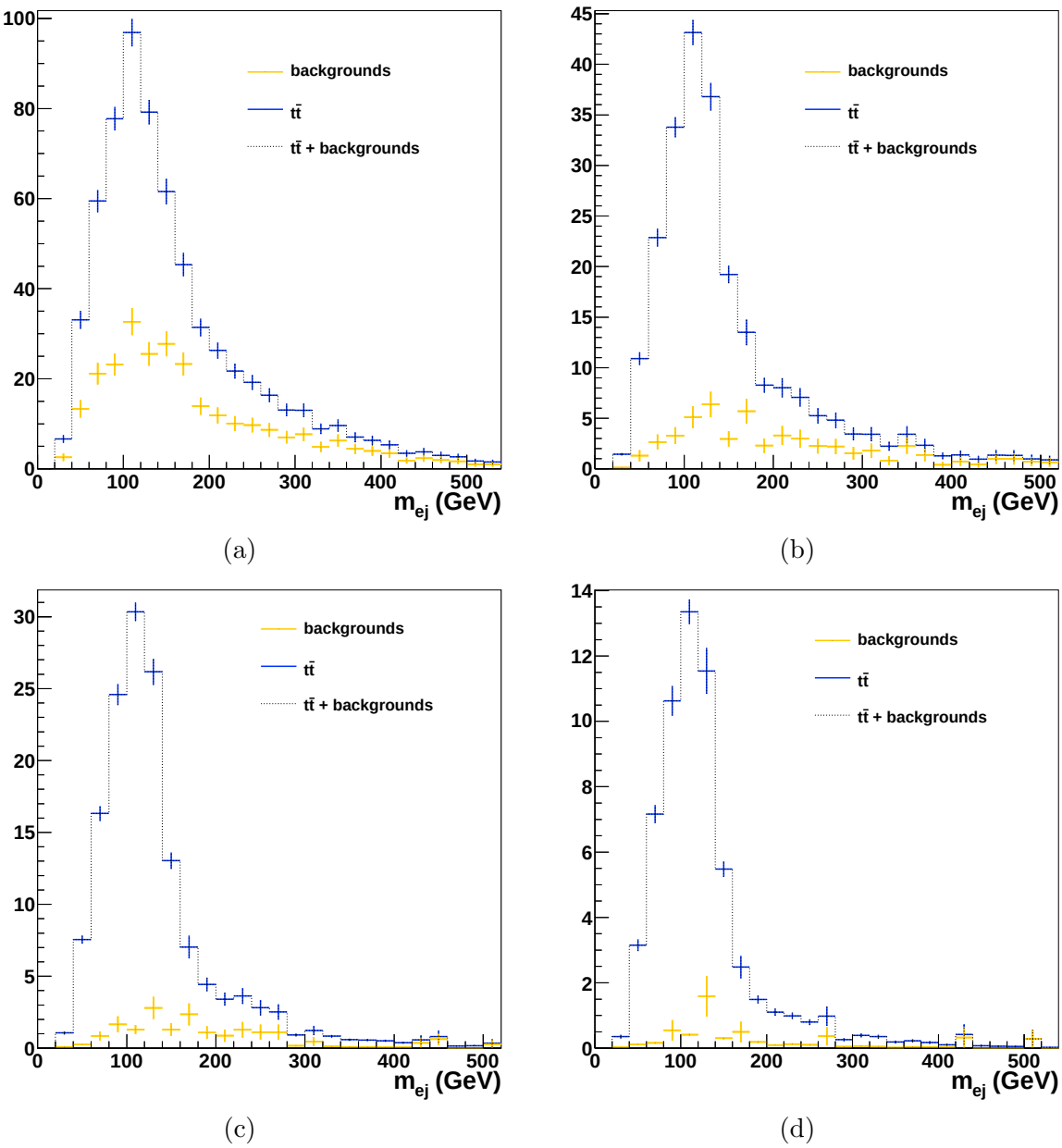
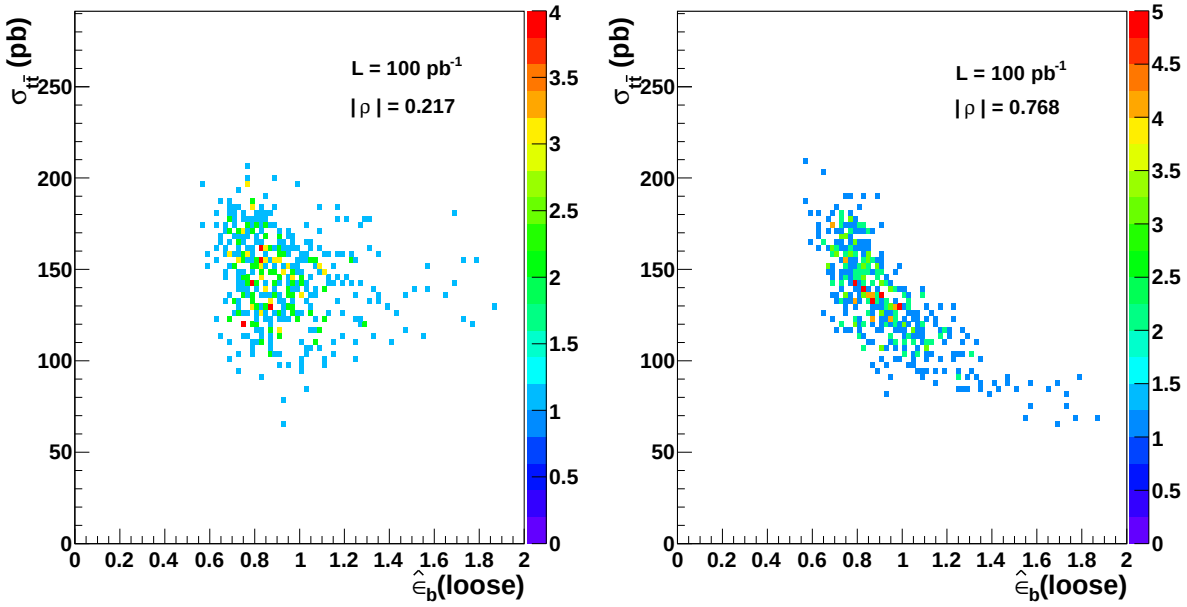


Figure 6.7: The effect of the TCHE b-tag selection on the signal and background contributions in the b-candidate jet sample: no b-tag selection, (a), together with loose, (b), medium, (c), and tight, (d), b-tag selections.

estimation in Section 6.3.1. Figure 6.8 (a) illustrates the estimated $t\bar{t}$ cross section versus the obtained b-tagging efficiency. As indicated on the plot itself, there is a very small correlation beween $\widehat{\sigma}_{t\bar{t}}$ and $\widehat{\epsilon_b}$ when the b-tagging requirement is not applied for the cross section measurement.

Another round of pseudo-experiments is run with the loose b-tag selection on the



(a) (b)

Figure 6.8: The estimated $t\bar{t}$ cross section versus the b-tagging efficiency for 500 pseudo-experiments at $L = 100\,pb^{-1}$. The estimation with no b-tagging requirement, (a), can be compared with the b-tagged cross section estimation at the loose b-tagging working point, (b).

b-candidate jet sample. The $\widehat{\sigma}_{t\bar{t}}^{tagged}$ is estimated in each pseudo-experiment according to Equation 6.7. Figure 6.8 (b) is the distribution of $\widehat{\sigma}_{t\bar{t}}^{tagged}$ versus $\widehat{\epsilon_b}$ for about 500 pseudo-experiments. A clear correlation appears for the combined $\widehat{\sigma}_{t\bar{t}}; \epsilon_b$ evaluation after the use of b-tagging.

Within the current set of pseudo-experiments, a histogram is filled with the estimated cross sections. The mean value and the width of a Gaussian fitted to this distribution are taken as an estimate for the $t\bar{t}$ cross section and its statistical uncertainty, respectively. This results in $\widehat{\sigma}_{t\bar{t}} = (161.2 \pm 17.5)\,pb$.

The combined measurement on data would result in a b-tagging efficiency of $\widehat{\epsilon_b}$ with an uncertainty of $\delta\epsilon_b$ as well as a $t\bar{t}$ cross section of $\widehat{\sigma}_{t\bar{t}}$ accompanied by $\delta\sigma_{t\bar{t}}$ uncertainty. A $\Delta\chi^2$ can be defined as

$$\Delta\chi^2 = \left(\frac{\sigma_{t\bar{t}} - \widehat{\sigma_{t\bar{t}}}}{\delta\sigma_{t\bar{t}}}\right)^2 + \left(\frac{\epsilon_b - \widehat{\epsilon_b}}{\delta\epsilon_b}\right)^2 \tag{6.9}$$

by which the uncertainty contours in the $\sigma_{t\bar{t}}; \epsilon_b$ plane are obtained around the central $(\widehat{\sigma_{t\bar{t}}}, \widehat{\epsilon_b})$ values. Figure 6.9 illustrates the possible results on data at $L = 100\,pb^{-1}$. Regarding the observed correlation, each measured value for the $t\bar{t}$ cross section con-

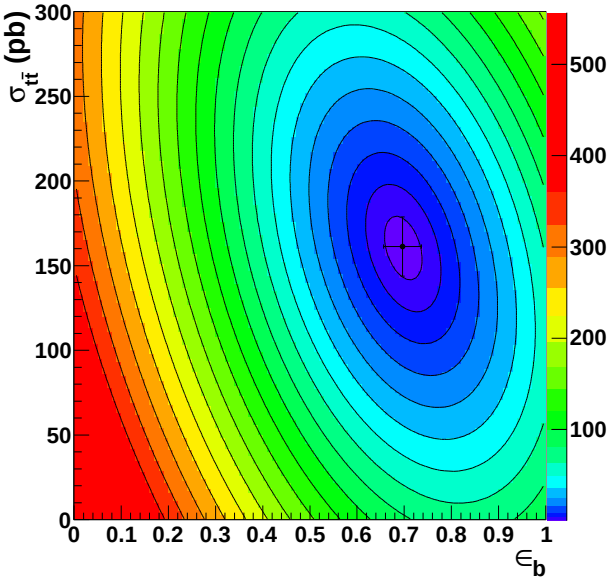


Figure 6.9: The estimated $t\bar{t}$ cross section together with the estimated b-tagging efficiency at the loose working point at $L = 100\,pb^{-1}$ integrated luminosity. The contours (Equation 6.9, $\Delta\chi^2 = \mathrm{C}^2$) illustrate the lines reflecting different standard deviations (i.e. C = {1,2,3, ...}).

tains a measurement for the b-tagging efficiency and implies an uncertainty on it. In another words, the systematic uncertainty on the b-tagging efficiency is already absorbed in the statistical uncertainty on the cross section. Therefore, an improved total uncertainty is achieved comparing to the one dimensional $\sigma_{t\bar{t}}$ measurement. The method is applicable for any b-tagging algorithm at every desired working point.

Appendix A

Pauli and Dirac matrices

Pauli matrices are the 2×2 unitary and complex Hermitian matrices referred to as the generators of the $SU(2)$ group,

$$\tau^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \tau^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \tau^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} .$$

They obey the following commutation relation

$$\left[\tau^i, \tau^j\right] = 2i\epsilon_{ijk}\tau^k$$

where ϵ_{ijk} is the totally anti-symmetric Levi-Civita tensor. Each of the Pauli matrices has two eigenvalues, +1 and -1 where the eigen vectors of τ^3 are

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

respectively. To represent the physical quantities like spin and isospin, the matrices are multiplied by $\frac{1}{2}$ giving $\pm\frac{1}{2}$ eigenvalues.
The $\begin{pmatrix} \nu_e \\ e \end{pmatrix}$ doublet in Section 1.1 can be written as a linear combination of

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L = \nu_{e,L} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + e_L \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Therefore, an isospin of $+(-)\frac{1}{2}$ is assigned to the neutrino (electron).

Dirac matrices These 4×4 Dirac matrices, $\{\gamma^0, \gamma^1, \gamma^3, \gamma^4\}$, act on the space of Dirac spinors to generate the infinitesimal Lorentz transformations in that space. They obey the anticommutation relation of

$$\{\gamma^i, \gamma^j\} = 2\eta^{ij} I,$$

where I is the unity matrix and η^{ij} is the Minkowski metric with the signature of (+,-,-,-). The γ matrices take the following forms in the Dirac representation,

$$\gamma^0 = \begin{pmatrix} I_{2\times2} & 0 \\ 0 & -I_{2\times2} \end{pmatrix} \quad \text{and} \quad \gamma^i = \begin{pmatrix} 0 & \tau^i \\ -\tau^i & 0 \end{pmatrix} \quad \text{for} \quad i = \{1, 2, 3\} .$$

The notaition $I_{2\times 2}$ stands for the 2×2 unity matrix and τ^i is the i'th Pauli matrix which is already introduced. The γ matrices are extensively used to develop the Lagrangian of spinors in quantum theories.

Appendix B

The lepton-b-quark correlation in the leptonic top-quark decay

Taken the $t \to bW \to bl\nu_l$ decay ($l = e, \mu$)in the top-qaurk rest frame and put the z-axis in the momentum direction of the W-boson (Figure B.1), one obtains the energy of the b-quark and the W-boson considering the 4-momentum conservation:

$$E_b = \frac{m_{top}^2 - m_W^2 + m_b^2}{2m_{top}}. \tag{B.1}$$

The m_b can be neglectad comparing the other masses and energies in the equations.

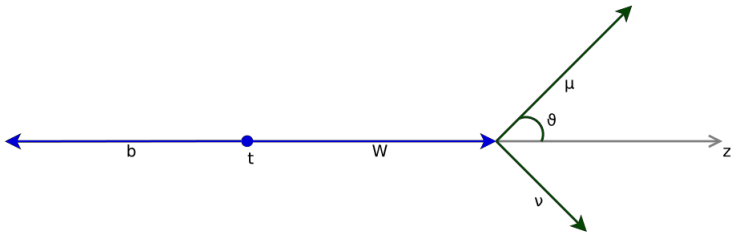


Figure B.1: The leptonic top-quark decay in its rest frame.

In the rest fram of the W-boson (Figure B.2) the energy of the lepton is calculated in a similar way. Taking $m_\nu \approx 0$ the lepton energy is

$$E_l' = \frac{m_W^2 + m_l^2}{2m_W} \tag{B.2}$$

where it can be simplified to

$$E_l' = \frac{1}{2} m_W \tag{B.3}$$

given $m_l \ll m_W$.

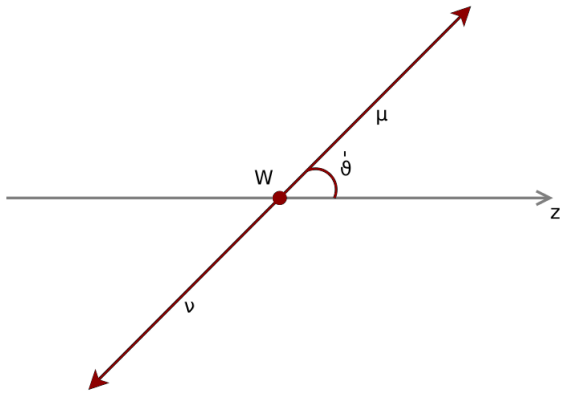


Figure B.2: The leptonic W-boson decay in its rest frame.

The invariant mass of the lepton and the b-quark in the top-quark rest frame can then be written as

$$M_{lb}^2 = (E_b + E_l)^2 - (\vec{p_b} + \vec{p_l})^2 , \tag{B.4}$$
$$= m_b^2 + m_l^2 + 2\left(E_b E_l + p_b p_l^z\right), \tag{B.5}$$

Given $m_b, m_l \ll E_b, E_l$, the M_{lb} becomes

$$M_{lb}^2 = 2E_b\left(E_l + p_l^z\right). \tag{B.6}$$

The Lorentz boost which transfroms the W-boson rest frame to the top-quark one is

$$\begin{bmatrix} E' \\ p'_x \\ p'_y \\ p'_z \end{bmatrix} = \begin{bmatrix} \gamma_W & 0 & 0 & -\beta_W\gamma_W \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta_W\gamma_W & 0 & 0 & \gamma_W \end{bmatrix} \begin{bmatrix} E \\ p_x \\ p_y \\ p_z \end{bmatrix}, \tag{B.7}$$

where $\beta_W \equiv v_W$ and $\gamma_W = \frac{1}{\sqrt{1-\beta_W^2}}$. Hence the energy and momentum of the lepton can be expressed in the top-quark rest frame:

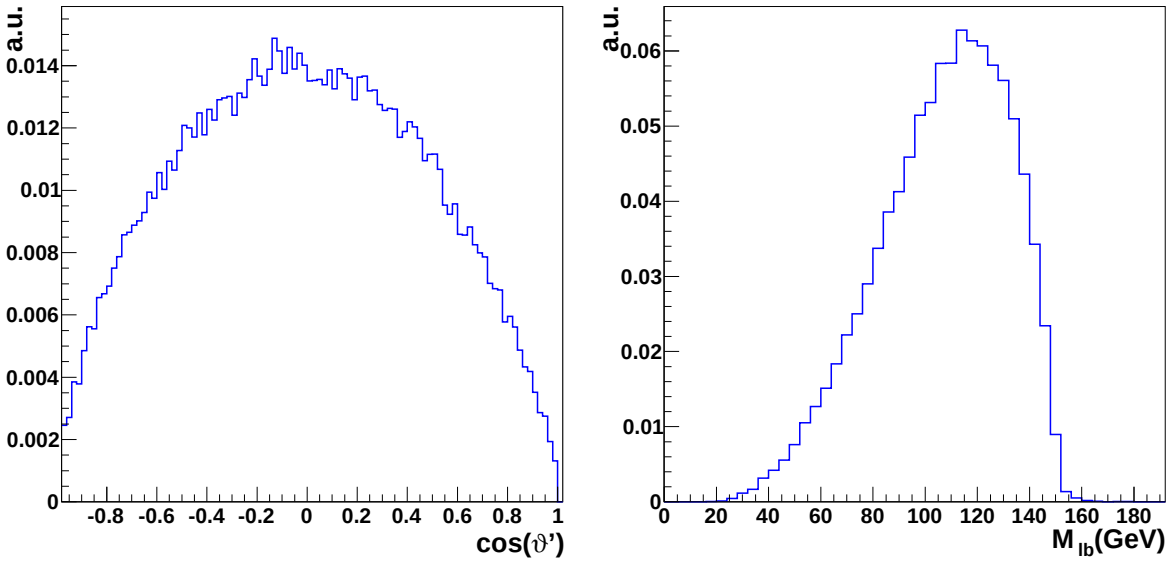
$$E_l = \gamma_W\left(E'_l + \beta_W p'^z_l\right) \tag{B.8}$$
$$= \gamma_W E'_l\left(1 + \beta_W \cos\vartheta'\right), \tag{B.9}$$
$$p_l^z = \gamma_W\left(\beta_W E'_l + p'^z_l\right) \tag{B.10}$$
$$= \gamma_W E'_l\left(\beta_W + \cos\vartheta'\right), \tag{B.11}$$
$$\tag{B.12}$$

where ϑ' is the angle between the muon momentum direction in the rest frame of the W-boson and the W-boson direction in the top-quark rest frame. Combining with Equation B.6, the M_{lb} can finally be written as

$$M_{lb}^2 = \frac{m_{top}^2 - m_W^2}{2}\left(1 + \cos\vartheta'\right). \tag{B.13}$$

If the $\cos\vartheta'$ had a uniform probability distribution function, the M_{lb} distribution would have been increasing, stopped sharply at $M_{lb} = \sqrt{\frac{m_{top}^2 - m_W^2}{2}}$. However, the angular momentum conservation in the W-boson rest fram and the fact that the ν_l is left handed impose some restrictions on the $\cos\vartheta'$ values and make the its p.d.f non-uniform as in Figure B.3 (a). Hence, the invariant mass between the lepton and the b-quark takes the shape as in Figure B.3 (b).

This distribution plays a key role in the analysis presented in this thesis since the M_{lj}



(a) (b)

Figure B.3: The normal distribution for the $\cos\vartheta'$ (a) and the lepton-b-quark invariant mass (b).

for all leptonic b-jet candiates that are truly originated from a b-quark are expected to follow this shape. Therefore, being limited to the bulk area, the the b-jet candiates sample can be enriched in true b-jets. Of course specially in the presence of other background processes, this jet sample needs to be purified as explained in Section 5.3.1.

Bibliography

[1] M. Peskin and D. Schroeder, *Introduction to quantum field theory*. Advanced Book Program. Addison-Wesley Pub. Co., 1995.

[2] T. Morii, C. Lim, and S. Mukherjee, *The physics of the standard model and beyond*. World Scientific, 2004.

[3] K. Nakamura *et al.*, J. Phys. G **37** (2010), no. 7A, 075021.

[4] H. Inazawa and T. Morii, Physics Letters B **203** (1988), no. 3, 279 – 282.

[5] M. J. Strassler and M. E. Peskin, Phys. Rev. D **43** (Mar, 1991) 1500–1514.

[6] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49** (1973) 652–657.

[7] M. Bustamante, L. Cieri, and J. Ellis, *Beyond the Standard Model for Montaneros*, `0911.4409`.

[8] WMAP Collaboration, J. Dunkley *et al.*, Astrophys. J. Suppl. **180** (2009) 306–329.

[9] Supernova Search Team Collaboration, A. G. Riess *et al.*, Astron. J. **116** (1998) 1009–1038.

[10] J. Ellis and D. Ross, Physics Letters B **506** (May, 2001) 331–336.

[11] R. N. Mohapatra *et al.*, Rept. Prog. Phys. **70** (2007) 1757–1867.

[12] L. Evans and P. Bryant, JINST **3 S08001** (2008).

[13] *The CDF experiment*, `http://www-cdf.fnal.gov/physics/physics.html`.

[14] *The D0 experiment*, `http://www-d0.fnal.gov/`.

[15] The CDF Collaboration, Phys. Rev. Lett. **74** (Apr, 1995) 2626–2631.

[16] The D0 Collaboration, Phys. Rev. Lett. **74** (Apr, 1995) 2632–2637.

[17] *The Tevatron Collider*, `http://www-bdnew.fnal.gov/tevatron/`.

[18] CDF Collaboration Collaboration, T. Aaltonen *et al.*, Phys.Rev.Lett. **103** (2009) 092002.

[19] D0 Collaboration, V. M. Abazov *et al.*, Phys. Rev. Lett. **103** (2009) 092001.

[20] T. E. W. Group, CDF, and D. Collaborations, *Combination of CDF and DO results on the mass of the top quark using up to 5.8 fb-1 of data*, 1107.5255.

[21] The ATLAS Collaboration, JINST **3 S08003** (2008).

[22] The CMS Collaboration, JINST **3 S08004** (2008).

[23] Atlas Collaboration, G. Aad *et al.*, Eur. Phys. J. **C71** (2011) 1577.

[24] CMS Collaboration, V. Khachatryan *et al.*, Phys. Lett. **B695** (2011) 424–443.

[25] CMS Collaboration, S. Chatrchyan *et al.*, *Measurement of the Top-antitop Production Cross Section in pp Collisions at sqrt(s)=7 TeV using the Kinematic Properties of Events with Leptons and Jets*, 1106.0902.

[26] CMS Collaboration, S. Chatrchyan *et al.*, JHEP **07** (2011) 049.

[27] CMS Collaboration, S. Chatrchyan *et al.*, *Measurement of the t-channel single top quark production cross section in pp collisions at sqrt(s) = 7 TeV*, 1106.3052.

[28] The D0 Collaboration Collaboration, V. M. Abazov *et al.*, Phys. Rev. Lett. **106** (Jan, 2011) 022001.

[29] Y. Grossman and I. Nachshon, JHEP **07** (2008) 016.

[30] R. D. Peccei, S. Peris, and X. Zhang, Nuclear Physics B **349** (1991), no. 2, 305 – 322.

[31] J. Ellis and G. Fogli, Physics Letters B **231** (1989), no. 1-2, 189 – 194.

[32] ALEPH Collaboration, *et al.*, *Precision Electroweak Measurements and Constraints on the Standard Model*, 1012.2367.

[33] The CMS Collaboration, *Search for Resonances in Semi-leptonic Top-pair Decays Close to Production Threshold*, CMS PAS TOP-10-007.

[34] The CMS Collaboration, *Measurement of the charge asymmetry in top quark pair production with the CMS experiment*, CMS PAS TOP-10-010.

[35] CMS Collaboration, S. Chatrchyan *et al.*, Phys. Lett. **B701** (2011) 204–223.

[36] DØ. public page, *Useful Diagrams of Top Signals and Backgrounds*, http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages.

[37] P. Van Mulders, *Calibration of the jet energy scale using top quark events at the LHC*, CMS TS-2011/003.

[38] J. Baglio and A. Djouadi, JHEP **10** (2010) 064.

[39] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, and R. Tanaka (Eds.), CERN-2011-002 (CERN, Geneva, 2011).

[40] H. M. Georgi, M. E. Glashow, and D. V. Nanopoulos, Phys. Rev. Lett. **40** (1978) 692.

[41] M. Dhrssen *et al.*, Phys. Rev. **D 70** (2004) 113009.

[42] V. Hankele, D. Zeppenfeld, and T. Figy, Phys. Rev. **D 74** (2006) 095001.

[43] CMS Collaboration, Phys. Lett. B **699** (2011) 25–47.

[44] J. Ellis, *Searching for Particle Physics Beyond the Standard Model at the LHC and Elsewhere*, arXiv:1102.5009v1[hep-ph]. Presented at the 11th conference on "Frontiers of Fundamental Physics", Paris, July 2010.

[45] CMS Collaboration, Physics Letters B **698** (2011), no. 3, 196 – 218.

[46] ATLAS Collaboration, *Search for supersymmetry using final states with one lepton, jets, and missing transverse momentum with the ATLAS detector in* $\sqrt{s}$*=7 TeV pp collisions*, arXiv:1102.2357[hep-ex].

[47] CMS Collaboration, Phys. Rev. Lett. **105** (2010) 211801.

[48] ATLAS Collaboration, Phys. Rev. Lett. **105** (2010) 161801.

[49] S. B. Giddings and S. D. Thomas, Phys. Rev. **D 65** (2002) 056010.

[50] S. Di-mopoulos and G. L. Landsberg, Phys. Rev. Lett. **87** (2001) 161602.

[51] C. M. Harris, M. J. Palmer, M. A. Parker, P. Richardson, A. Sabetfakhri, and B. R. Webber, JHEP **0505** (2005) 053.

[52] CMS Collaboration, Phys. Lett. B.

[53] *Taking a cloder look at the LHC*, http://www.lhc-closer.es/.

[54] J. Jeanneret, D. Leroy, L. Oberli, and T. Trenkler, *Quench levels and transient beam losses in LHC magnets*, LHC-Project-Report-44.

[55] J. Jeanneret, D. Leroy, L. Oberli, and T. Trenkler, *Equilibrium beam distribution and halo in the LHC*, LHC-Project-Report-592 and EPAC02.

[56] *LHC Collimation Project*, http://lhc-collimation-project.web.cern.ch/.

[57] *The CERN Hadron Ion Sources*, http://linac2.home.cern.ch/linac2/sources/source.htm.

[58] *CERN Hadron Linacs*, http://linac2.home.cern.ch/linac2/.

[59] The LHCb Collaboration, JINST **3 S08005** (2008).

[60] The ALICE Collaboration, JINST **3 S08002** (2008).

[61] The TOTEM Collaboration, JINST **3 S08007** (2008).

[62] The LHCf Collaboration, JINST **3 S08006** (2008).

[63] *Pierre Auger Observatory*, http://www.auger.org/.

[64] *Telescope Array*, http://www.telescopearray.org/.

[65] ALICE Collaboration, Phys. Rev. Lett.

[66] LHCb Collaboration, Phys. Lett. B.

[67] *CERN Document Server*, http://cdsweb.cern.ch/.

[68] CMS Collaboration, Eur. Phys. J.

[69] CMS Collaboration, V. Khachatryan *et al.*

[70] CMS Collaboration, *Strange Particle Production in pp Collisions at* $\sqrt{s}$ *= 0.9 and 7 TeV*, arXiv:1102.4282v1[hep-ex].

[71] CMS Collaboration, JHEP **03** (2011) 090.

[72] CMS Collaboration, JHEP **01** (2011) 080.

[73] CMS Collaboration, Phys. Lett. **B699** (2011) 25–47.

[74] CMS Collaboration, JHEP **09** (2010) 091.

[75] CMS Collaboration, *Search for Physics Beyond the Standard Model in Opposite-Sign Dilepton Events at sqrt(s) = 7 TeV*, arXiv:1103.1348[hep-ex].

[76] CMS Collaboration, *Search for Supersymmetry in pp Collisions at sqrt(s) = 7 TeV in Events with Two Photons and Missing Transverse Energy*, arXiv:1103.0953v2[hep-ex].

[77] CMS Collaboration, Phys. Rev. Lett. **105** (2010) 262001.

[78] CMS Collaboration, *Search for a W' boson decaying to a muon and a neutrino in pp collisions at sqrt(s) = 7 TeV*, arXiv:1103.0030[hep-ex].

[79] CMS Collaboration, *Search for Neutral MSSM Higgs Bosons Decaying to Tau Pairs in pp Collisions at* $\sqrt{s}$ *= 7 TeV*, arXiv:1104.1619v1[hep-ex].

[80] CMS Collaboration, *Observation and studies of jet quenching in PbPb collisions at* $\sqrt{s_{NN}}$ *= 2.76 TeV*, arXiv:1102.1957v2[hep-ex].

[81] *CMS physics Technical Design Report.*

[82] CMS Collaboration, Journal of Instrumentation **5** (2010) T03009.

[83] The CMS Collaboration, *Inclusive total and differenctial production cross sections of J/ψ and b-hadron production in pp collisions at $\sqrt{s} = 7$ TeV with the CMS experiment*,.

[84] W. Adam *et al.*, *Track reconstruction in the CMS tracker*, CMS NOTE-2006/041.

[85] R. Frühwirth, Nuclear Instruments and Methods in Physics Research A **262** (1987) 444.

[86] The CMS Collaboration, *Tracking and Vertexing Results from First Collisions*,.

[87] The CMS Collaboration, *Tracking and Primary Vertex Results in First 7 TeV Collisions*,.

[88] W. Waltenberger, R. Frühwirth, and P. Vanlaer, *Adaptive Vertex Fitting*, CMS NOTE-2007/008.

[89] *Vertex Fitting in the CMS Tracker*, CMS NOTE-2006/032.

[90] *Sensitivity of Robust Vertex Fitting Algorithms*, CMS NOTE-2004/002.

[91] CMS Collaboration, *Electromagnetic calorimeter commissioning and first results with 7 TeV data*, CMS NOTE-2010/012.

[92] A. Kyriakis and E. Petrakou, *Electron Position Resolution at the CMS Endcaps with the Preshower Detector*, CMS AN-2009/116.

[93] The CMS HCAL Collaboration, Eur. Phys. J. **C55** (2008) 159.

[94] CMS Collaboration, Journal of Instrumentation **5** (2010) T03001.

[95] The CMS Collaboration, *Jets in 0.9 and 2.36 TeV pp Collisions*, CMS PAS JME-10-001.

[96] The CMS Collaboration, *Performance of Missing Transverse Energy Reconstruction in $\sqrt{s}$ = 900 and 2360 GeV pp Collision Data*, CMS PAS JME-10-002.

[97] The CMS Collaboration, *Jet Performance in pp Collisions at $\sqrt{s}$=7 TeV*, CMS PAS JME-10-003.

[98] The CMS Collaboration, *Missing Transverse Energy Performance in Minimum-Bias and Jet Events from Proton-Proton Collisions at $\sqrt{s}$=7 TeV*, CMS PAS JME-10-004.

[99] E. Meschi *et al.*, *Electron Reconstruction in the CMS Electromagnetic Calorimeter*, CMS Note-2001/034.

[100] N. Adam *et al.*, *Electron Reconstruction at Low p_T*, CMS Note-2009/074.

[101] CMS Collaboration, Phys. Lett. **B692** (2010) 83–104.

[102] J. Bernardini *et al.*, *Momentum analysis with cosmics at B=0*, CMS Note-2009/065.

[103] CMS Collaboration, JINST **5** (2010) T03022.

[104] The CMS Collaboration, *Performance of muon identication in pp collisions at* $\sqrt{s}$*=7 TeV*, CMS PAS MUO-10-002.

[105] for the CMS Collaboration, M. Chiorboli, *Trigger Issues for New Physics Searches in the CMS Experiment*, Presented at ICHEP2010: 35th ICHEP conference.

[106] R. Covarelli, *The CMS High-Level Trigger*, Tech. Rep. CMS-CR-2009-188. CERN-CMS-CR-2009-188, CERN, Geneva, Jul, 2009.

[107] *CMSSW Application Framework*, `https://twiki.cern.ch/twiki/bin/view/CMS/WorkBookCMSSWFramework`.

[108] *ROOT, A Data Analysis Framework*, `root.cern.ch`.

[109] T. Orimoto, *Central Skims and Secondary Datasets 2011*,.

[110] A. Fanfani *et al.*, J. Grid Computing (2010) 1572–9814.

[111] *Worldwide LHC Computing Grid (WLCG)*, `http://lcg.web.cern.ch/LCG/public/default.htm`.

[112] C. Collaboration, Journal of Instrumentation **5** (2010), no. 03, T03006.

[113] G. Codispoti *et al.*, PoS **ACAT08** (2008) 029.

[114] D. Spiga *et al.*, Journal of Physics: Conference Series **219** (2010), no. 7, 072019.

[115] L. Tuura, A. Meyer, I. Segoni, and G. D. Ricca, Journal of Physics: Conference Series **219** (2010), no. 7, 072020.

[116] A. Afaq *et al.*, Journal of Physics Conference Series **119** (July, 2008).

[117] *Top commissioning: Operations*, `https://twiki.cern.ch/twiki/bin/view/CMS/TWikiTopQuarkComWG4`.

[118] D. J. Gross and F. Wilczek, Phys. Rev. Lett. **30** (Jun, 1973) 1343–1346.

[119] H. D. Politzer, Phys. Rev. Lett. **30** (Jun, 1973) 1346–1349.

[120] T. Sjostrand, S. Mrenna, and P. Z. Skands, JHEP **05** (2006) 026.

[121] T. Stelzer and W. F. Long, Comput. Phys. Commun. **81** (1994) 357–371.

[122] F. Maltoni and T. Stelzer, JHEP **02** (2003) 027.

[123] J. M. Campbell, J. W. Huston, and W. J. Stirling, Rept. Prog. Phys. **70** (2007) 89.

[124] S. Frixione and B. R. Webber, JHEP **06** (2002) 029.

[125] S. Frixione, P. Nason, and B. R. Webber, JHEP **08** (2003) 007.

[126] M. A. Dobbs *et al.*, ArXiv High Energy Physics - Phenomenology e-prints (Mar., 2004).

[127] J. C. Collins, D. E. Soper, and G. F. Sterman, Adv. Ser. Direct. High Energy Phys. **5** (1988) 1–91.

[128] M. L. Mangano, *Two lectures on heavy quark production in hadronic collisions*, `hep-ph/9711337`.

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

[130] D. Stump *et al.*, JHEP **10** (2003) 046.

[131] Y. L. Dokshitzer, Sov. J. Phys. JETP **46** (1977) 641.

[132] V. N. Gribov and L. N. Lipatov, Sov. J. Nucl. Phys. **15** (1972) 438.

[133] G. Altarelli and P. G, Nucl. Phys. **B126** (1977) 298.

[134] E. Norrbin and T. Sjöstrand, European Physical Journal C **17** (Oct., 2000) 137–161.

[135] S. Hoche, F. Krauss, N. Lavesson, L. Lönnblad, M. Mangano, A. Schalicke, and S. Schumann, *Matching Parton Showers and Matrix Elements*, `arXiv:hep-ph/0602031`.

[136] M. Mangano, M. Moretti, F. Piccinini, and M. Treccani, *Matching matrix elements and shower evolution for top-quark production in hadronic collisions*, `arXiv:hep-ph/0611129`.

[137] B. Andersson, Acta Phys. Polon. **B32** (2001) 3993–4011.

[138] C. Peterson, D. Schlatter, I. Schmitt, and P. M. Zerwas, Phys. Rev. **D27** (1983) 105–111.

[139] G. Altarelli, N. Cabibbo, G. Corbo, L. Maiani, and G. Martinelli, Nucl. Phys. **B208** (1982) 365–380.

[140] T. Sjöstrand and P. Skands, JHEP **03** (2004) 053.

[141] R. Field, ArXiv e-prints (Oct., 2010).

[142] The CMS Collaboration, *Measurement of the Underlying Event Activity at the LHC with $\sqrt{s} = 7$ TeV and Comparison with $\sqrt{s} = 0.9$ TeV*,.

[143] W. Beenakker, W. V. Neerven, R. Meng, G. Schuler, and J. Smith, Nuclear Physics B **351** (1991), no. 3, 507 – 560.

[144] J. M. Campbell and R. K. Ellis, Phys. Rev. **D65** (2002) 113007.

[145] *Standard Model Cross Sections for CMS at 7 TeV*, `https://twiki.cern.ch/twiki/bin/view/CMS/StandardModelCrossSections`.

[146] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, Eur. Phys. J. **C64** (2009) 653–680.

[147] F. Demartin, S. Forte, E. Mariani, J. Rojo, and A. Vicini, Phys. Rev. **D82** (2010) 014002.

[148] N. Kidonakis, Phys. Rev. **D82** (2010) 114030.

[149] V. Ahrens, A. Ferroglia, M. Neubert, B. D. Pecjak, and L. L. Yang, Journal of High Energy Physics **9** (Sept., 2010) 97–+.

[150] GEANT4 Collaboration, S. Agostinelli *et al.*, Nucl. Instrum. Meth. **A506** (2003) 250–303.

[151] W. Adam *et al.*, *Electron reconstruction at CMS*, CMS AN-2009/164.

[152] The CMS Collaboration, *Particle-Flow Event reconstruction in CMS and Performance for Jets, Taus, and* $E_{\mathrm{T}}^{\mathrm{miss}}$, CMS PAS PFT-09/001.

[153] M. Pioppi, *Electron Pre-identification in the Particle Flow framework*, CMS AN-2008/032.

[154] H. Bethe and W. Heitler, Proceedings of the Royal Society of London. Series A **146** (1934), no. 856, 83–112.

[155] R. Frühwirth, Computer Physics Communications **100** (1997), no. 1-2, 1 – 16.

[156] R. Frühwirth and S. Frühwirth-Schnatter, Computer Physics Communications **110** (1998), no. 1-3, 80 – 86.

[157] W. Adam *et al.*, *Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC*, CMS AN-2005/001.

[158] The CMS Collaboration, *Electron reconstruction and identication at* $\sqrt{s}$*=7 TeV*, CMS PAS EGM-10-004.

[159] *Electron Identification*, `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideElectronID`.

[160] S. Baffioni *et al.*, *Electrons Identification in CMS*, CMS AN-2009/178.

[161] A. Vartak, M. LeBourgeois, and V. Sharma, *Lepton Isolation in the CMS Tracker, ECAL and HCAL*, CMS AN-2010/106.

[162] D. Barge *et al.*, *Study of photon conversion rejection at CMS*, CMS AN-2009/159.

[163] The CMS Collaboration, *Measurement of the $t\bar{t}$ Pair Production Cross Section at $\sqrt{s}$=7 TeV using b-quark Jet Identification Techniques in Lepton + Jet Events*, CMS PAS TOP-10-003.

[164] F. Blekman *et al.*, *Top Lepton Plus Jets: Electron and Muon Efficiency Measurements for 2010 Dataset*, CMS AN-10/362.

[165] J. Berryhill *et al.*, *Electron Efficiency Measurements with 2.88 pb^{-1} of pp Collision Data at $\sqrt{s}$= 7 TeV*, CMS AN-10/323.

[166] The CMS Collaboration, *Performance of Jet Reconstruction with Charged Tracks only*, CMS PAS JME-08-001.

[167] The CMS Collaboration, *Jet Plus Tracks Algorithm for Calorimeter Jet Energy Corrections in CMS*, CMS PAS JME-09-002.

[168] The CMS Collaboration, *Commissioning of the Particle-Flow Reconstructionin Minimum-Bias and Jet Events from pp Collisions at 7 TeV*, CMS PAS PFT-10-002.

[169] G. P. Salam and G. Soyez, JHEP **05** (2007) 086.

[170] M. Cacciari and G. P. Salam, PHYS.LETT.B **641** (2006) 57.

[171] M. Cacciari, G. P. Salam, and G. Soyez, JHEP **04** (2008) 063.

[172] The CMS Collaboration, *Performance of Jet Algorithms in CMS*, CMS PAS JME-07-003.

[173] A. Heister *et al.*, *Measurement of Jets with the CMS Detector at the LHC*, CMS NOTE-2006/036.

[174] The CMS Collaboration, *Jet Energy Calibration and Transverse Momentum Resolution in CMS*, CMS PAPER JME-10-011.

[175] A. Harel and P. Schieferdecker, *Calorimeter Jet Quality Criteria for the First CMS Collision Data*, CMS AN-2009/087.

[176] ATLAS Collaboration, *Search for supersymmetry in pp collisions at sqrts = 7 TeV in final states with missing transverse momentum and b-jets*, `arXiv:1103.4344v1 [hep-ex]`.

[177] The CMS Collaboration, *Search for Supersymmetry in Final States with b-Jets and Missing Energy at the LHC*, CMS PAS SUS-10-011.

[178] The CMS Collaboration, *Algorithms for b Jet Identification in CMS*, CMS PAS BTV-09-001.

[179] *Inclusive secondary vertex reconstruction in jets*, CMS NOTE-2006/027.

[180] The CMS Collaboration, *Performance Measurement of b-tagging Algorithms Using Data containing Muons within Jets*, CMS PAS BTV-07-001.

[181] The CMS Collaboration, *Evaluation of udsg Mistags for b-tagging using Negative Tags*, CMS PAS BTV-07-002.

[182] The CMS Collaboration, *Commissioning of b-jet identification with pp collisions at* $\sqrt{s}$*=7 TeV*, CMS PAS BTV-10-001.

[183] *Measurement of the Identification Efficiency for b-Quark Jets in 2011 Data using the Relative Transverse Momentum of Muons*, CMS NOTE AN-11/207.

[184] S. Bhattacharya *et al.*, *Measurement of the b-tagging efficiency using the System8 Method with 2011 Data*, CMS AN AN-11/195.

[185] CMS Collaboration, J. DHondt *et al.*, *Offline Calibration of b-Jet Identification Efficiencies*, CERN-CMS-NOTE-2006-013.

[186] J. Maes, *Estimation of the b-tag efficiency using top quarks at CMS*, CMS TS-2011/028.

[187] S. Beauceron, J. DHondt, and M. Zeinali, *Jets-Electron Cleaning via CaloTowers Selection*, CMS AN -2009/149.

[188] F. Caola, J. M. Campbell, F. Febres Cordero, L. Reina, and D. Wackeroth, *NLO QCD predictions for W+1 jet and W+2 jet production with at least one b jet at the 7 TeV LHC*, 1107.3714.

[189] S. Lowette, *B-Tagging as a Tool for Charged Higgs Boson Identification in CMS*, CMS TS-2007/003.

Summary

The Large Hadron Collider (LHC) set up a record for high energy collisions on March 30th 2010, by colliding proton beams at a center-of-mass energy of 7 TeV. The LHC physics program is to reveal the physics beyond the Standard Model and to search for the Higgs particle which is believed to be responsible for Electroweak Symmetry Breaking where the data is also used to ascertain the Standard Model of particle physics. Based on this program, different experiments like the Compact Muon Solenoid (CMS) experiment are designed to collect and analyze the LHC collision data.

In many data analyses, the jets originating from b-quarks are of special importance in discriminating between the physics signal of interest and the background processes that need to be discarded. This necessitates the development of algorithms to identify the b-quark jets using their distinct properties. In addition, data driven methods are needed to calibrate the performance of the b-jet identification algorithms.

Using top quarks which are produced in pair at a very high rate at the LHC, a data driven method is described in this thesis to measure the performance of the b-jet identification algorithms developed in CMS. This method is applied for the first time on the LHC collision data collected in 2010 by the CMS experiment.

Searching for the semi-electron final state of $t\bar{t}$ events, $t\bar{t} \rightarrow qq'b\bar{b}e\nu_e$, a dedicated selection is performed to prepare an event sample enriched with top-like events. Considering the fact that the top quark decays almost all the time to a b-quark and a W boson, the prepared sample is a rich source of b-quark jets and is well suited for measuring the performance of b-jet identification algorithms.

The non-b-quark jets present in the event together with one of the b-quark jets are considered as being the decay products of one of the top quarks using a jet-parton matching algorithm. The jet-parton matching algorithm uses the mass of the top quark and W boson as constraints. A jet sample is formed by the remaining jet out of four for which the b-quark jet content is $\sim 30\%$. This jet sample is further divided into a b-dominated and a b-depleted jet sample based on the kinematic correlations between the jet and the electron present in the final state. The b-dominated jet sample has a b-purity of $\sim 39\%$ while for the b-depleted jet sample, the b-purity is $\sim 11\%$. The b-dominated jet sample is purified even more using the information of non-b-quark jets in the b-depleted jet sample. In the purification of the b-dominated jet sample, the knowledge from the b-depleted sample is complemented by the information obtained form another jet sample, the control sample, to make the method absolutely independent from simulation. The control sample is constructed using the jets associated to the W boson by the jet-parton matching algorithm.

The b-jet identification (b-tagging) efficiency for the Track Counting High Efiiciency

b-tagging algorithm is estimated within the purified b-dominated jet sample. For an integrated luminosity of $100\,pb^{-1}$ at 7 TeV center-of-mass energy, it is expected to achieve an absolute (relative) statistical uncertainty of 3.6%(5.1%), 3.4%(7.2%) and 2.9%(11%), for the b-jet identification efficiency of about 75%, 50% and 25% , respectively. A conservative estimation of the systematic uncertainties leads to an absolute (relative) systematic uncertainty of 3.3%(4.7%), 2.5%(5.3%), 0.8%(3.3%).
The method is applied on the CMS 2010 data equivalent to an integrated luminosity of $36\,pb^{-1}$, resulting in a b-tagging efficiency of

$$\widehat{\epsilon_b}(\text{loose}) = 0.73 \pm 0.36\,(\text{stat.}),$$
$$\widehat{\epsilon_b}(\text{medium}) = 0.42 \pm 0.26\,(\text{stat.}),$$
$$\widehat{\epsilon_b}(\text{tight}) = 0.20 \pm 0.13\,(\text{stat.}).$$

The systematic uncertainties are not mentioned for this measurement since they are much smaller than the statistical uncertainties. More accurate results are expected with more accumulated data.
For an integrated luminosity of $100\,pb^{-1}$ the possibility of combining results of the semi-electron decay channel with the semi-muon final state of $t\bar{t}$ is investigated where it is shown that such combination can lead to a statistical uncertainty reduced by a factor of $1/\sqrt{2}$.
The prospect of extending the method towards a simultaneous $t\bar{t}$ cross section and b-jet identification efficiency measurement is also discussed. Due to the correlation between the $t\bar{t}$ cross section and the b-jet identification efficiency, the systematic uncertainty introduced by the b-jet identification is absorbed in the statistical uncertainty on the cross section by the simultaneous measurement. For an integrated luminosity of $100\,pb^{-1}$, the $t\bar{t}$ cross section with the use of b-jet identification is obtained from a template fit, resulting in a statistical uncertainty around $18\,pb^{-1}$.
It should be noted that for the $t\bar{t}$ cross section measurement in the semi-electron final state, the event selection includes the presence of one prompt electron in the event. The electron selection efficiency needs to be accounted for in the final determination of the cross section. The data driven measurement of the electron selection efficiency is performed using a Tag&Probe method within the $Z \to ee$ processes. It is assumed that the scale factors defined as the ratio between the electron efficiencies in data and simulation are the same in $Z \to ee$ and $t\bar{t}$ events. For an integrated luminosity of $36\,pb^{-1}$, the scale factors in $Z \to ee$ events are

$$SF_{id} = 0.98 \pm 0.02\,(\text{syst.} + \text{stat.}),$$
$$SF_{iso} = 1.009 \pm 0.007\,(\text{syst.} + \text{stat.}),$$

where id (iso) stands for the identification (isolation) requirements applied on the electron. Thesimulation-driven electron efficiency in the $t\bar{t}$ events is corrected with the scale factors. Using the Tag&Probe results in the calculation of the $t\bar{t}$ cross section, a systematic uncertainty of about 6% is introduced to cover the possible differences in the characteristics of the $Z \to ee$ and $t\bar{t}$ events.