Vrije Universiteit Brussel

Faculteit Wetenschappen
Departement Natuurkunde

Estimation of the b-tag efficiency using top quarks at CMS

Joris Maes

Promotor: Prof. Dr. Jorgen D'Hondt

Proefschrift ingediend met het oog op het behalen van
de academische graad Doctor in de Wetenschappen

November 2010

Cover illustration
Different visualizations of the reconstructed particles of an event passing the top quark event selection criteria in 7 TeV data detected by the CMS detector.

Print: Silhouet, Maldegem

2010 Uitgeverij VUBPRESS Brussels University Press
VUBPRESS is an imprint of ASP nv (Academic and Scientific Publishers nv)
Ravensteingalerij 28
B-1000 Brussels
Tel. +32 (0)2 289 26 50
Fax +32 (0)2 289 26 59
E-mail: info@vubpress.be
www.vubpress.be

ISBN 978 90 5487 817 9
NUR 924
Legal Deposit D/2010/11.161/140

Doctoral examination commission

Chair: Prof. Dr. Robert Roosen (VUB)
Supervisor: Prof. Dr. Jorgen D'Hondt (VUB)
Prof. Dr. Freya Blekman (VUB)
Prof. Dr. Jan Danckaert (VUB)
Prof. Dr. Stefaan Tavernier (VUB)
Prof. Dr. Gilles De Lentdecker (ULB)
Prof. Dr. Daniel Bloch (IPHC)
Prof. Dr. Ivo van Vulpen (NIKHEF)

Research funded by a Ph.D. grant of the Agency for innovation by Science and Technology (IWT).

Contents

Introduction

The Standard Model of elementary particles describes the smallest constituents of matter and the fundamental interactions between them, except gravity. In general the Standard Model is found to be in very good agreement with experimental observations except for the yet undiscovered Higgs boson. The Higgs boson is introduced in the Standard Model to give mass to the particles and is the only missing piece to complete the Standard Model. Despite the great success of the Standard Model, it does not provide a satisfying answer to several fundamental questions and should be extended towards a more general theory. To search for the Higgs boson and to discover unexpected phenomena, large particle collider experiments are needed. In collider experiments nature is probed at its smallest scale by colliding particles onto each other at very high energies. The current highest-energy particle collider is the Large Hadron Collider (LHC) located at the CERN laboratory near Geneva and collides proton beams at center-of-mass energies of 7 TeV. This allows the large particle detectors near the LHC to search for the Higgs boson and discover new physics phenomena.

One of the two main detectors operational at the LHC is the Compact Muon Solenoid (CMS) detector. For many studies, performed by the CMS experiment, b quark jets play an important role in separating interesting signal events from the large amount of background events. Therefore an accurate identification of b quark jets is crucial. Based on the specific properties of b quark jets, several dedicated b quark jet identification algorithms or b-tagging algorithms have been developed at the CMS experiment. The calibration of these algorithms is essential for the success of the physics program at the LHC. The top quark is known to decay nearly always to a b quark and can, given the large production rate of top quarks at the LHC, be used as a calibration tool. In this thesis a data-driven method is developed to calibrate the performance of b-tagging algorithms using top quarks.

In Chapter 1 the Standard Model is introduced together with a brief overview of the properties and the importance of the top quark. The method developed in this thesis to calibrate b-tagging algorithms is developed to be applied on proton collisions detected by the CMS detector at the LHC which are introduced in Chapter 2. The LHC is operational only since a few months, therefore the method is studied on simulated proton collisions, which are described in Chapter 3. The reconstruction of the physics objects based on the electronic signals recorded by the CMS detector is described in Chapter 4, emphasizing the b-tagging algorithms and their expected performance. In Chapter 5 the selection and the reconstruction of top quark events is studied. The method to estimate the efficiency of the b-tagging algorithms based on the selected top quark events is introduced in Chapter 6. In Chapter 7 this method is extended towards a differential estimation of the b-tag efficiency as a function of the kinematic properties of the jets. Finally in Chapter 8 the potential and the perspectives of the method are summarized.

Chapter 1

The Standard Model and the Top Quark

The Standard Model of elementary particle physics or shortly the Standard Model gives a theoretical description of the most elementary particles and the interactions between them, except for the gravitational interaction. This theory was developed in the middle of the 20^{th} century and is widely accepted due to its ability to accurately describe a wide variety of experimental results. In Section 1.1 the Standard Model is briefly overviewed and some of its basic concepts are introduced. Although its great success in describing and predicting high energy physics experimental results the Standard Model is assumed to be incomplete. Some of its shortcomings and possible extensions are described in this section.

A key player in this thesis is the top quark. This quark, discovered about 15 years ago, has a special place in the Standard Model since it is the heaviest fundamental particle discovered to date. It plays an important role in the consistency tests of the Standard Model. Therefore section 1.2 is dedicated to the production, the decay and the properties of this quark. Some of its properties have been measured with a good precision over the last years. This allows the experiments at the Large Hadron collider to use top quarks, that will be abundantly produced, as a calibration tool.

1.1 The Standard Model of elementary particles

The Standard Model describes the elementary particles and the three forces operating between them [1–3]. These forces are the electromagnetic force, the weak force and the strong force. In Section 1.1.1 the matter particles, fermions, and the force carrying particles, bosons, are introduced. In the Section 1.1.2 the Standard Model is described in the framework of quantum field theory. In this framework the interactions in the Standard Model emerge from requiring the theory to be invariant under a specific set of symmetry transformations. These symmetry transformations are introduced together with the symmetry breaking mechanism that is expected to be responsible for the mass of the particles. To make concrete predictions based on the Standard Model the notion of renormalization is introduced. In Section 1.1.3 the overview of the Standard Model is finalized by listing some of the shortcomings to be a theory of everything and some extensions towards more complete theories are given.

1.1.1 Fermions and bosons

In the Standard Model it is assumed that all known matter is composed of 12 elementary particles. These 12 matter particles are spin $\hbar/2$ particles and are called fermions. A list of them is given in Table 1.1 together with their electrical charge[1]. Each of these fermions f has an anti-particle $\bar{f}$ which is an exact copy of the particle, i.e. it has the same quantum numbers, but which has an opposite electrical charge. Anti-particles are denoted with a bar over their symbol except for the electron, the muon and the tau particle for which their respective anti-particles, the positron, the anti-muon and the anti-tau particle are denoted by e^+, μ^+ and τ^+ emphasizing their positive charge. Among the fermions two main categories can be distinguished based on the interactions they participate in, namely the leptons and the quarks. Leptons like the electrons, muons and tau particles interact through the electromagnetic force and the weak force while the neutrinos only interact through the weak force. Quarks interact through the electromagnetic and the weak force and as well interact through the strong force which has no impact on the leptons. Furthermore the quarks and leptons can be subdivided in three generations. Each generation of two quarks and two leptons is an identical copy of the first generation except for an increasing mass. All visible, stable matter in the universe is composed of quarks and leptons from the first generation. Protons are composed of two up quarks and one down quark while neutrons are composed of one up quark and two down quarks. Protons and neutrons together with the electrons form atoms which in their turn make up all the known chemical elements.

	electrical charge	1st generation	2nd generation	3rd generation
quarks	$+2/3$ $-1/3$	up u down d	charm c strange s	top t bottom b
leptons	0 -1	electron neutrino ν_e electron e^-	muon neutrino ν_μ muon μ^-	tau neutrino ν_τ tau τ^-

Table 1.1: An overview of the 3 generations of fermions and their charge in the Standard Model of elementary particles.

The strong and electroweak interactions between the fermions are mediated through force carrying particles with spin $1\hbar$. These so-called bosons are listed in Table 1.2 together with their measured mass [4]. The electromagnetic interaction is mediated by the photon γ, the weak interaction is mediated by the $W^{\pm}$ bosons and the Z boson and the strong interaction is mediated by gluons. The gravitational interaction is not included in the Standard Model due to its very low interaction strength compared to the other three interactions.

[1] The electrical charge is given in multiples of the absolute value of the charge of an electron.

A spin $0\hbar$ boson, the Higgs boson, is predicted in the Standard Model and plays a crucial role in the mechanism to introduce the masses of the fermions and bosons but its existence has not been experimentally confirmed.

interaction	force carrier	mass (GeV/ c^2)
electromagnetic	photon γ	0
weak	W^- and W^+	80.398 ± 0.023
weak	Z^0	91.1876 ± 0.0021
strong	gluons g	0

Table 1.2: An overview of the bosons and their measured mass [4] in the Standard Model of elementary particles.

1.1.2 Quantum field theory

Quantum field theory combines two great achievements of physics in the 20^{th} century, namely quantum mechanics and special relativity. In the quantum field theory description of the Standard Model first the fermion fields are introduced and then by requiring the Lagrangian of the model to be gauge invariant under certain local symmetry transformations the interactions between the fermion fields are generated. This section starts with a summary how a general local invariance induces interactions between the fields. In the second part of the section the three symmetry groups defining the Standard Model interactions are introduced.

Gauge symmetries and interactions

Fermions are represented by a Dirac-spinor field ψ. The Dirac Lagrangian of a free fermion field is formally written as

$$\mathcal{L}_{Dirac} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi, \tag{1.1}$$

which describes the quantum mechanical equivalent of the equations of motion.

In the framework of quantum mechanics physically observable quantities depend on $|\psi|^2$, therefore the Dirac Lagrangian is required to be invariant under a local phase transformation

$$\psi' = U\psi = e^{i\epsilon^a(x)\cdot\frac{\tau^a}{2}}\psi, \tag{1.2}$$

with rotation parameters $\epsilon^a(x)$ in an internal space, characterized by the generators τ^a, $a = 1, ..., n$ of a given Lie-group with dimension n. To assure the Dirac Lagrangian to be invariant under a local phase transformation or a so-called local gauge symmetry, the following covariant derivative is introduced

$$\mathcal{D}_{\mu} = \partial_{\mu} - ig\frac{\tau^a}{2}A^a_{\mu}\ , \tag{1.3}$$

where g is the interaction strength or coupling constant associated to the new interacting gauge fields A^a_μ. The Dirac Lagrangian becomes now

$$\begin{aligned}\mathcal{L} &= i\bar{\psi}\gamma^\mu\mathcal{D}_\mu\psi - m\bar{\psi}\psi \\ &= i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi + g\bar{\psi}\gamma^\mu\frac{\tau^a}{2}A^a_\mu\psi. \end{aligned} \tag{1.4}$$

The last term in the equation now expresses the coupling between the new vector fields and the fermion field.

From these observations it is found that requiring a theory to be invariant under a local phase transformation introduces gauge fields that generate the dynamics of the fermion fields. In the case the gauge transformation is represented by an Abelian[2] group only interactions between the fermion fields and the gauge fields are allowed. By imposing the invariance under a gauge transformation represented by a non-Abelian group, couplings among the gauge fields themselves are present.

Gauge symmetries in the Standard Model

To cope with the experimental observations in high energy physics, three gauge symmetries are required to build the Standard Model. They are generated by the following gauge groups

$$G_{SM} = SU(3)_C \otimes SU(2)_L \otimes U(1)_Y \tag{1.5}$$

where the first group describes the strong interaction and the last two describe the unified electroweak interaction.

- **The electroweak interaction:** To describe the electroweak interaction a local gauge invariance under $SU(2)_L \otimes U(1)_Y$ is required [5–7]. The gauge invariance under the Abelean group $U(1)_Y$ introduces a single field B_μ whereas the gauge invariance under the non-Abelean group $SU(2)_L$ introduces three gauge fields W^a_μ, $a = 1, 2, 3$. The covariant derivative to assure the Lagrangian to be invariant under the gauge symmetries is

  $$D_\mu = \partial_\mu - ig\frac{\tau^a}{2}W^a_\mu - ig'\frac{Y}{2}B_\mu \ , \tag{1.6}$$

  where g and g' are the interaction strengths, Y is the hypercharge and the matrices τ^a, $a = 1, 2, 3$ are the Pauli matrices. To explicitly incorporate the parity violating nature of the weak interactions the three boson fields W^i_μ can only couple to left-handed fermions.

  The B_μ and W^a_μ gauge fields do not correspond immediately to the photon, the Z boson and the W bosons. To obtain the physically observed bosons a linear combination of the gauge fields is needed,

  $$\begin{aligned} W^\pm_\mu &= \sqrt{\frac{1}{2}}\left(W^1_\mu \mp iW^2_\mu\right) \\ Z^0_\mu &= W^3_\mu \cos\theta_w - B_\mu \sin\theta_w \\ A_\mu &= W^3_\mu \sin\theta_w + B_\mu \cos\theta_w \ , \end{aligned} \tag{1.7}$$

[2] An Abelian group is defined by commuting generators τ, i.e. $[\tau_i, \tau_j] = 0$.

where θ_w is the Weinberg mixing angle, defined as

$$\tan\theta_w = \frac{g'}{g}. \tag{1.8}$$

Although the electromagnetic force and the weak force are unified in the Standard Model, this symmetry must be broken at low energies since the W bosons and the Z boson have a mass. Adding explicitly a mass term to the Lagrangian breaks the gauge invariance, therefore a mechanism to spontaneously break the symmetry is proposed known as the Higgs mechanism.

- **Quantum chromodynamics:** The theory of quantum chromodynamics (QCD) is described in the framework of quantum field theory by requiring the Lagrangian to be invariant under transformations of the non-Abelian gauge group $SU(3)_C$. This requirement introduces eight gauge fields G^a_μ, $a = 1, ..., 8$ known as gluons. To assure the Lagrangian is invariant under these gauge transformations, the following covariant derivative is required

$$D_\mu = \partial_\mu - ig_s\frac{\lambda^a}{2}G^a_\mu \ , \tag{1.9}$$

where g_s is the interaction strength and λ^a are the Gell-Mann matrices. A colour charge C is only present for quarks which appear as triplets under $SU(3)$ transformations. Gluons carry colour charge as well and interact with themselves due to the non-Abelian character of the symmetry group. Leptons do not interact with gluons as they carry no colour charge.

To account for the experimentally observed CP violation and the processes violating strangeness it is assumed that the eigenstates of quarks for the strong interactions differ slightly from the eigenstates for the weak interactions. The matrix defining the difference between both quantum states is the Cabibbo-Kobayashi-Maskawa (CKM) matrix

$$\begin{pmatrix} d^{weak} \\ s^{weak} \\ b^{weak} \end{pmatrix}_L = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}_L \tag{1.10}$$

The terms of the matrix express the probability, proportional to $|V_{qq'}|$, for one quark q to decay into another quark q' through the weak decay.

Spontaneous symmetry breaking

The gauge fields induced by the local gauge invariance of the Standard Model Lagrangian introduce the dynamics of the fermion fields. This allows the Standard Model as a quantum field theory to make predictions which can be validated by experiments. The local gauge invariance of the theory however forbids fermions and bosons to have a mass because of the presence of an explicit mass term like $-m^2A^\mu A_\mu$ in the Lagrangian breaks the local gauge invariance. A solution to this problem is the introduction of a scalar field which leaves the Lagrangian invariant but breaks the gauge symmetry of the vacuum [8–10].

The simplest way to break the $SU(2)_L \otimes U(1)_Y$ gauge symmetry is by introducing a scalar field ϕ that is a doublet in $SU(2)$ of complex scalar fields ϕ^+ and ϕ^0,

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \tag{1.11}$$

The following gauge invariant term can then be added to the Lagrangian

$$\mathcal{L}_{Higgs} = (\mathcal{D}^\mu \phi)^\dagger \mathcal{D}_\mu \phi - V(\phi) \tag{1.12}$$

$$= (\mathcal{D}^\mu \phi)^\dagger \mathcal{D}_\mu \phi - \mu^2 \left(\phi^\dagger \phi\right) - \lambda \left(\phi^\dagger \phi\right)^2, \tag{1.13}$$

where μ^2 is a mass parameter and $\lambda>0$ is the strength of the field's self interaction. Requiring that $\mu^2>0$ results in a minimum of the potential $V(\phi)$ that is at $\phi=0$. Requiring on the other hand that $\mu^2<0$ results in a minimum that is no longer unique. The potential has now a minimum for

$$\phi^\dagger \phi = \frac{|\mu^2|}{\lambda} = v^2, \tag{1.14}$$

with v the vacuum expectation value. The vacuum can now be arbitrarily chosen as a quantum fluctuation around the vacuum expectation value

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{1.15}$$

where the only remaining real field $h(x)$ is the so-called Higgs boson field. The other three fields from the complex doublet are absorbed by the W bosons and Z boson when acquiring a mass. The mass of the spin $0\hbar$ boson associated to the Higgs boson field has a mass $M_H = \sqrt{2\lambda v^2}$. The explicit calculation of Equation 1.13 for the electroweak Standard Model covariant derivative from Equation 1.6 leads to a mass term for the W bosons and the Z boson equal to

$$m_W = \frac{1}{2} v\, g \qquad\qquad m_Z = \frac{1}{2} v \sqrt{g^2 + g'^2}. \tag{1.16}$$

The masses of the fermions are not generated in a similar way as for the bosons. For the fermions to acquire a mass Yukawa coupling terms need to be added to the Lagrangian. These gauge invariant terms describing the interaction between the fermion fields and the Higgs boson field have the form

$$\mathcal{L}_{Yukawa} = g_{Yukawa} \phi \bar{\psi} \psi, \tag{1.17}$$

where g_{Yukawa} are the Yukawa coupling constants inducing a mass $m_{fermion} = g_{Yukawa} v/\sqrt{2}$. These additional parameters in the Standard Model regulate the strength of the coupling, introducing thus the mass of the fermions as free parameters in the theory.

Renormalization

The Standard Model in the framework of quantum field theory is built from a set of fermion fields representing the elementary particles. Requiring the Standard Model Lagrangian to be

gauge invariant under the $SU(2)_L \otimes U(1)_Y$ symmetry groups for the electroweak interaction and the $SU(3)_C$ symmetry group for the strong interactions introduced the gauge boson fields mediating the interactions between the fermions. The covariant derivative for the Standard Model Lagrangian to be invariant under these gauge transformations combines the two terms 1.6 and 1.9 into

$$D_\mu = \partial_\mu - ig\frac{\tau^a}{2}W_\mu^a - ig'\frac{Y}{2}B_\mu - ig_s\frac{\lambda^a}{2}G_\mu^a \ . \tag{1.18}$$

The mass of the particles in the Standard Model is governed by the spontaneous symmetry breaking mechanism for the bosons and by the introduction of Yukawa coupling terms for the fermions. The total Standard Model Lagrangian can thus be summarized as

$$\mathcal{L}_{SM} = \mathcal{L}_{fermions} + \mathcal{L}_{gauge\ \ bosons} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa}. \tag{1.19}$$

Despite this elegant description the calculation of concrete predictions in the Standard Model is non-trivial and requires arbitrary interventions. To simplify the calculations, a systematic approach is adopted based on Feynman diagrams and rules. To a given process a lowest order Feynman diagram can be assigned depicting the incoming and outgoing particles. When computing the probability of a certain process to occur, quantum mechanical corrections to the lowest order of the process are introduced by adding extra loops and vertices to the Feynman diagrams. To perform the perturbative calculation these higher order diagrams are ordered in increasing number of vertices which leads to a series in increasing coupling constant. When the coupling constant is smaller than unity the calculation can be performed up to a certain order leaving only a small deviation from the all-order result.

A problem however in this procedure is the presence of divergences in the calculations which lead to unphysical predictions. Within the Standard Model these divergences can always be absorbed into unobservable bare parameters by a technique known as renormalization [11]. Renormalization requires the introduction of a renormalization scale to cancel the divergences in the calculations of a given quantity. However the result should not depend on the choice of the renormalization scale. In general this scale is chosen close to the scale of the energy exchanged in the process.

1.1.3 Beyond the Standard Model

Limitations of the Standard Model

The Standard Model has been tested extensively in various experimental environments over the last decades. It is found that in general the experimental observations are in very good agreement with the theoretical predictions [12]. Although these agreements the Standard Model is lacking an explanation for several questions. A short, non exhaustive list of experimental and theoretical shortcomings of the Standard Model is given here.

- In the Standard Model, as it is described in the previous section, neutrinos are not considered to be massive. Experimental measurements however have pointed out that this assumption might not hold and neutrinos have a very small mass [13]. An extension of the Standard Model can be made rather easily to include those observations, leading mainly to a significant increase of the number of free parameters in the Standard Model.

- The Standard Model predicts the existence of a massive spin $0\hbar$ boson, the Higgs boson. Until today no experimental evidence has been found for its existence. It is however possible by combining the measurements within the Standard Model to limit the allowed mass range for the Standard Model Higgs boson. In Section 1.2.2 a short overview of the current limits on the Higgs boson mass is given.

- In cosmological experiments discrepancies have been observed between the visible mass of galaxies and the mass needed to explain the rotational speed of these galaxies [14]. These observations suggest the presence of massive particles that do not interacts through any of the known forces. These particles whose nature is unknown are denoted as dark matter. Non of the Standard Model particles is a good candidate to explain this matter.

- Another open issue is the so-called hierarchy problem. The scale at which the electroweak symmetry is broken is of the order of 10^2 GeV which is many orders of magnitude lower than the Planck scale, the scale of gravity, which is of the order of 10^{19} GeV. Within the Standard Model no new physics is expected beyond the scale of electroweak symmetry breaking up to the Planck scale. Together with a rather light Higgs boson, as expected by indirect measurements, this would require a very precise tuning of parameters to cancel all higher order corrections which is seen upon as unnatural.

- Unification of electricity and magnetism into the theory of electromagnetism and unification of the electromagnetism in its turn with the weak interaction led to the hopes of unifying the strong force with the electroweak interaction. It is preferred that these forces are unified at some very large energy scale which is spontaneously broken at the energy scales where they are observed as distinct forces. Currently the Standard Model does not support such a unification. A step further would be to embed the gravitational interaction, which is absent in the Standard Model, in a unified theory of everything. Although a priori no reason exists for a theory of everything where all observed laws of physics are special cases, it is very desirable.

Extensions of the Standard Model

One of the most promising extensions of the Standard Model is an extension that includes an additional symmetry, namely supersymmetry [15]. Supersymmetry solves the hierarchy problem and offers a candidate for dark matter. Supersymmetry adds to every particle in the Standard Model a superpartner which has the same properties, i.e. mass, charge, etc. but differs by $\hbar/2$ in spin. Fermions thus get a bosonic superpartner while bosons get a fermionic superpartner. However if supersymmetry would be exact these supersymmetric particles should have been discovered already due to their same mass as the Standard Model particles. Therefore it is assumed that supersymmetry is broken at some higher mass scale and the superparticles have a very high mass. The lightest supersymmetric particle is expected to be stable in the minimal supersymmetric extension of the Standard Model and offers thus a good candidate for explaining dark matter. The presence of superpartners for each particle differing by $\hbar/2$ in spin adds additional terms to the radiative corrections to the Higgs boson mass solving the hierarchy problem. Additionally the supersymmetric

extension of the Standard Model unifies the coupling constants at an energy close to the Planck mass scale which is desired for unifying the three forces described in the Standard Model. However for the theory to be consistent the Higgs boson should be accompanied by additional Higgs bosons and corresponding superpartners. This theory is not only very promising it also only works if the Higgs bosons and some of the superpartners have masses below or around the TeV scale which are within the reach of the discovery potential of the current particle collider experiments. Besides the extension towards supersymmetry many other models have been proposed such as models introducing extra space dimensions [16], technicolour [17], etc. which can provide a solution to the hierarchy problem as well.

To give answer to the open questions in the Standard Model and to indicate which extension of the Standard Model represents nature high energy physics experiments are needed. To study physics at high energy scales increasingly powerful particle colliders are being built. Currently the particle collider operating at the highest collision energy is the Large Hadron Collider (LHC) [18], operational since the end of 2009. It is expected that the experiments located near this collider can unravel some of the open issues in the field of elementary particle physics.

1.2 The top quark

The existence of the top quark was experimentally confirmed in 1995 by the DØ [19] and the CDF [20] experiments at the Tevatron collider [21] and is therefore the most recently discovered quark. Its discovery was a great success for the Standard Model which suggested its existence already in 1977 with the discovery of the b quark. The discovery took place nearly 20 years after the discovery of the b quarks mainly due to the very high mass of the top quark. The progress in constraining its mass and finally its discovery was driven by the construction of more powerful particle colliders. Until recently only the Tevatron collider had a high enough center-of-mass energy to create top quarks.

In Section 1.2.1 the production and the decay of top quarks are discussed. The production of top quarks is mainly dominated by pair production but single production of top quarks was experimentally confirmed a few years ago. Due to its very short lifetime the top quark has no time to hadronize but will decay almost exclusively to a b quark and a W boson. Due to its very high mass the top quark plays an important role in the consistency tests of the Standard Model, in Section 1.2.2 this is discussed together with the implications of the top quark mass on the Higgs boson mass. In Section 1.2.3 finally the use of the top quark as a calibration tool in particle detectors is discussed. Given the large production rate expected at the LHC top quarks can be used to calibrate the detectors.

1.2.1 Top quark production and decay

At hadron colliders top quarks can be produced via two mechanisms. The dominant mechanism is the simultaneous production of a top quark and an anti-top quark, a so-called top quark pair, via the strong interaction. Additionally via the electroweak interaction single top or anti-top quarks can be produced as well. In this thesis the focus is put on events where a top quark pair is produced, shortly denoted as $t\bar{t}$ events, therefore the single top quark production is not discussed here.

In Figure 1.1 the leading-order Feynman diagrams representing the production of a $t\bar{t}$ pair through gluon fusion $gg \rightarrow t\bar{t}$ and through quark anti-quark annihilation $q\bar{q} \rightarrow t\bar{t}$ are shown. The cross section quantifying the probability for this event to occur at the LHC is discussed in Section 3.5.

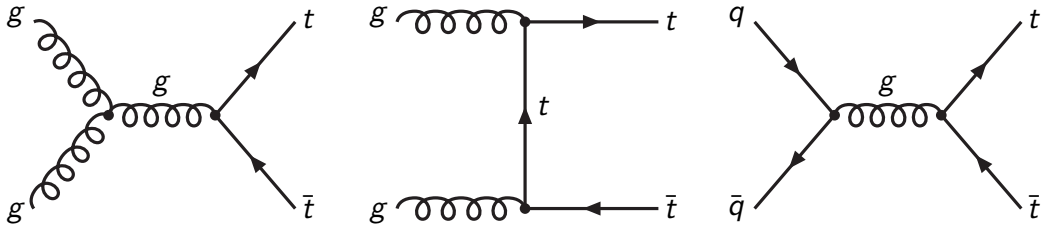

Figure 1.1: Leading order Feynman diagrams for $t\bar{t}$ production through gluon fusion (left and middle) and through quark anti-quark annihilation (right).

The top quark is predicted, due to its very large mass, to have a very short lifetime, of the order of 5.10^{-25}s [22]. This short lifetime is about 20 times smaller than the typical time scale for strong interactions. Therefore the top quark decays weakly before it is able to form hadrons. This property is unique among all the quarks in the Standard Model and offers a way to study the behavior of a bare quark.

The decay of top quarks[3] is mainly occurring via the process

$$t \rightarrow W^+ b \qquad \text{and} \qquad \bar{t} \rightarrow W^- \bar{b}, \tag{1.20}$$

for which the branching ratio is close to 1. Therefore the decay of a top quark to a lighter quark, like $t \rightarrow Ws$ and $t \rightarrow Wd$ is highly suppressed. The CKM matrix element $|V_{tb}|$ quantifying the probability of a top quark to decay to a W boson and a b quark is measured at the Tevatron experiments and is found to be 0.91 ± 0.13 at the CDF experiment and 1.07 ± 0.12 at the DØ experiment [23].

The $t\bar{t}$ events thus mainly decay into two W bosons and two b quarks. The W bosons in their turn can decay hadronically into a quark anti-quark pair, $W \rightarrow q\bar{q}$, with a branching ratio of $2/3$ or leptonically into a lepton and a neutrino, $W \rightarrow \ell\nu_\ell$, with a branching ratio of $1/3$. Since two W bosons are present three decay modes can be distinguished, the hadronic decay where both bosons decay hadronically, the leptonic decay where both W bosons decay leptonically and the semi-leptonic decay where one W boson decays hadronically and where the other W boson decays leptonically. The latter one, in the case where the lepton is a muon, is the relevant decay mode studied in this thesis and is schematically[4] denoted as

$$t\bar{t} \rightarrow \mathrm{WbWb} \rightarrow \mathrm{bqqb}\mu\nu_\mu. \tag{1.21}$$

The probability for this semi-muonic $t\bar{t}$ decay to occur is about 14.8% [4].

[3] When a particle like, e.g. the top quark, is mentioned in this thesis its antiparticle is implicitly assumed as well.

[4] For convenience the explicit distinction between particles and anti-particles is suppressed.

1.2.2 The top quark and the Standard Model

The top quark is the heaviest quark in the Standard Model, the latest measurement of its mass combining the results from several different analyses from the CDF experiment and the DØ experiment [24] results in $m_{top} = 173.3 \pm 1.1$ GeV/c^2. Using an integrated luminosity of 5.6 fb^{-1} at the Tevatron the relative precision on the top quark mass is about 0.6% and mainly dominated by systematic uncertainties. A precise measurement of the top quark mass plays an important role in consistency tests of the Standard Model.

At tree level the mass of the W boson is fully determined by the fine structure constant α, the Fermi coupling constant G_F and the mass of the Z boson m_Z. When including higher order corrections to the W boson mass m_W two main contributions are present, namely from the top quark mass m_t and the Higgs boson mass m_H. Based on the measurements of the fine structure constant, the Fermi coupling constant and the Z boson mass, Figure 1.2 shows the dependency of Higgs boson mass on the W boson mass and the top quark mass [12, 25]. The green band shows a range of possible Higgs boson masses. The direct measurements of the top quark mass and the W boson mass from LEP-2 and from Tevatron is indicated and clearly preferring a rather light Higgs boson. Additionally the top quark mass and W boson mass determined via indirect measurements using LEP-1 and SLD measurements is shown.

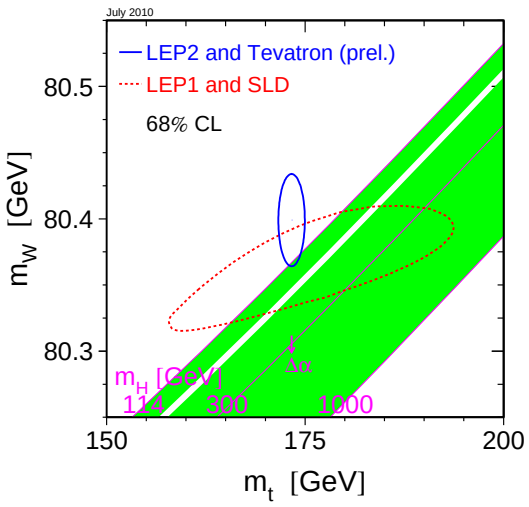

Figure 1.2: The measured top quark mass and W boson mass and the dependency of the Higgs boson mass on the top quark mass and W boson mass.

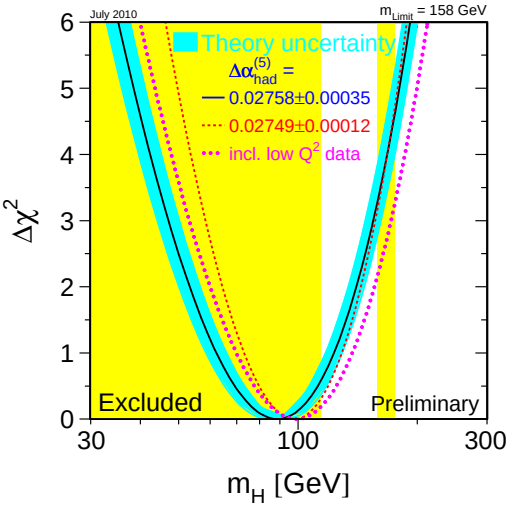

Figure 1.3: The preferred Higgs boson mass from the Standard Model consistency test and the experimentally allowed mass region.

The mass of the top quark and the W boson have been measured with a good precision so that it is possible to perform a χ^2 fit involving these measurements to constrain the Higgs boson mass. In Figure 1.3 the $\Delta\chi^2$ of this fit is shown, where the black line indicates the central value and the blue band indicates the total systematic uncertainty. The precision data depends on the extrapolation from the measured value of the fine structure constant at

low energy up to high energies. The uncertainty is mainly originating from the uncertainty related to low-energy QCD measurements. The red dashed line indicates an alternative central value associated with a different treatment of the low-energy QCD data. The dotted magenta line indicates the effect of the central value after inclusion of NuTeV data which shows discrepancies with other electroweak data. The direct exclusion limits in yellow indicates that the Higgs boson should be more heavy than 114 GeV/ c^2 at 95% confidence level based on LEP results. Recently a Higgs boson mass in the region 158-175 GeV/ c^2 has been excluded at 95% confidence level by the Tevatron experiments.

1.2.3 The top quark as a calibration tool

At the Large Hadron Collider top quarks will be produced at a very high rate resulting in a large sample of top quarks. This large sample offers the opportunity to perform more detailed measurements of the properties of the top quark extending the knowledge obtained at the Tevatron collider. Additionally, because some of the properties of the top quark have been measured already with good precision, the top quark can now be used to calibrate the detector. When reconstructing[5] the properties, such as e.g. the four-momenta of the particles created in the proton collision, the measured quantities can be biased due to various detector effects. Instead of using these observed quantities to measure the properties of the top quark, the properties of the top quark can be used to modify the measured quantities in such a way that they fulfill the top quark properties which now act as a constraint. In this way correction factors can be deduced to calibrate the reconstruction algorithms based on top quark events. Several analyses have been developed over the last years and are ready to be carried out using real detector data.

- **Jet energy scale calibration:** Due to the precise measurement of the top quark mass and the W boson mass in previous experiments these values can be used to calibrate jets [26]. A kinematic fit can be applied in top quark events forcing the jets in the $t \rightarrow bW \rightarrow bqq$ decay to fulfill the W boson mass and the top quark mass requirement. This provides a data-driven estimation of the jet energy scale correction factors for b quark jets and light quark jets simultaneously. These correction factors can then be applied to calibrate the energy scale of the reconstructed jets.

- **b-tag efficiency calibration:** Due to the branching ratio $|V_{tb}|$ of nearly 1 for a top quark to decay into a b quark and a W boson, the top quark offers an ideal environment to calibrate the efficiency of b-tagging algorithms. Previous studies [27] have shown that it is feasible to estimate the efficiency of b-tagging algorithms based on the selection of a jet sample with an enriched b quark jet content. An important aspect in the proposed method in [27] is the control of the contribution of the background, i.e. the non-b quark jets, which is obtained from simulations.

In this thesis a data-driven method is developed to estimate the efficiency of b-tagging algorithms based on the large sample of b quark jets present in top quark events. The

[5] The reconstruction of physics objects, such as jets, muons, etc. in CMS is overviewed in Chapter 4 where concepts like e.g. the jet energy scale calibration and the efficiency of b-tagging algorithms are introduced.

control of the contribution of the background is obtained from a jet sample with a low b quark jet content in combination with a data-driven control sample.

Chapter 2

The Large Hadron Collider and the CMS experiment

In the quest to study particle physics at the smallest possible scale increasingly powerful particle colliders are being built. The design of these accelerators is driven by the need for colliding particles at very high rates at the highest possible energies. The Large Hadron Collider is currently the highest-energy collider in the world.

The analysis in this thesis is developed to be applied on data collected by the Compact Muon Solenoid detector at the Large Hadron Collider. Both are briefly discussed in this chapter. In Section 2.1 an overview of the design of the Large Hadron Collider is given. In Section 2.2 the Compact Muon Solenoid (CMS) detector is discussed. The CMS detector is composed of several subdetectors and each of them is shortly described. The final section of this chapter is dedicated to the online selection and data acquisition system. A short overview is given of the computing infrastructure for processing and storing the data produced by the CMS detector.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [18] is installed in a 26.7 km long quasi circular tunnel in the CERN (European Organization for Nuclear Research) research laboratory near Geneva on the French-Swiss border. The tunnel, excavated in 1985-88, was housing the LEP accelerator from its start in 1989 until its shutdown in 2000. The LHC project was approved in 1994 and is designed to accelerate proton beams, in each direction of the accelerator, up to 7 TeV to generate proton-proton collisions with a center-of-mass energy of 14 TeV. The accelerator is operational since autumn 2009 and after a commissioning phase it is currently, since March 2010, colliding protons at a center-of-mass energy of 7 TeV.

2.1.1 Design of the Large Hadron Collider

The LHC is designed to generate proton collisions a center-of-mass energy and a rate that has never been precendented. Its design center-of-mass energy of 14 TeV is about 7 times higher than the previous highest energy particle accelerator, the Tevatron [21], located near Chicago in the United States of America. The LHC machine luminosity, $\mathcal{L}$, is designed to

be 10^{34} cm^{-2}s^{-1} which is a factor 100 more than the Tevatron. This high luminosity will be achieved by colliding two beams, containing about 10^{11} protons, every 25 ns. The number of events occurring each second is given by the product of the luminosity and the cross section for a given process, $N_{event} = \mathcal{L}\ \sigma_{event}$. With a total proton-proton cross section of roughly 100 mb at a center-of-mass energy of 14 TeV [28], this leads to 10^9 inelastic events per second at design luminosity. At this luminosity on average 22 simultaneous inelastic collisions will occur which implies on the order of 1000 particles emerging from the interaction region every 25 ns putting stringent requirements on the detectors built near the interaction points.

Due to the large number of particles needed in each proton beam to achieve the designed luminosity the LHC collides protons on to protons rather than on to anti-protons. Due to the same charge of the particles in the colliding beams two separate beam rings are needed to keep the proton bunches on their trajectory. Since the tunnel was originally built for the LEP experiment, a particle anti-particle collider, the size of the tunnel is of the order of 4 meters. This limited space puts constraints on the design of the LHC magnets, leading to the choice of a twin-bore magnet design as depicted in Figure 2.1. The dipoles guiding the protons in opposite direction along the beam-line are cooled below 2K to become superconducting to generate a magnetic field of more than 8 Tesla to bend the 7 TeV beams. In total 1232 superconducting magnets are installed and are complemented by over 2500 conventional magnets for the focusing and the cleaning of the beams.

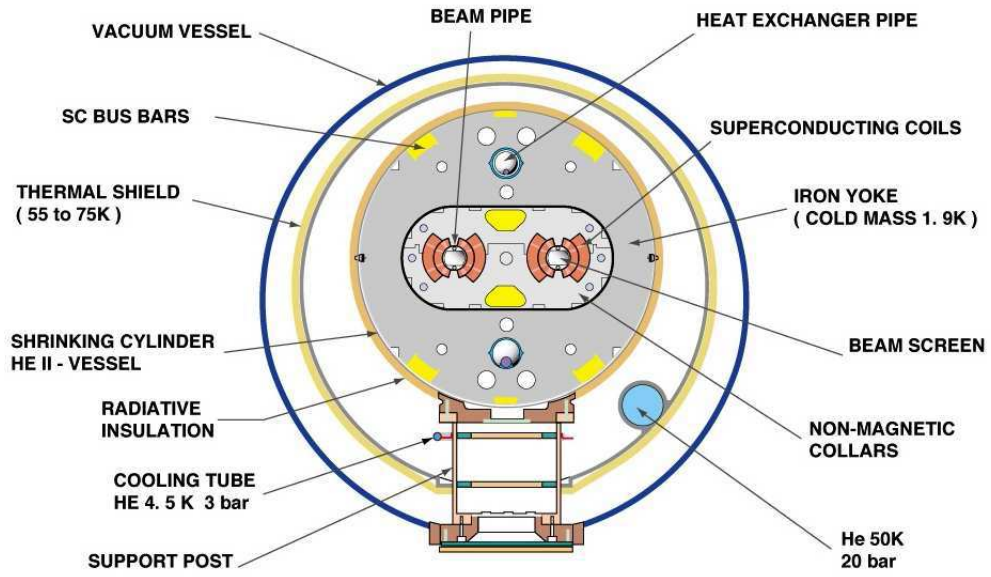

Figure 2.1: A schematic view the cross section of an LHC superconducting dipole.

Before the protons are injected in the LHC they are pre-accelerated in the CERN accelerator complex. This complex consists of several smaller accelerators, schematically depicted in Figure 2.2. The acceleration of the protons starts with a linear accelerator (LINAC) delivering protons to the booster. After acceleration in the booster they are fed into the Proton Synchrotron (PS) which accelerates the protons up to 25 GeV and which delivers bunches of protons with a spacing of 25 ns to the Super Proton Synchrotron (SPS). The SPS then accelerates the proton bunches up to 450 GeV which are then fed into the LHC

where they are accelerated up to 7 TeV. The minimal time needed to pre-accelerate the protons to 450 GeV and inject them into the LHC is about 16 minutes. The time needed to ramp up the beam energy from 450 GeV to 7 TeV is approximately 20 minutes. An additional 20 minutes are needed to ramp down the magnets again to 450 GeV after the beam is dumped. This implies a minimal turn-around time of approximately 70 minutes for the LHC operation.

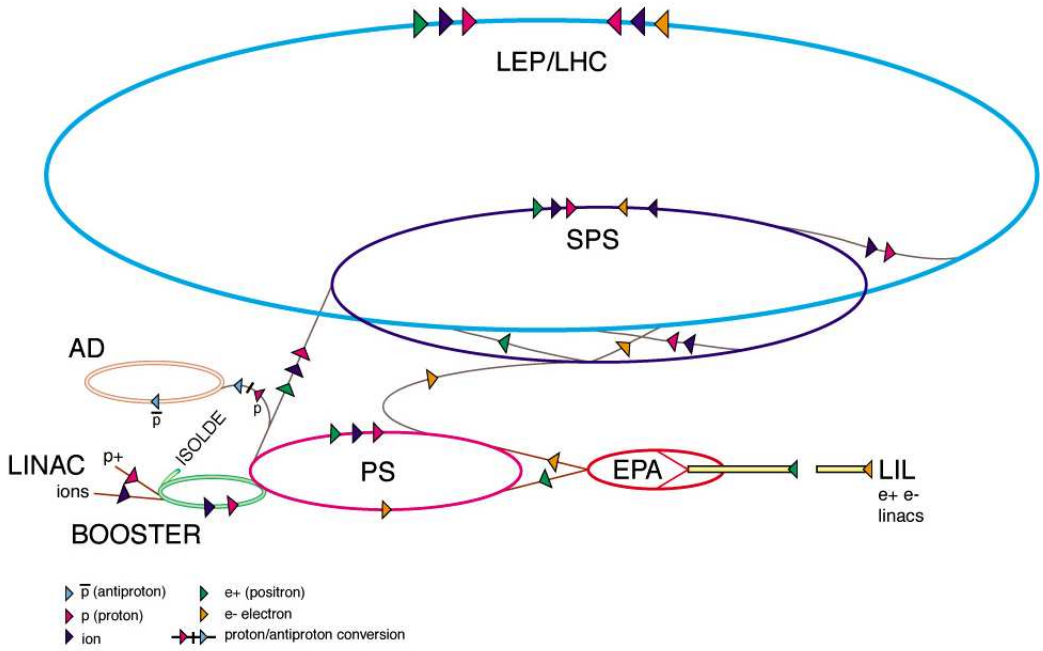

Figure 2.2: A schematic view of the CERN accelerator complex.

2.1.2 Physics motivation and experiments

One of the primary objectives of the physics program of the LHC project is to discover and study the Higgs boson that is expected to be responsible for the mechanism behind the electroweak symmetry breaking. The search for the Higgs boson will be performed by two general-purpose detectors; the CMS [29] and the ATLAS [30] experiments. The design of these experiments is guided by the optimization of the potential to discover the Higgs boson.

The LHC provides, besides the searches beyond the Standard Model, also a very good environment to constrain the Standard Model further and perform precision measurements of Standard Model predictions. Besides CMS and ATLAS other experiments will contribute to dedicated measurements. The LHCb [31] experiment is designed to study physics involving B mesons and will particularly focus on the CP violation phenomenon.The ALICE [32] experiment will study heavy ion collisions to obtain more experimental knowledge on the quark-gluon plasma. The TOTEM [33] detector aims at measuring the total proton-proton cross section and to study elastic proton scattering and diffractive processes. The LHCf [34] experiment will focus on the particles created in the very forward regions of the proton

collisions. By measuring the energy and numbers of neutral pions in the forward direction it will contribute to the understanding of ultra-high energy cosmic rays. In Figure 2.3 a schematic overview is given of the four interaction points and the location of the four main LHC experiments.

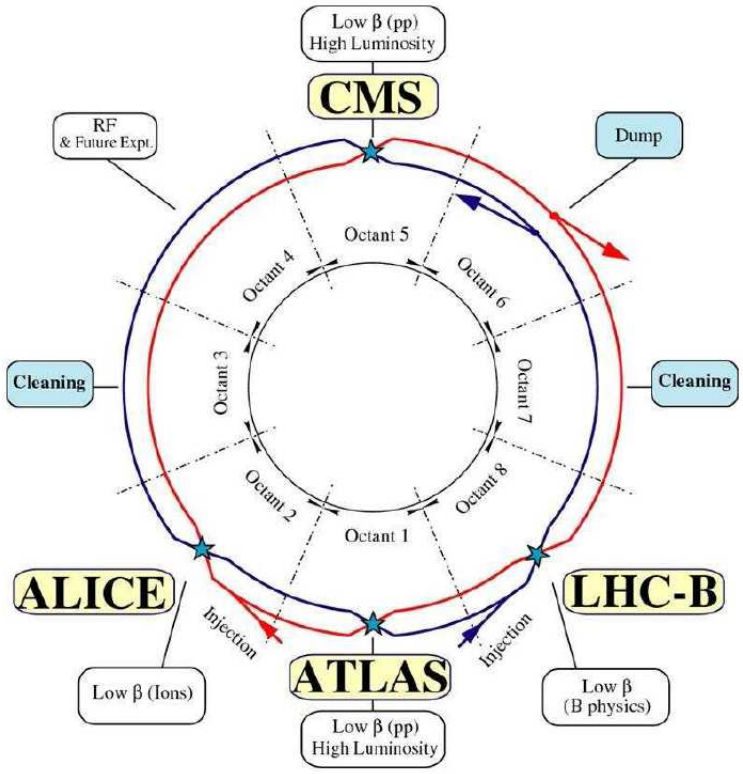

Figure 2.3: A schematic view of the Large Hadron Collider and the location of the four main experiments

2.2 The Compact Muon Solenoid experiment

The Compact Muon Solenoid (CMS) is one of the four large detector facilities operating at the LHC. Together with the ATLAS detector it will search for the Higgs boson and potential signals from physics beyond the Standard Model. The CMS detector consist of several subdetectors each with a dedicated task. In this section an overview is given of the various subdetectors. One of the great challenges for detectors at the LHC is to reduce the large rate of proton collisions, one bunch crossing every 25 ns corresponding to about 10^9 collisions per second, to a manageable rate. This section is finalized with a short introduction to the online selection system dedicated to this task and a brief overview of the computing infrastructure dealing with the offline processing and storage of the data.

2.2.1 General concept of the CMS detector

An overview of the CMS detector is shown in Figure 2.4 where the multi-layered structure, typical for detectors at collider experiments, is visible [35]. The CMS detector is, with its length of 21 m and diameter of 15 m, significantly compacter, hence its name, than the ATLAS detector which has a length of 46 m and a diameter of 25 m. The total weight of the CMS detector is approximately 12500 ton while the ATLAS detector weighs approximately 10000 ton.

One of the primary aspects driving the design of the CMS detector is to obtain a precise measurement of the muon momentum. To achieve this, a strong magnetic field to bend the muons is required together with a precise tracking system and a performant muon system. The strong magnetic field is generated by the superconducting solenoid magnet which has a length of 12.5 m and an inner diameter of 6 m. The magnetic field generated in the superconducting solenoid is about 4 Tesla and is twice as large as the magnetic field in the ATLAS detector. The magnetic field is closed by an iron return yoke surrounding the magnet and supporting the muon detectors at the outer layers of the CMS detector. The weight of the iron return yoke is approximately 10000 ton, dominating the total weight of the CMS detector.

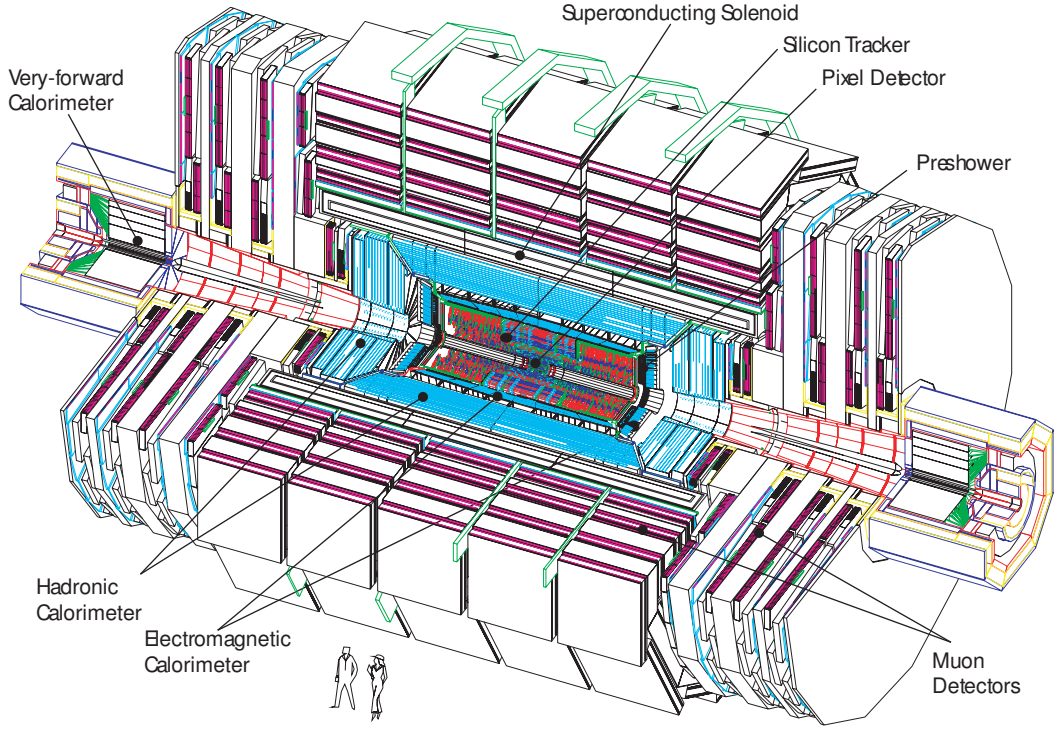

Figure 2.4: An overview of the Compact Muon Solenoid detector layout.

The bore of the magnet coil is large enough to accommodate the tracking system, the electromagnetic calorimeter system and the hadronic calorimeter system. The inner tracking system is consisting of three layers of pixel detectors and ten layers of highly granular silicon strip detectors and is located the closest to the beam-line. Together they provide, besides a precise reconstruction of the charged particle trajectories, a good measurement of the

impact parameters of charged particle tracks as well as the position of secondary vertices, in a very dense track multiplicity environment. Surrounding the inner tracking system, the electromagnetic and the hadronic calorimeter are located. The electromagnetic calorimeter is a homogeneous scintillating crystal detector providing an excellent energy resolution for reconstructed photons and electrons[1] which meets another important design criterion of the CMS detector. The electromagnetic calorimeter is surrounded by the sampling hadronic calorimeter which alternates brass and active scintillating material.

The nominal interaction point at the center of the detector is adopted as the origin of the CMS coordinate system. The y-axis is pointing upwards to the sky, the x-axis is pointing towards the center of the LHC and the z-axis is chosen to yield a right-handed coordinate system. The azimuthal angle ϕ is measured in the (x, y)-plane starting from the x-axis and ranging from 0 to 2π, the radial coordinate in this plane is denoted by r. The polar angle θ is measured from the z-axis ranging from -π to π. The pseudo-rapidity is defined by

$$\eta = -ln\left(tan\frac{\theta}{2}\right), \tag{2.1}$$

which is Lorentz invariant for a boost along the beam direction.

2.2.2 The inner tracking system

Closest to the beam-line, at the center of the CMS detector, the inner tracking system is located. Its main purpose it to reconstruct the trajectories of the charged particles emerging from the proton collisions. The charged particles are bent in the magnetic field allowing a measurement of their charge and momentum. The tracking volume of the silicon tracking system is a cylinder with a length of 5.8 m and a diameter of 2.6 m. The most inner part of the tracking detector is the pixel detector which is surrounded by the silicon strip detectors. The 3 pixel layers, schematically depicted in Figure 2.5, are located at a radius of 4.4 cm, 7.3 cm and 10.2 cm and are complemented by two disks of pixel modules at each side extending from 6 cm to 15 cm in radius and are located at z=$\pm$34.5 cm and z=$\pm$46.5 cm, covering a pseudo-rapidity range of $|\eta|$<2.5. Each pixel cell has a size of 100 $\times$ 150 μm^2 in order to obtain a low occupancy and a spatial resolution per pixel hit in the range of 15-20 μm. The total number of pixels in the CMS detector is 66 millions.

The silicon strip detector consists of a total of 9.3 million silicon strip sensors divided in three different subsystems extending from a radius of 20 cm up to 116 cm and is schematically depicted in Figure 2.6. The Tracker Inner Barrel and Disks (TIB/TID) are composed of four barrel layers and three disks at each end extending to about $\pm$55 cm in the z direction. The TIB/TID is surrounded by the Tracker Outer Barrel (TOB) consisting of six layers and extends in z to $\pm$118 cm. Beyond this range the Tracker EndCaps (TEC$\pm$) are located. They are composed of nine layers, covering the region from $\pm$124 cm up to $\pm$282 cm in the z-direction and a radius reaching from 22.5 cm up to 113.5 cm. The single point resolutions are, for the TIB, 23-35 μm on the r-ϕ measurement and 230 μm on the z measurement. For the TOB the single point resolutions are 35-53 μm on the r-ϕ measurement and 530 μm on z measurement.

[1] Positrons are implicitly considered.

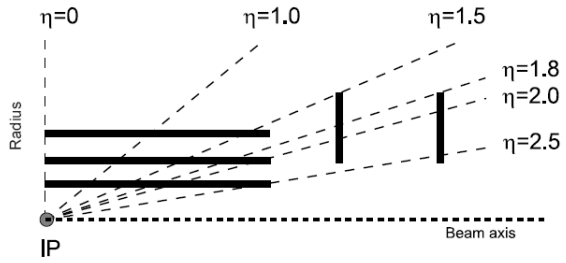

Figure 2.5: Schematic view of a part of the CMS pixel detector.

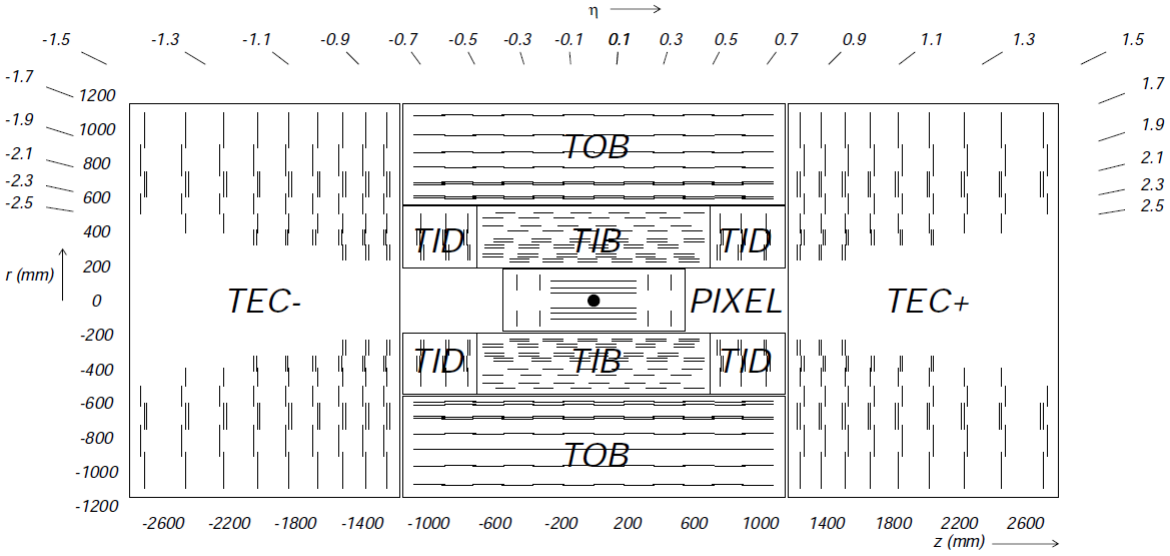

Figure 2.6: Schematic view of the CMS silicon tracking system.

Track reconstruction

Based on the hits reconstructed in the pixel detector and the silicon strip tracker, the reconstruction of tracks is performed in a series of four steps; seed generation, trajectory building, ambiguity resolving and the final track fit [36].

Seed generation: after the reconstruction of the hits in the tracker, track seeds are generated providing initial trajectory candidates for the full track reconstruction. These seeds are composed from at least three hits in the tracker or two hits and a beam constraint. The parameters of the seed provide an initial estimate of the five parameters of the helix track. Due to the low occupancy and the precise two-dimensional position determination the best seeds are provided by using the pixel hits. Additionally building seeds from pixel hits and silicon strip hits further improves the track reconstruction efficiency.

Trajectory Building: the trajectory building starts from the coarse estimate of the track parameters of the seed using a combinatorial Kalman filter method [37]. The trajectory

is extrapolated towards the compatible layers in the tracker based on the equations of motion of a charged particle in a constant magnetic field, taking into account the multiple scattering and the energy loss in the traversed material. Based on this extrapolation one or more compatible hits can be found in the compatible layer. For each compatible hit a trajectory candidate is formed. One additional trajectory is created in which no measured hit is used to take into account the possibility that the track did not leave a hit on that particular layer. Each trajectory is then updated with the additional hit and the procedure of finding compatible hits is continued until the outermost layer of the tracker is reached. To limit the number of trajectory candidates, only tracks are retained which are passing a cut on the normalized χ^2 of the fit and a cut on the number of valid and invalid hits.

Resolving Ambiguities: It is possible that the same track is reconstructed from different seeds or that a given seed leads to multiple trajectory candidates. To resolve these ambiguities, tracks sharing too much hits are discarded, retaining only the track with the most hits. This procedure is applied twice, first for all trajectory candidates from a single seed and once again for all trajectory candidates from all seeds.

Final track fit: At the end of the trajectory building phase, after resolving the possible ambiguities, each trajectory has an estimate of the track parameters. These estimates can be biased due to the constraints applied to obtain the tracks seeds. To avoid a bias, the tracks are refitted with the Kalman filter using all hits associated to the track. This procedure yields the optimal estimates of the track parameters at the surface of each hit, specifically, at the first and the last hit of the trajectory. Estimates on other surfaces, e.g. at the impact point, are derived by extrapolation from the closest hit.

The resolution on the momentum of muons which have a momentum of the order of 100 GeV/c is expected to be 1-2% up to a pseudo-rapidity of $|\eta|<1.6$. For higher transverse momenta the resolution becomes worse. For isolated muons in the range of 1-100 GeV/c a reconstruction efficiency of more than 99% is measured over the full η-range of the tracker acceptance [38].

Vertex reconstruction

After the reconstruction of tracks, the reconstruction of the primary vertex is performed in two steps [35]. The first step, vertex finding, involves the grouping of tracks into vertex candidates. The second step, vertex fitting, determines the best estimate of the vertex position and further constrains the parameters of tracks associated to the vertex by imposing the vertex position as a constraint on the trajectory.

Vertex finding. Primary vertex finding using fully reconstructed tracks provides a first estimation of the vertex position and its covariance matrix as well as a list of tracks associated to the primary vertex. The vertex finding starts with the preselection of tracks. For each track the transverse impact parameter significance is calculated, this value is the ratio of the distance (in the (x, y)-plane) of closest approach of the track to the beam, the so-called transverse impact parameter, divided by the uncertainty on the transverse impact parameter. Based on their transverse impact parameter significance and their transverse momentum p_T, incompatible tracks are discarded. Tracks close in longitudinal impact parameter significance are grouped together. For each of these clusters a fit is executed discarding tracks which are incompatible with the primary vertex. Primary vertices are sorted in decreasing order of p_T^2 of the associated tracks, discarding vertices with a poor fit

quality and vertices incompatible with the beam-line. For the case of $t\bar{t}$-events an efficiency close to 100% is expected for finding the primary vertex assuming low-luminosity pile-up.

Vertex fitting. Vertex fitting algorithms have the purpose, starting from a set of tracks, to compute the best estimates of the vertex parameters as well as the quality (χ^2) of the fit. For this purpose the Kalman fitter can be used. This is the optimal estimator, based on the minimization of a global χ^2, in the case when the uncertainties are Gaussian and there are no mismeasured tracks or tracks from other vertices, so called outliers. A more robust method in the case of outliers, used in the CMS collaboration, is the Adaptive Vertex Fitter [39]. This algorithm adapts the weight of tracks according to their distance to the primary vertex, down-weighting outliers. Optionally the beamspot constraint can be taken into account in the fit procedure resulting in an improved vertex resolution. Besides primary vertices often secondary and tertiary vertices are present in the event. The reconstruction of secondary and tertiary vertices are the subject of Section 4.3.2.

2.2.3 The calorimeter system

Surrounding the tracking system, still within the superconducting solenoid, the calorimeter system is located. It is composed of two parts, the inner part is the electromagnetic calorimeter, consisting of homogeneous scintillating crystals while the outer part, the hadronic calorimeter, consists of a layered structure of scintillating plastic and brass absorber material.

The electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) aims to collect all energy of electromagnetic interacting particles, i.e. charged particles like electrons and neutral particles like photons. Lead tungstate crystals are used for the ECAL since they have a high density and a short radiation length resulting in a fine granularity and a compact design needed to fit within the solenoid. The scintillation decay time of lead tungstate is of the order of the LHC bunch crossing time, 80% of the light is emitted withing 25 ns, making the ECAL fast and thus useable for triggering purposes. The layout of the geometry of the ECAL is shown in Figure 2.7. The barrel section of the ECAL (EB) extends up to a pseudo-rapidity of 1.479 and has an inner radius of 129 cm. The crystals are installed quasi projective, albeit slightly tilted over an angle of 3° to minimize the energy loss of particles traversing exactly on the boundary of two crystals. The crystals in the barrel are 23 cm long and have a front area of 20×20 mm^2 or 0.0174×0.0174 rad^2 in (η, ϕ)-space. A total of 61200 individual crystals are contained in the barrel ECAL. The endcap section of the ECAL (EE) consists of 14648 crystals ranging, in pseudo-rapidity, from 1.479 up to 3 and are located at $\pm$315.4 cm along the z-axis. The crystals in the endcaps have a front surface of 28.62×28.62 mm^2 and are 220 cm long. Additionally a pre-shower is installed in front of the ECAL endcaps for discriminating between photons from neutral pion decays and photons produced in eg. Higgs boson decays. It also improves the position determination of electrons and photons.

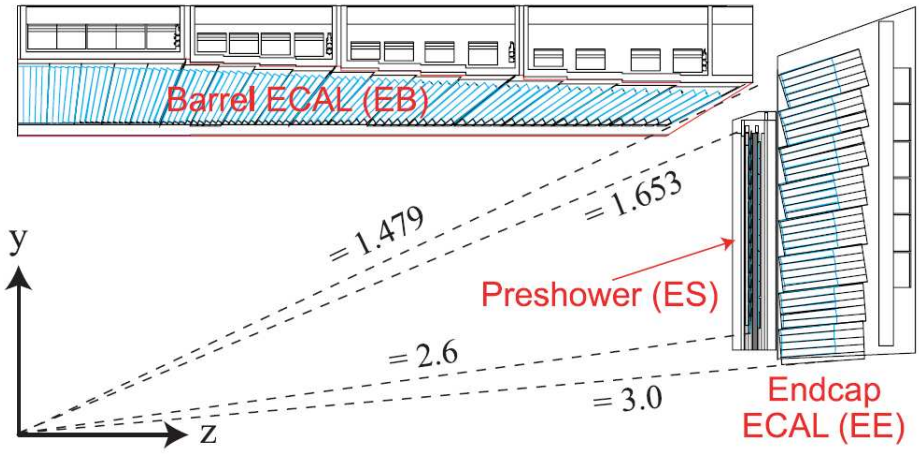

Figure 2.7: Schematic view of a part of the CMS electromagnetic calorimeter.

The hadronic calorimeter

The purpose of the hadronic calorimeter (HCAL), surrounding the ECAL, is to collect the energy of hadronic jets. It plays as well a role in determining the missing energy from neutrinos and potential exotic particles not interacting with the quasi hermetic detector. The HCAL is a sampling calorimeter alternating brass absorber plates with scintillating plastic tiles. Around 70000 of these tiles are installed in the CMS detector. A schematic overview of the HCAL is given in Figure 2.8. The barrel section of the HCAL (HB) is covering a pseudo-rapidity range up to 1.3 and consists of towers of 0.087 $\times$ 0.087 rad^2 in (η, ϕ)-space. The endcap section of the hadronic calorimeter (HE) covers a pseudo-rapidity range from 1.3 up to 3.0 and has a larger granularity. Due to the size limitations of the hadronic calorimeter, it has to fit within the solenoid, but the stopping power of the HB might not contain all energy of the hadrons. Therefore it is complemented by the outer calorimeter (HO). At a pseudo-rapidity up to approximately 5 the forward calorimeter is installed (HF), about 11 m away from the interaction point, to measure the forward hadronic activity. The calorimeter is designed to have an energy resolution of $100\%/\sqrt{E}$ + 5% for an energy measurement E in GeV.

Calorimeter towers

The energy deposits in the electromagnetic and hadronic calorimeters will be the input objects for the jet reconstruction algorithms introduced in Section 4.2. Due to the different granularity of the ECAL and the HCAL, several ECAL towers are merged with one HCAL tower to obtain a calorimeter tower. In the barrel region 5$\times$5 ECAL crystals are merged with one HCAL tower leading to a calorimeter tower with dimension $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ rad^2. In the endcap region a more complex association of ECAL cells to HCAL cells is applied due to the different geometry. The distribution of the total number of towers is shown in Figure 2.9. Energy level thresholds are applied to each individual cell according to the scheme in Table 2.1 to reject calorimeter noise [40]. An additional overall tower threshold of E_T>0.5 GeV is applied to suppress energy contributions from the underlying event.

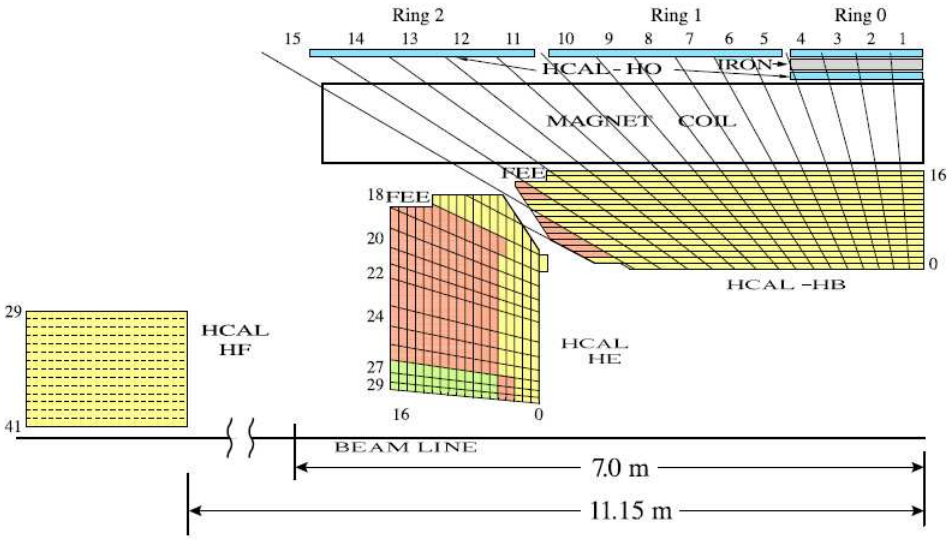

Figure 2.8: Schematic view of a part of the CMS hadronic calorimeter.

HB	HO	HE	$\sum$EB	$\sum$EE
0.90	1.10	1.40	0.20	0.45

Table 2.1: Energy thresholds (in GeV) for calorimeter noise suppression. $\sum$EB and $\sum$EE denote to the sum of the energy deposits associated with a tower, respectively in the barrel and the endcap.

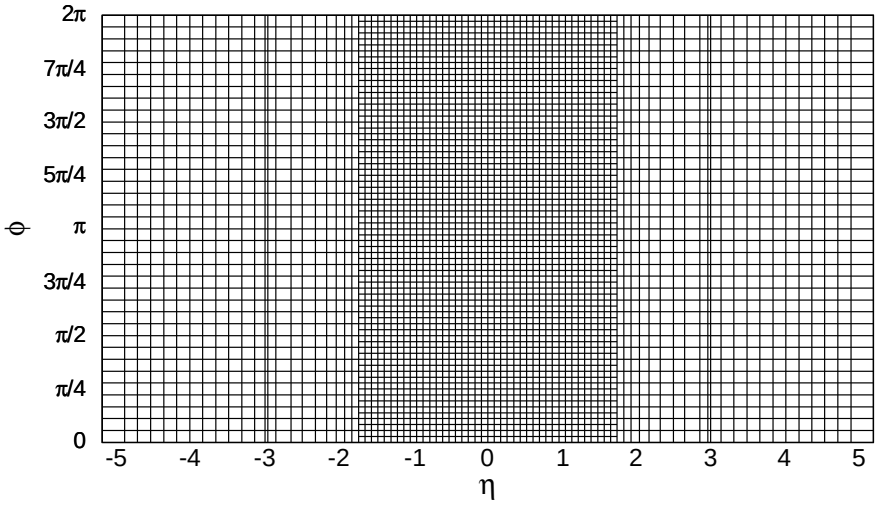

Figure 2.9: Map in η-ϕ of the calorimeter towers.

2.2.4 The muon system

The muon system of the CMS detector is located outside the solenoid in its iron return yoke. It aims at identifying efficiently muons and measuring their transverse momentum with a good resolution. An important design property of the muon system is the fast reconstruction and identification of muons to be used for triggering purposes.

There are three types of gaseous detectors in the muon system as depicted in Figure 2.10. In the barrel region, up to a pseudo-rapidity of 1.2, four layers of Drift Tubes (DT) are interleaved with the layers of the magnetic flux return plates. In the endcap region, for a pseudo-rapidity between 0.9 and 2.4, four layers of Cathode Strip Chambers (CSC) are located. The cathode strips of each chamber run radially outwards to provide a precise measurement in the r-ϕ bending plane. The DT's and CSC's are complemented with Resistive Plate Chambers (RPC) which have a good time resolution, of the order of 1 ns, but a coarser position resolution. They are mainly used for triggering and the measurement of the beam crossing time. The DT provides a spatial resolution of 100 μm for the position measurement and a resolution of 1 mrad on the direction. The typical resolution for the CSC is a spatial resolution of about 200 μm and an angular resolution of the order of 10 mrad.

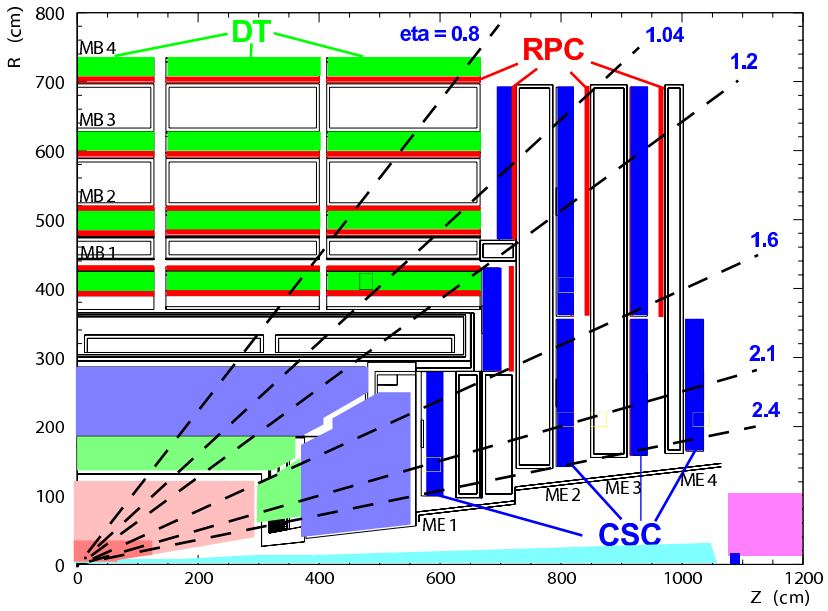

Figure 2.10: Schematic view of a part of the CMS muon system.

2.2.5 The online selection system

Operating at design luminosity the LHC will collide protons at a rate of 40 MHz, creating about 10^9 inelastic proton collisions each second. Given the storage capacity of the CMS experiment, the maximal rate of events that can be stored is around 100 Hz. Therefore a strong and adequate filtering is needed to reduce the number of events with a factor of approximately 10^7 without rejecting potentially interesting events. Due to the short time

between two consecutive collisions, the trigger decisions to accept or reject an event needs to be taken fast, i.e. every 25 ns. The CMS online trigger system performing this task is composed of two physical levels, the level-1 trigger which is implemented on dedicated hardware and the high level trigger which makes use of the offline reconstruction software running on commercial computing units.

Level-1 trigger

The level-1 (L1) trigger will provide the first reduction step, decreasing the event rate to about 100 kHz. The L1 trigger is implemented on dedicated programmable hardware and bases its decision only on calorimeter and muon system information. The trigger hardware is placed in the service cavern located next to the CMS cavern resulting in a time needed to transfer data from the front-end detector electronics to the trigger hardware of about 2 μs. The total time allowed for making a trigger decision is limited to 3.2 μs, limiting therefore the time to reconstruct trigger primitive objects to about 1 μs. Dedicated reconstruction algorithms have been developed to perform this task. Based on the properties of the trigger primitives a decision to keep or reject an event is made every 25 ns, with a delay of 3.2 μs. A pipeline of 128 slots is available to buffer the events during this delay time.

High Level trigger

As soon as the L1 trigger made the decision to keep the data for further analysis the data from the buffer pipelines is transferred to the High Level Trigger (HLT). The HLT is implemented on processor farms located at the ground level in the technical area above the CMS cavern. In the HLT process all available information of the event is gathered and transferred to one computing unit where the offline reconstruction software reconstructs physics objects like muons, jets, electrons, etc. To use the available infrastructure efficiently the HLT aims to reject unwanted data as soon as possible. Therefore several levels of reconstruction are implemented. In the first place only muon and calorimeter information is used while in the second stage the tracker information is included as well. The trigger used in this thesis, as a part of the event selection procedure is described in Section 5.1.1.

2.2.6 The CMS computing environment

The CMS experiment, and by extension all detectors located at the LHC, need very high processing power and storage capacity for the data produced by the detectors. Also for simulated data this infrastructure is required. This detector and simulated data has to be accessible for many physicists geographically distributed all over the world. To meet these specific goals an innovative computing model is adopted by the LHC community [41–43].

Physical structure

The computing resources for the LHC experiments are organized in a tier-like structure distributed worldwide and is called the Worldwide LHC Computing GRID (WLCG). These tier centers are, in general, computer clusters located at research laboratories contributing to the LHC experiments and are based on commercial hardware. The centers are connected

by high throughput networks which meet specific metrics depending on the place in the hierarchy. The functionality of the tier centers can be shortly summarized as follows.

Tier-0: The highest level in the hierarchical structure is the unique Tier-0 center located at CERN. It receives the detector data directly from the CMS online selection system, archives it and performs the first prompt reconstruction. The reconstructed data, together with the raw data, is distributed to the Tier-1 centers. Additional to these activities also high-priority analyses are performed at the CERN Analysis Facility (CAF) integrated in the Tier-0 center.

Tier-1: The next level in the structure consists of 11 Tier-1 centers. From these 11 tier centers 7 are used by CMS, they are located in France, Germany, Italy, Spain, Taiwan, the United Kingdom and the United States of America. One of their main goals is to archive a part of the data produced by the detector and to serve as a backup for the data stored at the Tier-0 center. At least one additional copy of each dataset located at the Tier-0 will be available at a Tier-1. Also resources are dedicated for a few high priority analysis. The Tier-1 centers are used to perform further reconstruction of the detector data. They serve as access-points for the data for the Tier-2 centers. Another important goal is to store the large amount of simulated data produced at Tier-2 centers.

Tier 2: Over 50 Tier-2 centers, on a total of 130 Tier-2 centers, are linked to the CMS Tier-1 centers. These centers are very diverse in size and processing power and their computing resources are used to produce the simulated event samples. Their main aim is however to provide computing resources to the end user physicist to execute analyses and store analyzed results.

Computing resource and data management

The GRID resources are steered by middelware. The services offered by the middleware cope with the very heterogeneous soft- and hardware configurations of the various computing tiers and makes them transparent for the grid user.

The reconstructed and simulated data is distributed over the many tier centers all over the world. Therefore a database, the Dataset Bookkeeping System (DBS), provides a way to locate data and retrieve metadata information, such as the configuration settings, for all available data in the CMS collaboration. The placement and transfer of data is performed with the PhEDEx software which is used to define, execute and monitor the data movement. To be able to transport data efficiently to the tier centers, good quality and performant transfer links need to be available. These links are tested by the Debugging Data Transfers (DDT) task force [44]. This task force is responsible for commissioning new transfer links, perform tests on the existing links, troubleshoot problematic links and document common problems.

In order to create and process simulated data by using the grid infrastructure, two workload management applications are used in the CMS collaboration. For user analyses this is named the CMS Remote Analysis Builder (CRAB). CRAB will take care of several tasks; it will locate the data using the DBS database and it prepares and submits the analysis jobs on the grid. Furthermore it takes care of the monitoring and it handles the output such as the final results and the log files. Analyzed data files can be stored on the storage element of the users affiliated Tier-2 center or can be copied to the user interface. Two possible implementations of a CRAB client exist, a stand-alone and a server client, where

the latter has the advantage of automatic resubmission after job failure and central tracking of errors but adding, as a disadvantage, an extra layer of complexity. For production of simulated events, the Production Agent software (ProdAgent) has been developed. A local ProdAgent instance is run by several teams in order to execute and monitor the progress of production of simulated samples [45, 46]. To be able to run grid-wide analyses the different version of the CMS analysis software (CMSSW) need to be available at all tier centers. This task of executing and monitoring the installation of the CMS software is centrally organized by a few persons divided in two teams serving routinely about 60 sites.

Chapter 3

Event generation and simulation

An accurate simulation of physical quantities and effects, based on theoretical and phenomenological models, is essential to design experiments to validate or falsify these models. It is also needed to define the research strategies and the discovery potentials of the experiments. The aim of this chapter is to overview the tools needed to simulate the proton collisions in the CMS experiment. They are designed and optimized to reproduce as closely as possible the collisions in the real detector. The step by step approach which is deployed, has the advantage that each of the consecutive steps in the simulation chain, going from the modeling of the physics in the collision up to the simulation of the signals expected in the detector, can be studied and tuned independently.

3.1 Event generation and simulation chain

The generation and simulation of events in hadron collider experiments can in general be factorized in several consecutive steps [47, 48]. This chain is illustrated in Figure 3.1, where the steps from the initial colliding protons until the decay of the long lived particles are illustrated.

- The simulation chain starts with the interaction of two protons. In most of the cases this interaction will be soft, resulting in diffraction or elastic scattering of the protons, referred to as *minimum bias* events. The collisions of interest for this thesis however are hard interactions (cf. Chapter 3.2) where two partons from either protons interact and the protons are destroyed in the collision. The simulation of these *hard interactions* by means of matrix element generators is discussed in Section 3.2.1.

- Each of the colliding protons consists out of three valence quarks and many sea quarks and gluons. The distribution of the momenta of these quarks and gluons, so-called partons, are described by the *parton density functions* in Section 3.2.2.

- Both the incoming and outgoing partons will radiate gluons and quarks. Emissions originating from the incoming partons are called Initial State Radiation (ISR), while emissions originating from the final state partons are called Final State Radiation (FSR). The modeling of radiation will be done by the *parton shower* approach in Section 3.3.

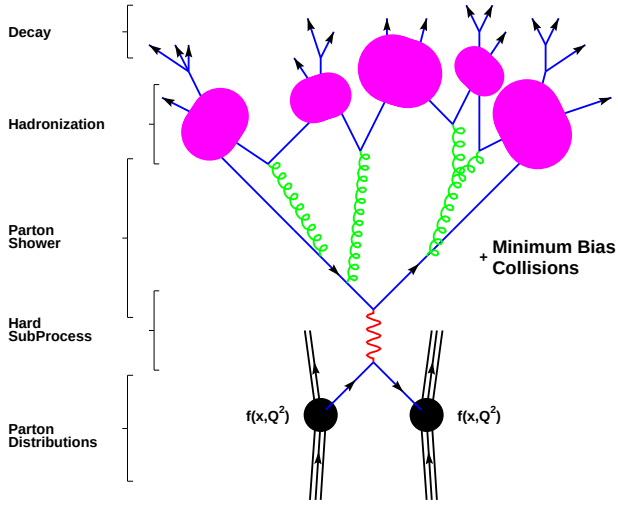

Figure 3.1: Schematic overview of the event generation chain, the time evolution of the event goes from bottom to top.

- Radiation can also be generated during the computation of the hard process by the matrix element generator. This could potentially lead to double counting of events. To avoid this, a matching procedure to eliminate double counted events, is introduced in Section 3.3.3.

- The partons formed in the parton shower process will move further away from each other. After a certain time it is not possible anymore to describe the evolution of the partons further within the frame of perturbative QCD and *hadronization* models need to be introduced. In Section 3.4 one such model is described, the Lund string model.

- Many of the hadrons formed in the hadronization process are unstable and *decay* further. Some of them have a long lifetime, long enough to be detected in the tracking system of CMS.

- Only a small part of the colliding protons take part in the hard interaction. The remnants of these protons will continue traveling in approximately the same direction as their initial direction. Since a fraction of the colour of the proton is taken by the interacting parton these remnants are also coloured and thus radiate and hadronize. The simulation of the so-called *underlying event* is described in Section 3.4.3.

It is only after completing all the previous steps that the simulation of the interaction of the particles with the detector will take place. A fully detailed simulation of the particles passing through each layer of the CMS detector is implemented using the software package `GEANT4` [49]. A parametrized detector simulation is also commonly used and both detector simulation packages are described in Section 3.6.

In the last step of the simulation chain the readout signals produced by the detector through which the particles are passing is packed into a data format which will serve as

the input for further physics analyses. The structure and content of the simulated detector readout is identical to the content of the observed collisions collected during the running of the experiment. The advantage of this similarity is that analyses based on simulated events are immediately applicable on real data. The event format is introduced in Section 3.7 making the connection with the framework for off-line analyses.

3.2 The hard interaction

To calculate the hard scattering process in hadron collisions a factorized approach can be followed [28]. In this approach the top pair production cross section $\sigma^{t\bar{t}}$ in proton collisions is calculated as follows

$$\sigma_{pp\to t\bar{t}}(\sqrt{s}, m_{top}) = \sum_{i,j=q,\bar{q},g} \int dx_i dx_j f_i(x_i, Q^2) f_j(x_j, Q^2) \times \widehat{\sigma}^{ij\to t\bar{t}}(\sqrt{s}, m_{top}, x_i, x_j, Q^2) \quad (3.1)$$

where $f_i(x_i, Q^2)$ and $f_j(x_j, Q^2)$ are the parton distribution functions (pdf's) for the incoming protons. The indices i and j run over the possible $q\bar{q}$ and gg subprocesses leading to a $t\bar{t}$ final state with differential cross section $\widehat{\sigma}_{ij\to t\bar{t}}(\sqrt{s}, m_{top}, x_i, x_j, Q^2)$. The parton distribution functions give the probability to find a parton i with momentum fraction x_i in the proton if it is probed at a scale Q^2. The scale Q^2 of the interaction indicates the momentum transfer in the process.

In the theory of quantum chromodynamics (QCD) in the regime where the coupling constant α_s is significantly smaller than unity the cross section can be calculated perturbative in orders of α_s. The simplest prediction of the top quark pair production cross section is made by only including the leading order Feynman diagrams, shown in Section 1.1.

The lowest order calculation of the $t\bar{t}$ cross section gives a prediction that is significantly underestimating the experimentally observed cross section. For a better accuracy the inclusion of higher order terms is needed. Such higher order calculations of the cross section are technically challenging. Significant progress has been made over the last years and next-to-leading order (NLO) calculations are now standard while next-to-next-to-leading order (NNLO) calculations are becoming available.

Many event generators are on the market, eg. `PYTHIA` [50] and `HERWIG` [51] which are general-purpose event generators of a wide range of processes limited to lowest order. At hadron colliders experimentalists are interested in higher order processes with additional hard partons in the final state. In event generators such as `PYTHIA` and `HERWIG` additional partons are simulated in the parton shower step. A more precise approach is to use matrix element event generators which are capable of generating tree level processes with more final state partons. This procedure gives a better description of the kinematics of the events, since for example in the $t\bar{t}$ final state, processes like $t\bar{t} + 1$ parton, $t\bar{t} + 2$ partons, etc. are explicitly generated from their matrix element rather than from the parton shower. By generating final states with additional partons, real corrections to the leading order Feynman diagram are included, going to higher orders in α_s. However to calculate a complete higher order, virtual or loop corrections need to be included as well. The calculations of these virtual corrections are difficult and results including full next-to-leading order corrections are obtained successfully in event generators like `MC@NLO` [52].

3.2.1 Matrix Element generators

In this section a description of two LO matrix element generators, MadGraph/MadEvent [53] and ALPGEN [54] is given. Both generators compute tree-level matrix elements with a fixed number of final state partons. The main limitation to the maximum number of final state partons is given by restrictions in computing power. Including more and more partons in the final state leads to a rapid growth of the CPU intensive calculations, therefore only a limited number of additional partons can be generated.

The ALPGEN generator

The ALPGEN matrix element generator provides mainly Standard Model processes. Among the available processes, mainly the top quark pair production with additional partons, $t\bar{t} + N$ *partons* (with $N \leq 4$) is of interest for the studies performed in this manuscript. The generation of the desired hard process is performed in a two step procedure.

In the first step the calculation of the cross section takes place in an iterative way to reduce the CPU-heavy calculations. In the first iteration the distribution of the cross section in the kinematic phase space of the process is explored. This is performed on an event by event basis by randomly selecting a parton-subprocess and a point in the phase space. A weight is then computed for each event by integrating the LO matrix element over the phase space. At the end of the first iteration a map of the cross section among the different subprocesses and in the phase space is available. In the subsequent iterations the phase space and subprocesses are randomly sampled. After a fixed, user-defined number of iterations an optimal sampling grid is obtained and serves as input for a final, large-statistics, run to obtain the weighted events. This first step in the matrix element generation technique provides the total cross section and the weighted events. When adding more partons in the final state the number of subprocesses grows rapidly making the computation more CPU intensive. Therefore only a limited amount of final state partons can be included. To limit the file size as well, only the seed of the random number of the weighted event is stored.

The second step is to unweight the weighted events and store them for further processing. Unweighted events are preferred to facilitate the physics analyses. The random selection of a weighted event is based on the maximum weight of the sample and the weight of the respective individual event. When an event is selected it becomes unweighted and the stored random seed is used to construct all information about the momenta, the flavour and the colour flow of the event. The unweighted events are then stored in the Les Houches Event (LHE) format [55] and are ready to be interfaced with parton shower tools.

The MadGraph/MadEvent generator

The MadEvent event generator is a multi-purpose event generator based on the MadGraph matrix element generator. To generate events the desired physics model needs to be specified first. In the current version of the software [56] the option exists to replace the Standard Model by other built-in or user-defined models beyond the Standard Model. Based on the model and process MadGraph will generate the Feynman diagrams and amplitudes for all the relevant subprocesses. The integration of the squared amplitudes over the phase space is performed using Monte Carlo techniques. With respect to other methods MadEvent

uses information from the Feynman diagrams to simplify the integration and speed up the computation. After the integration step the process-dependent information is combined into a stand-alone code allowing the user to calculate the cross section and to generate the unweighted events in the LHE format. The stand-alone code can be transported and processed on an arbitrary computing cluster without the need of additional software which is an advantage in the distributed computing environments used by high energy physics experiments. Limitations to the physics processes calculated by the generator are related to the maximum number of final state particles since computations become very heavy for many-parton final states. For the generation of $t\bar{t} + N$ jets up to $N \leq 3$ additional partons can be explicitly computed.

3.2.2 Parton distribution functions

To calculate the $pp \rightarrow t\bar{t}$ cross section the distribution of the momentum fractions of the partons inside the proton needs to be known. These functions, known as the parton distribution functions (pdf's), cannot be derived from first principles but need to be determined from global fits on data. These fits are carried out by groups like CTEQ [57] and MRST [58] who provide updated fit functions when new data or theoretical improvements become available.

The data used in the global fit is obtained from measurements of deep-inelastic scattering in lepton-hadron scattering and lepton-pair production eg. from the *HERA* experiments. In addition to this data, experimental information obtained in hadron-hadron scattering, eg. from Tevatron experiments [19, 20], is combined in the global fit and mainly constrain the gluon distribution function. The current PDF global fits are carried out at next-to-leading order on more than 2000 data points and a good agreement is found between the data and the fit for both CTEQ and MRST pdf's. Although data is available in a very broad range of momentum fraction x and energy scale Q^2, extrapolations of the pdf's are needed to the kinematic regions accessible at the LHC. The evolution of the pdf's Q^2 is known as the DGLAP equations (cf. Section 3.3). Calculations of this extrapolation show that an accuracy of a few percent is found for very large ranges of x and Q^2. This result is only approximately true and increased uncertainties need to be accounted for in regions where the momentum fraction x is either very large or very small. The parton distributions for CTEQ6.5 [59, 60] are displayed in Figure 3.2 for a Q^2 value of $(350\ \mathrm{GeV}/\mathrm{c}^2\ )^2$ [61], corresponding to the invariant $t\bar{t}$ mass. The uncertainties on the pdf's are calculated using a Hessian technique [62, 63] where a large matrix, with a dimension equal to the number of free parameters in the fit, is diagonalized. In the case of CTEQ this results in 20 orthonormal eigenvector directions, providing the basis for the uncertainty determination for any cross section from the pdf uncertainty. The method presented in this thesis to estimate the b-tag efficiency is based on the kinematics of the final state and is thus not expected to depend on the pdf uncertainty which mainly induces an uncertainty on the top quark production cross section. Therefore these uncertainties are not evaluated.

From Figure 3.2 it is clear that for small momentum fractions x the gluon pdf dominates. This implies that top quark pairs will be produced mainly by gluon fusion at the LHC with a center of mass energy $\sqrt{s}$ ranging from 7 TeV up to 14 TeV. To produce a top quark pair at rest at least enough energy $\widehat{s} = x_i x_j s$, in the interacting parton pair is needed so that

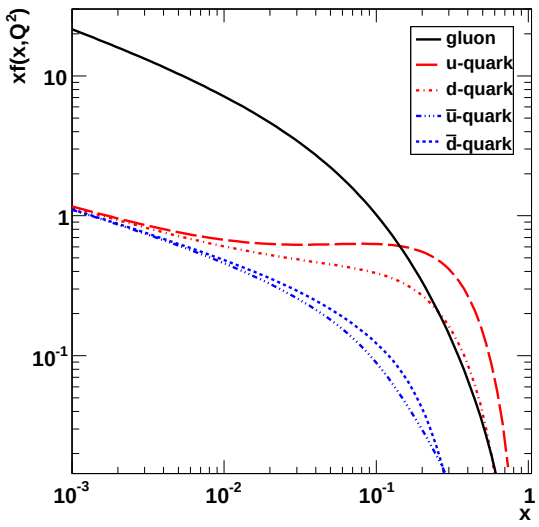

Figure 3.2: Pdf's for the up and down quarks and anti-quarks and gluons from CTEQ6.5 at a Q^2 value of (350 GeV/ c^2)^2.

$\hat{s} \approx 4m_t^2$. Setting $x_i \approx x_j \equiv x$, gives typical values of $x \approx 2m_t/\sqrt{s}$ to estimate the fraction of $t\bar{t}$ production from gluon fusion and quark anti-quark annihilation. Typical values of x and the fractions of top quark pair production with m_t=175 GeV/ c^2 are given in Table 3.1.

$\sqrt{s}$(TeV)	x	$gg \rightarrow t\bar{t}$	$q\bar{q} \rightarrow t\bar{t}$
7	0.05	92%	8%
10	0.035	95%	5%
14	0.025	96%	4%

Table 3.1: Fraction of $pp \rightarrow t\bar{t}$ production through gluon fusion and quark anti-quark annihilation for a top quark mass of m_t=175 GeV/ c^2. The momentum fraction x to produce a top quark pair at rest is given.

3.3 The parton shower

The initial and final state partons can radiate quarks and gluons. Rather than including increasingly higher orders in the perturbative expansion of the matrix element calculation, a parton shower approach is applied to describe this. This showering or branching evolves the partons towards a lower energy scale until $\alpha_s \approx 1$ and the evolution can no longer be described perturbative but hadronization models are needed to further describe the evolution towards observable particles. Three types of radiation can be distinguished, the branching of a gluon in a quark anti-quark pair $g \rightarrow q\bar{q}$, the splitting in a gluon pair $g \rightarrow gg$ and the radiation of a gluon $q \rightarrow qg$. The successive branching is described formally by adding a Sudakov factor [64] in the solutions of the DGLAP evolution equations [65–68]. The

Sudakov factor is needed to handle the cancellation between real and virtual divergences in the Feynman diagrams.

3.3.1 The parton shower approach

The parton shower describes the successive branching of quarks and gluons. Assume now the branching of parton a in a pair of partons b and c, schematically depicted in Figure 3.3.

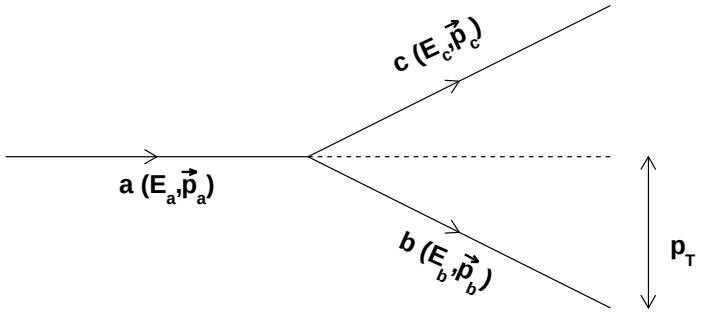

Figure 3.3: Kinematics of parton a branching into two partons b and c

The fraction z of energy carried by parton b with respect to the initial parton a is given by

$$z = \frac{E_b}{E_a} = 1 - \frac{E_c}{E_a}. \tag{3.2}$$

The DGLAP equations describe the splitting probability of parton a into partons b and c,

$$d\mathcal{P}_{a\to bc} = \frac{\alpha_s}{2\pi}\frac{dQ^2}{Q^2}P_{a\to bc}(z)dz, \tag{3.3}$$

where $P_{a\to bc}$ are the splitting functions given by,

$$P_{\mathrm{q}\to\mathrm{qg}} = \frac{4}{3}\frac{1+z^2}{1-z} \tag{3.4}$$

$$P_{\mathrm{g}\to\mathrm{gg}} = 3\frac{(1-z(1-z))^2}{z(1-z)} \tag{3.5}$$

$$P_{\mathrm{g}\to\mathrm{q\bar{q}}} = \frac{n_f}{2}(z^2+(1-z)^2) \tag{3.6}$$

where n_f is the number of quark flavours. From the DGLAP equations it can be seen that a divergence occurs for soft gluon radiation where $z \to 1$ leading to an unphysical cross section. A divergence appears as well for $Q^2 \to 0$ where the radiated parton becomes collinear with the initial parton. To regulate this divergence a cut-off scale Q^2_{min} is introduced

in the calculations. Below this scale, typically of the order of $\mathcal{O}(1\,\mathrm{GeV})$, no more radiation is allowed with the parton shower approach and the non-perturbative regime is entered. Section 3.4 deals with the description of this non-perturbative regime using fragmentation models.

Sudakov form factor

Even though this cut-off scale regulates the divergences it is still possible that the branching probabilities are greater than unity. This is settled by introducing a Sudakov form factor in the DGLAP equations

$$d\mathcal{P}_{a\to bc} = \frac{\alpha_s}{2\pi}\frac{dQ^2}{Q^2}P_{a\to bc}(z)dz\ exp\left(-\sum_{b,c}\int_{Q^2}^{Q^2_{max}}\frac{dQ'^2}{Q'^2}\int\frac{\alpha_s}{2\pi}P_{a\to bc}(z')dz'\right) \tag{3.7}$$

where the exponent is the Sudakov factor. The Sudakov factor can be interpreted as the probability of evolving from initial scale Q^2_{max} to a smaller scale Q^2 without radiation of a parton. The evolution of the partons can now be seen as a cascade of partons. For Final State Radiation starting from an initial scale Q^2_{max} the partons evolve through radiation until a lower cut-off scale is reached. The successive branching is based on a random choice of the branching type and based on the respective probabilities. At each branching, the energy, the momentum and the flavour are conserved from the initial parton a to the branched partons b and c. For the description of Initial State Radiation the situation is more complex since the incoming protons have an internal structure influencing the ISR process. The simulation of the ISR evolution is implemented backwards reconstructing what happened before the hard interaction by using a conditional probability. If a parton b is present at scale Q^2, what is the probability that it would have originated from a branching $a \to bc$ at a smaller scale.

Initial and final state radiation in PYTHIA

For all generated events used in this thesis, the parton shower evolution is carried out by PYTHIA. The values of the relevant parameters and their corresponding uncertainties are based on the proposals in [69]. These recommendations have been cross checked with the parameters adopted by the CDF experiment. The nominal values for the relevant parameters are the following,

- Λ_{QCD}: PARP(61)=0.25 GeV, PARP(72)=0.25 GeV, PARJ(81)=0.29 GeV. The Λ_{QCD} parameter defines the scale of the running coupling constant α_s. PARP(61) regulates the amount initial state radiation where PARP(72) regulates the final state radiation except in the decay of a resonance, then PARP(81) is used.

- Q^2_{max}: PARP(67)=4, the Q^2 scale of the hard scattering is multiplied with this value to define the maximum virtuality of initial state radiation.

- $k^2_\perp$: PARP(64)=0.2, the transverse momentum evolution scale $k^2_\perp = (1-z)Q^2$ is multiplied with this factor for initial state showers.

To study the impact of the uncertainty on ISR/FSR modeling on the method described in this thesis, the values of these parameters are altered in additional $t\bar{t}$ samples. These samples reflect the effect of increased or decreased initial and final state radiation in the top quark production process. The uncertainties in the analysis due to an incomplete knowledge about ISR and FSR are evaluated by comparing the estimated b-tag efficiency between these event samples and the nominal event sample. The event sample with decreased ISR/FSR was generated with a modification to PARP(67)=2.5, while in the sample with increased ISR/FSR the following settings have been used, PARP(61)=0.35 GeV, PARP(72)=0.35 GeV, PARJ(81)=0.35 GeV, PARP(64)=1.0.

3.3.2 Heavy quarks in the parton shower process

The presence of charm and bottom quarks, in this section referred to as heavy quarks, in a parton shower has mainly two sources. On one hand their presence could be due to the production of heavy quarks in the hard process but on the other hand it is also possible that they are produced during the showering process. Gluons radiated in the parton shower have a probability to split in a heavy quark pair if their virtuality Q^2 is larger than twice the mass of the heavy quark pair, $Q^2 > 4m^2$. In this case the process can be calculated perturbative [70]. The mean number of heavy quark pairs, indicated by $R_{q\bar{q}}$, per gluon parton shower, is given by the transition probability times $n_g(E, Q^2)$, where $n_g(E, Q^2)$ is the number of gluons with off-shellness Q^2 in a jet produced at energy scale E,

$$R_{q\bar{q}} = \int_{4m^2}^{E^2} \frac{dQ^2}{Q^2} \frac{\alpha_s(Q^2)}{2\pi} \int_{z_-}^{z_+} \frac{1}{2} \left(z^2 + (1-z)^2 + \frac{2m^2}{Q^2} \right) dz . n_g(E^2, Q^2), \qquad (3.8)$$

where the z-integrand is the splitting function for gluon into a quark anti-quark pair $g \rightarrow q\bar{q}$, generalized to massive quarks. The integration limits are given by $z_\pm = (1 \pm \beta)/2$ with $\beta = \sqrt{(1 - 4m^2/Q^2)}$. The probability of a gluon to branch into a heavy quark pair as a function of the gluon energy is given in Figure 3.4. It is predicted that a gluon with an energy of 50 GeV has a probability of 3% to split into a $b\bar{b}$ pair and a probability of about 8% to split into a $c\bar{c}$ pair. For a gluon with an energy of 100 GeV this increases to probabilities of respectively 4% and 11%. Measurements have been performed of the production rate of respectively charm quarks and bottom quarks in Z boson decays with the OPAL, DELPHI, ALEPH and L3 experiments [71–74].

3.3.3 Matching Matrix Element and parton shower

A leading order matrix element generator is based on a systematic expansion in powers of α_s. In this formal calculation of the process the emission of quarks and gluons in soft and collinear regions leads to diverging probabilities, resulting in unphysical radiation. For this reason matrix element generators need to be interfaced with parton shower programs which perform better in describing the soft radiation of quarks and gluons which occurs when partons reach a lower energy scale. Combining the matrix element generator and the parton shower evolution approach is thus needed to obtain a sensible event description but can lead to double counting. It might happen that the radiation of a hard parton in the parton shower process results in a jet final state generated as well by the matrix element.

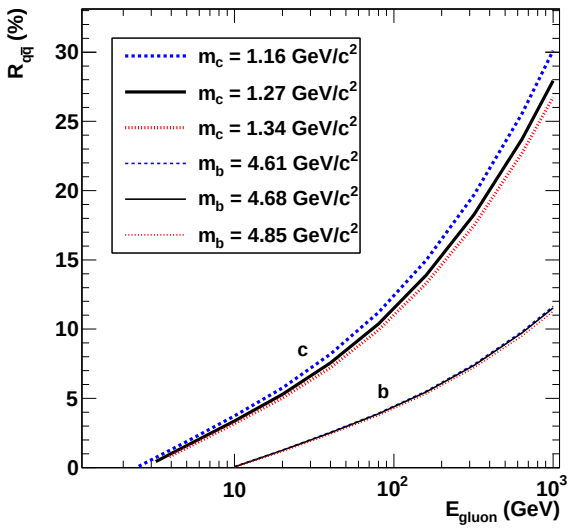

Figure 3.4: Probability $R_{q\bar{q}}$ to split into a charm or bottom quark pair per gluon jet at energy E

It is possible to obtain a $(n+1)$-jets final state in two ways. In the first place the $(n+1)$-parton final state can be explicitly computed in the matrix element. These partons are then interfaced with a showering program and each of them leads to a corresponding cascade of particles, a jet, resulting in a $(n+1)$-jets final state. For the second possibility a n-parton matrix element final state could lead to a $(n+1)$-jet final state when the emission of a sufficiently hard parton leading to an additional jet occurred in the evolution of the parton shower. Two solutions to solve this double counting problem are described here, the MLM matching scheme and the MC@NLO approach. The MLM matching scheme is applied in the samples used in this thesis.

The MLM matching scheme

The idea of the MLM matching scheme [75, 76] is to veto shower evolutions leading to multi-parton final states already described by the matrix element computation. The scheme can be summarized as follows:

- The final state partons generated in the matrix element computation are constrained by acceptance cuts,

$$p_T^{part} > p_T^{min},\ |\eta^{part}| < \eta^{max},\ \Delta R > \Delta R_{min}, \tag{3.9}$$

  where p_T^{part} and η^{part} are the transverse momentum and the pseudo-rapidity of the partons and ΔR is their minimal angular distance in the (η, ϕ) space. The exclusive n-parton sample is defined as the collection of events where exactly n partons pass the acceptance cuts.

- The partons produced in the parton shower process are clustered using a generic clustering algorithm with a jet cone size R^{clus}. The final clusters are called jets and are accepted if their transverse energy exceeds a minimum energy E_T^{clus}.

- Each parton from the matrix element is then matched to a clustered jet in a uniquely defined way:
  - Starting from the highest p_T parton the closest jet in ΔR is looked for. If $\Delta R < \Delta R_{match}$, the matching radius, the parton is matched to the jet.
  - The jet is removed from the list of available jets to prevent a jet to be matched to two partons.
  - The next p_T-ordered parton is selected and the iteration continues until all partons are matched.
- Only events where all partons match the present jets are kept in the sample. Defining the n-jet exclusive samples.
- In the ME computation, only up to N partons are generated due to computational limitations, constraining n to $n \leq N$. Therefore, in the case $n = N$, events are kept where all jets are matched but softer clusters are present. This will define the N-jet inclusive sample.

As a last step all exclusive event samples and the N-jet inclusive sample can be combined into a fully inclusive sample. The MLM matching scheme takes care of the proper merging of the LO matrix element calculations with the parton shower approach. This is shown in [75] where the addition of several exclusive samples is found to yield smooth distributions for $t\bar{t}$ related variables. Going beyond leading order and applying a matching to next-to-leading order is a more ambitious and complex approach and is pursued in the MC@NLO generator.

The MC@NLO approach

The aim of the matching scheme in the MC@NLO generator [52] is to get not only the real but also the virtual corrections correctly included up to next-to-leading order (NLO). In the MC@NLO approach the hard emissions are generated up to NLO in the matrix element while soft and collinear emissions are handled by the parton shower. The MC@NLO scheme works as follows:

- The matrix element of a n-parton process is calculated up to NLO, including $(n+1)$-parton real corrections and n-parton virtual corrections.
- Then it is calculated analytically how a first branching in the showering process starting from a n-parton topology would populate the $(n+1)$-parton phase space.
- From the $(n+1)$ matrix element the shower expression is subtracted to obtain the 'true' $(n+1)$ events. The rest is considered as belonging to the n-parton events. The parton shower and matrix element overlap in the soft and collinear regions, so the singularities cancel, leaving finite cross sections for the n- and $(n+1)$-partons events.
- Now both n- and $(n+1)$-parton events are interfaced with the parton shower.

3.3.4 Comparison of event generators for top quark physics

In this section a comparison is made between $t\bar{t}$ event samples generated with the PYTHIA, ALPGEN and MadGraph event generator. The latter two explicitly generate extra partons in the matrix element calculations whereas in the former all additional jets are formed in the parton shower process. In ALPGEN and MadGraph the combination of the matrix element calculation and the parton shower approach leads to double counting. To avoid this the MLM matching scheme is applied with matching thresholds $p_T^{min} = 30$ GeV/c, $\eta_{max} = 5$ and ΔR =0.7. The $t\bar{t}$ event samples are generated with different values for the top quark mass, in MadGraph a top quark mass of 170.9 GeV/c^2 was used, while in PYTHIA a top quark mass of 172.4 GeV/c^2 and in ALPGEN a top quark mass of 175 GeV/c^2 is used.

In Figure 3.5 the comparison is shown between some kinematic properties of b quarks produced in $t\bar{t}$ events, only $t\bar{t}$ events are accepted where one of the W bosons decays as $t \rightarrow bW \rightarrow b\mu\nu$, so-called leptonic decaying top, and the other decays as $t \rightarrow bW \rightarrow bq\bar{q}$, so-called hadronic decaying top. In the first row the transverse momentum p_T^b and the pseudo-rapidity η^b of the bottom quarks from the hadronic and leptonic decaying top quarks are shown. For these kinematic properties there is an overall good agreement between the different generators. On the second row the angle in (θ, ϕ)-space between the muon and the bottom quark from the leptonic decaying top quark, $\Delta\Omega^{\mu b}$, as well as the mass $m_{\mu b}$ of this system are shown. In general there are no discrepancies between the spectra indicating that the kinematics of a system of two final state particles in the $t\bar{t}$ events is in agreement among the different event generators. The mass $m_{\mu b}$ shifts towards higher values for PYTHIA and ALPGEN compared to MadGraph reflecting the higher top quark mass used in the generation of the events. The plots in the third row show the transverse momentum $p_T^{t\bar{t}}$ of the top quark pair and the angle in (θ, ϕ)-space between the top quark and the associated bottom quark, $\Delta\Omega^{tb}$. In the transverse momentum a discrepancy is seen between MadGraph and ALPGEN on one hand and PYTHIA on the other hand. This is due to the different handling of the matrix element generation between PYTHIA and MadGraph/ALPGEN. In PYTHIA only 2→2 matrix element generation is performed while in MadGraph and ALPGEN additional final state particles can be generated. These additional particles absorb part of the transverse momentum of the top quark pair, resulting in a softer transverse momentum spectrum for top quark pairs generated with MadGraph and ALPGEN. Based on the conclusion that rather similar kinematics are found for all three generators it is motivated to use the MadGraph event generator for generating the $t\bar{t}$ events used in this thesis.

3.4 Hadronization and underlying event

The parton shower process is stopped when the momentum of the partons reach a scale of the order of 1 GeV. At this point the perturbative approach breaks down and the non-perturbative regime is entered. Practically this means that the coloured partons in the shower are transformed into colourless hadrons. This process is called hadronization and cannot be calculated from first principles but is described by phenomenological models. The initial step in the hadronization process described here is the fragmentation of the partons formed in the parton shower. In PYTHIA the Lund string model [77] is the default fragmentation model and is used for all samples in this thesis. After the fragmentation

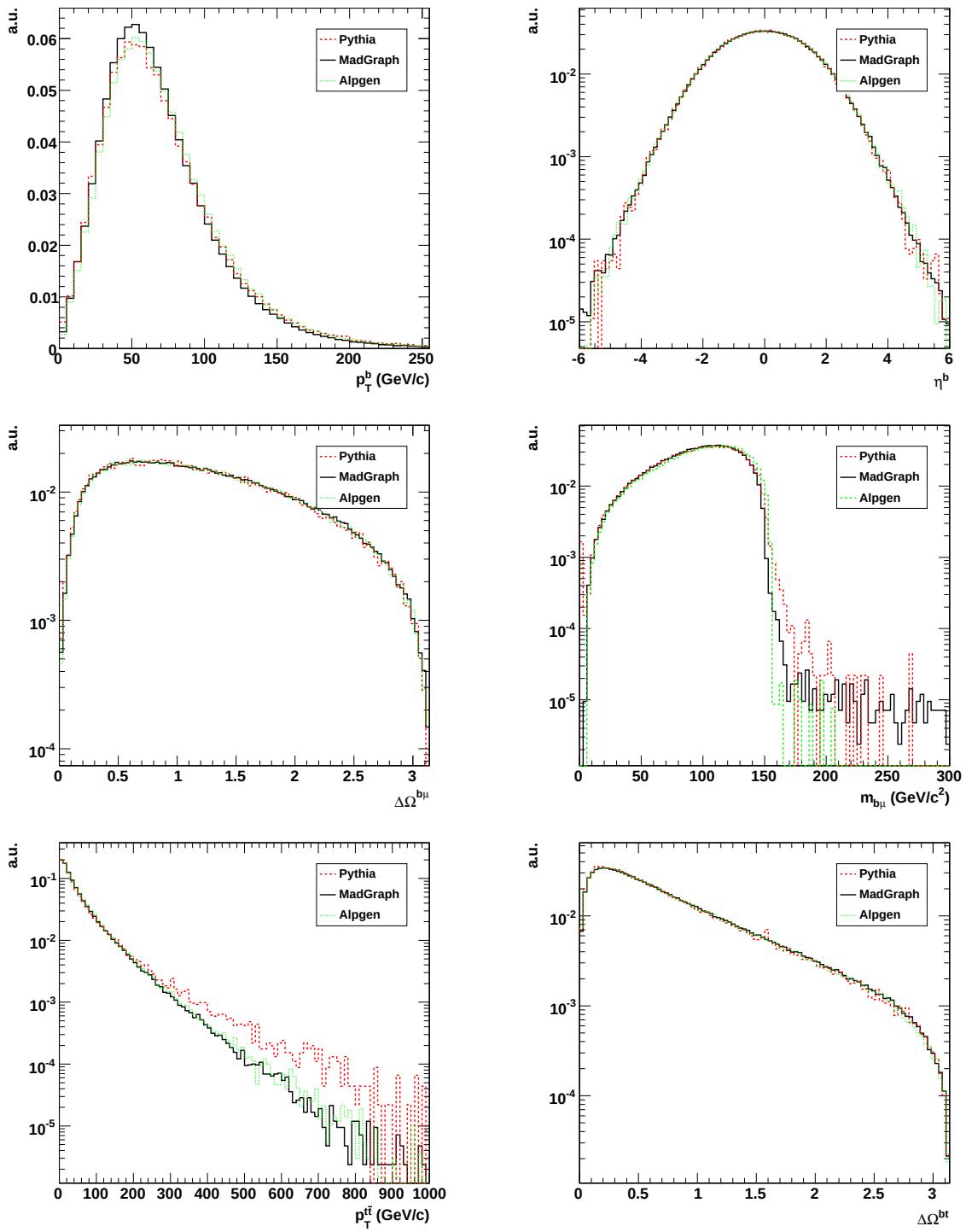

Figure 3.5: The kinematic variables of the bottom quarks, the mass and the angle between the muon and the bottom quark, the angle between the bottom quark and the top quark and the transverse momentum of the $t\bar{t}$ pair produced with the ALPGEN event generator, the PYTHIA event generator and the MadGraph event generator.

step, partons are grouped together to form the colourless particles like mesons and baryons. Some of these particles are unstable and the final step is their decay into the observed final state particles.

3.4.1 The Lund string model

In fragmentation models, like the Lund string model [50], it is assumed that consecutively new $q'\bar{q}'$ pairs are formed from an initial $q\bar{q}$ pair. In the Lund string model this is based on the assumption of linear confinement in QCD. It is assumed that the energy stored in the field between both partons in the $q\bar{q}$ pair increases linearly when the colour-charged partons q and $\bar{q}$ separate. A physical picture is that of a string between the two partons with a string constant, $\kappa \approx 1$ GeV/ fm, corresponding to the energy per unit length. As the q and $\bar{q}$ move further away from each other the potential energy in the string increases and it may break up giving rise to a new $q'\bar{q}'$ string. If the invariant mass of either string pieces is large enough further string breaking might occur. The break-up process is continued until only on-mass-shell hadrons remain. The probability to generate a $q\bar{q}$ pair with transverse mass m_T and transverse momentum p_T is described by the Schwinger mechanism based on quantum tunneling [78] and is proportional to

$$exp(\frac{-\pi m_T^2}{\kappa}) = exp(\frac{-\pi m^2}{\kappa})exp(\frac{-\pi p_T^2}{\kappa}) \tag{3.10}$$

The factorization of the transverse momentum and the mass leads to a flavour independent Gaussian spectrum for the p_T spectrum of the $q\bar{q}$ pair. It is found in experiments that the average transverse momentum of the particles is slightly higher than predicted in this model. This is understood as coming from unresolved soft gluon radiation. The mass term in the expression also implies a suppression of heavy quarks produced in the string break-up. The ratio of different quark flavours is $u : d : s : c \approx 1 : 1 : 0.3 : 10^{-11}$. Therefore charm and more heavy quarks like bottom and top quarks are not formed in the fragmentation process. From these final quarks and anti-quarks after fragmentation a random choice is made to form the mesons and baryons reflecting the abundances observed in data.

The longitudinal momenta of the hadrons are determined from the symmetric Lund fragmentation function

$$f(z) \propto \frac{(1-z)^a}{z} exp(\frac{-bm_T^2}{z}) \tag{3.11}$$

which expresses the probability to generate a given momentum fraction z taken by the hadron from the original $q\bar{q}$ pair. The parameters in PYTHIA are set to a=PARJ(41)=0.3 and b=PARJ(42)=0.58 GeV$^{-2}/c^2$ while the width of the transverse momentum p_T distribution of the hadrons is set to PARJ(21)=0.36 GeV/ c. This fragmentation function agrees well with experimental results for up, down and strange quarks. For charm and bottom quarks it is however found in experiments that a harder fragmentation function is needed. Based on the energy transfer in the break-up process of a fast moving heavy quark in a heavy flavoured meson and a light quark [79], the Peterson/SLAC function

$$f(z) \propto \frac{1}{z(1-\frac{1}{z}-\frac{\epsilon_q}{1-z})^2}, \tag{3.12}$$

is derived providing the best known fragmentation function for heavy quarks. The parameter ϵ_q for bottom and charm quarks is a free parameter in this model and is expected to scale like $\epsilon_q \propto 1/m_q^2$. The parameters in PYTHIA have been set to ϵ_c=PARJ(54)=-0.05 and ϵ_b=PARJ(55)=-0.005 [69].

3.4.2 Decay of hadrons with focus on bottom quarks

In the hadronization process a fraction of the formed particles is unstable and need to be decayed. For the samples in this thesis the decay is conducted by PYTHIA and, although a part of the extensive list of involved parameters, like decay widths and mass distributions of the unstable particles, are poorly known the final physics outcome of the simulation is not altered much when changing these parameters within their uncertainties.

In the fragmentation of a bottom quark[1] typically the following four B-hadrons are formed, the $B^-(\bar{u}b)$-meson, the $\bar{B}^0(\bar{d}b)$-meson, the $\bar{B}_s^0(\bar{s}b)$-meson and the B-baryons ($\Lambda_b^0(udb), \Xi_b^0(usb), \Xi_b^-(dsb)$). An overview of their measured branching ratios Γ_i/Γ and the corresponding mean lifetime τ is given in Table 3.2.

B-hadrons	fraction (Γ_i/Γ)	mean lifetime τ
$B^-(\bar{u}b)$	$(39.9 \pm 1.1)\%$	$(1.638 \pm 0.011) \times 10^{-12}$ s
$\bar{B}^0(\bar{d}b)$	$(39.9 \pm 1.1)\%$	$(1.530 \pm 0.009) \times 10^{-12}$ s
$\bar{B}_s^0(\bar{s}b)$	$(11.0 \pm 1.2)\%$	$(1.470^{+0.026}_{-0.027}) \times 10^{-12}$ s
B-baryons	$(9.2 \pm 1.9)\%$	$(1.208 \pm 0.051) \times 10^{-12}$ s

Table 3.2: The main decay modes of a bottom quark [4]

The three B-mesons have a very similar lifetime τ, of the order of 1.5 pico-seconds. This substantial long lifetime corresponds to $c\tau = 450\ \mu m$, nearly half a millimeter. Due to the high mass of the B-hadrons their decay products tend to have a high transverse momentum giving rise to a displaced vertex not compatible with the primary vertex of the hard interaction. The similar lifetime justifies the use of the spectator model to perform the B-hadron decays in PYTHIA. The advantage of the spectator model is that all B-hadron decays can be treated in the same way. The additional quark besides the bottom quark, the so-called spectator, is not taking part in the bottom decay. The only role of the spectator quark is to define the flavour content of the decayed particles. The bottom quark in the B hadron will mainly decay in a W boson and a charm quark since $|V_{cb}| \gg |V_{ub}|$. This charm quark will create a D-hadron like the $D^+(\bar{d}c)$-, $D^0(\bar{u}c)$- and $D_s^+(\bar{s}c)$-mesons and the Λ_c^0-baryon which have a lifetime similar to the B hadrons but a significant lower momentum. These decays can initiate an extra displaced vertex with respect to the primary interaction and secondary B-hadron decay vertex. Due to the lower momentum of the D-hadron the effective flight distance of the D-hadron before decaying will be on average lower than for the flight distance of the B-hadron.

Another property of bottom quarks, interesting for the experimental determination of the flavour of a quark is the leptonic decay of bottom and charm quarks. The decay of a bottom quark, through an intermediate W-boson, to a lepton $b \rightarrow \ell$ has a branching

[1] The properties of the anti-bottom quark are equivalent.

ratio of about 10% for each lepton flavour. The branching ratio of the bottom quark decay through an intermediate charm quark, $b \to c \to \ell$, has a branching ratio of 10% for each parton flavour. The leptons that are produced in this mechanism are non-isolated inside the parton shower. Their energy and relative momentum with respect to the direction of the parton can be used to distinguish them from other sources of leptons in parton showers which are mainly coming from inflight decays of charged π's and K's, Dalitz decays of π^0's, from γ conversions and mis-identified leptons. The properties of bottom quark jets discussed in this section are crucial for b-flavour identification algorithms (cf. Section 4.3) exploited in the CMS collaboration.

3.4.3 Underlying event

In the hard scattering process only a fraction of the incoming protons is involved, leaving behind a coloured beam remnant which will hadronize as well. Also, due to the composite nature of the proton, additional hard and semi-hard interactions might occur in the proton-proton collision, this is known as multiple interactions. The combination of the multiple interactions and the beam remnant is referred to as the underlying event. In the high luminosity phase of the LHC it is possible to have several collisions between protons in one and the same beam crossing leading to pile-up events. These pile-up collisions are not simulated in this thesis and not further considered here.

The underlying event cannot be derived from first principles and phenomenological models, which are tuned to experimental results, are needed to describe this. In PYTHIA the evolution of the beam remnant is described by a few components. Given that the proton is a colour-singlet particle, the flavour and colour of the beam remnant is reconstructed from the flavour and colour of the initiator particle involved in the hard interaction. Due to total energy and momentum conservation energy and momentum of the beam remnant is determined by the primordial transverse momentum $k_\perp$ of the initiator parton. In PYTHIA the primordial transverse momentum $k_\perp$ is assumed Gaussian, the width of the Gaussian is PARP(91) $= 2.1$ GeV/c. An upper cut-off for $k_\perp$ is applied and set to PARP(93) $= 15$ GeV/c.

The rate of multiple interactions, as a function of the transverse momentum scale $p_\perp$ of the scattering is assumed to be described in the frame of perturbative QCD. This is certainly true for reasonably large $p_\perp$ values, but the extension towards low $p_\perp$ regions leads to particular difficulties. For $p_\perp \to 0$ the cross section diverges and a cut-off scale $p_{\perp,min}$ and a regularization scale $p_{\perp,0}$ need to be introduced. The parameter $p_{\perp,min}$ is explicitly dependent on the center-of-mass energy and thus needs to be extrapolated from Tevatron to LHC energies. Here the PYTHIA parameters have been tuned to a cut-off scale $p_{\perp,min}$ = PARP(81) $= 1.9$ GeV/c and regularization scale $p_{\perp,0}$ = PARP(82) $= 1.838$ GeV/c. The tuning of the parameters for multiple interactions and underlying event is the D6T tune obtained with data from the Tevatron and the UA6 experiment [50]. To ultimately tune these parameters for processes at the LHC tunings need to be performed with data collected at the LHC experiments. The first study of the transverse momentum and pseudo-rapidity distribution of charged hadrons at center-of-mass energy of 0.9 TeV [80] shows a comparison between the data and the current phenomenological models. From the comparison between data and the used models it is concluded that better tunings are

needed to describe the transverse momentum and pseudo-rapidity of charged hadrons more accurately to comply with the observations in the collision. While this work is in progress the current tuning is used for the simulations of the sample in this thesis.

3.5 Cross section of $t\bar{t}$ pair production

The prediction of the top quark production cross section at the Tevatron and the LHC has received large attention over the past years. Currently the cross section is fully calculated at NLO [81] at center-of-mass energies of 10 TeV and 14 TeV. Several attempts are made to calculate the cross section at next-to-next-to-leading order (NNLO). An approach where this is done by adding soft gluon contributions to the NLO calculation is pursued in [82]. For the top quark event samples used in this thesis NLO cross sections are assumed. The cross sections are calculated for a top quark mass of $m_t = 171$ GeV/c^2 and are compared here using two different pdf sets, CTEQ6.5 [60] and MRST2006 [83], for the samples used in this thesis the CTEQ6.5 pdf set is used.

The NLO cross section for top quark pair production at the LHC at a center-of-mass energy of 10 TeV with the CTEQ6.5 pdf is

$$\sigma^{NLO}_{pp\to t\bar{t}}(10\,\text{TeV}, m_t = 171\,\text{GeV}/c^2, \text{CTEQ6.5}) = 414^{+36+20}_{-38-18}\ \text{pb}, \tag{3.13}$$

while with the MRST2006 pdf the cross section is

$$\sigma^{NLO}_{pp\to t\bar{t}}(10\,\text{TeV}, m_t = 171\,\text{GeV}/c^2, \text{MRST2006}) = 446^{+20+8}_{-42-8}\ \text{pb}. \tag{3.14}$$

At a center-of-mass energy of 14 TeV the NLO cross section with the CTEQ6.5 pdf increases to

$$\sigma^{NLO}_{pp\to t\bar{t}}(14\,\text{TeV}, m_t = 171\,\text{GeV}/c^2, \text{CTEQ6.5}) = 908^{+82+30}_{-85-29}\ \text{pb}, \tag{3.15}$$

while with the MRST2006 pdf the cross section is

$$\sigma^{NLO}_{pp\to t\bar{t}}(14\,\text{TeV}, m_t = 171\,\text{GeV}/c^2, \text{MRST2006}) = 961^{+89+11}_{-91-12}\ \text{pb}. \tag{3.16}$$

The first uncertainty on the predicted cross section is due to variations of the renormalization scale while the second uncertainty is due to the uncertainties on the pdf sets. The contribution of the scale uncertainties on the predicted cross section is the most important one. In the calculations of the cross section two scales have to be set, the factorization scale which separates the long and short distance physics and the renormalization scale for removing divergences in the higher order calculations. Both scales have been set to the top quark mass m_t. To evaluate the uncertainty induced by these scales they have been varied independently around the top quark mass by setting them equal to $m_t/2$ or $2m_t$. The uncertainty on the cross section due to scale variations is found to become smaller when taking into account NNLO corrections in calculations.

The uncertainty on the cross section due to uncertainties on the parton distribution functions, which is the dominant uncertainty on the theoretical cross section at the Tevatron [84], is found to be less important at the LHC. The reason is that at the LHC the x values for top quark production are much smaller then at the Tevatron. The experimental knowledge for quark and gluon pdf's is better constrained by data at lower values of x. The same is

found when going from a center-of-mass energy of 10 TeV to 14 TeV where the relative uncertainty due to pdf uncertainties becomes smaller with increasing center-of-mass energy. Also the mass dependency of the cross section is found to be smaller at the LHC than at the Tevatron since the top quark pair is produced further away from the production threshold [82].

In order to combine the NLO predictions of the cross section with the LO generators used in this thesis a K-factor is introduced. This factor is calculated as the ratio of the LO and the NLO cross section and serves as a scaling factor to rescale the obtained results at LO to NLO. This K-factor can depend on the kinematic properties of the event but in practice the K-factor varies only slowly and can be approximated by one single number [28]. The cross sections for the samples used in this thesis are given in Tables 3.3 and 3.4 in Section 3.8.

3.6 Simulating the CMS detector

The complete simulation of the digital response of the CMS detector is a complex task. To perform this task a detailed description of the geometry and material of the detector is needed to exactly simulate all the trajectories and interactions of the particles traversing the various detector components. Therefore the geometry of the detector and the materials of the detector components are stored in a database together with a map of the magnetic field. In addition to this information a modeling is needed of physics processes like eg. scattering and absorption of the particles in the different detector components. Apart from the interaction with the 'dead' material like support- and cooling-structures, a simulation is needed for the active detector components. These simulated electronic signals produced by the active detector materials are then digitized to result in a data-stream reflecting the output of real detector data. A widely used, object oriented, software package suited for this task is the `GEANT4` simulation toolkit [49] and is used in CMS to produce a detailed simulation of the detector response.

The propagation of each individual particle through the complex and dense sub-detectors is a computing intensive task making the full simulation very time consuming. To speed up the simulation an alternative, fast detector simulation is used in parallel to the full detector simulation, better know as fastsim [85]. In fastsim a much less computing intensive simulation is achieved by using a simplified detector geometry, dedicated parametrizations for the calorimeters systems and an alternative tracking algorithm. The fast simulation is used to generate large event samples for example needed for studying systematic effects since it is of the order of 100 times faster than the full simulation. The general outcome of the fast simulation is in good agreement with the results obtained in the full simulation [86]. In Figure 3.6 a comparison is made between jets in semi-muonic $t\bar{t}$ events, simulated with `MadGraph` either with a full or a fast simulation of the CMS detector. Jets reconstruct the direction and energy of the initial partons by clustering the partons produced in the parton shower process and hadronization, a more elaborate description of jets is given in Section 4.2. The figures show the transverse momentum p_T and the pseudo-rapidity η of the first leading p_T jet and the fourth leading p_T jets ordered in descending transverse momentum. Only jets with a transverse momentum $p_T > 15$ GeV/c and a pseudo-rapidity $|\eta| < 2.4$ have been included. A good agreement is found for the pseudo-rapidity of the first leading

and the fourth leading jet. For the transverse momentum of the jets a discrepancy is seen, this is due to the different jet energy scale correction applied in the jet reconstruction. Despite this difference, the samples with the fast detector simulation have been used in this thesis because of their larger number of events. The difference between the energy scale of the jets will only affect the number of selected events in the event selection but does not influence the kinematic properties of the events relevant for this thesis. The effect of the uncertainty on the jet calibration will be evaluated as a systematic uncertainty.

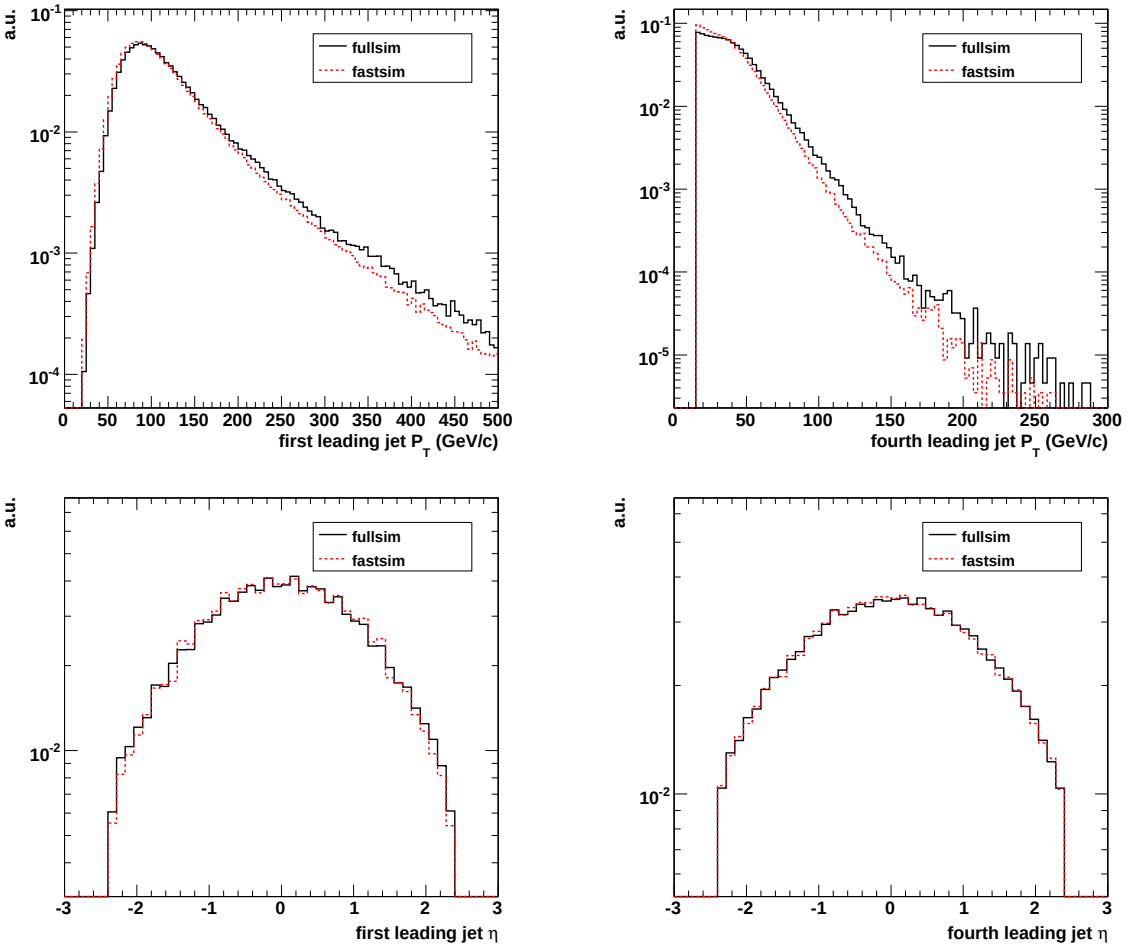

Figure 3.6: The transverse momentum (top) and the pseudo-rapidity (bottom) are shown for the first leading jet (left) and the fourth leading jet (right) comparing full simulation and fast simulation of the detector.

3.7 Event data model in CMS

The final step in the event simulation chain is to build from the digital response of the detector, either coming from simulation or from real collisions, the high-level physics objects suitable for analysis. This has the advantage that the same software tools, developed on simulated collisions, can be applied on real detector data. In this section an overview is

given of the framework used to perform this task. Starting from the digital output of the detector all information is gathered in a data format, a so-called event, which is the central concept of the CMS event data model (EDM). This event is processed until all basic event information is reconstructed in objects suitable for physics analysis.

RAW data format

Based on the digital output of the detector, the Trigger and Data Acquisition System (TriDAS) selects interesting events to be saved for physics analysis. This system packs all the digital information into the so-called RAW events data format. Basically this format contains detector output, the level 1 trigger results, the result of the high level trigger and some higher level objects needed in the high level trigger processing. The typical size of a RAW event is of the order of 1.5 MB/event.

RECO data format

The RAW events serve as basis for the reconstruction of higher level objects, like tracks, jets, missing energy, etc. by means of the CMS software framework (CMSSW). The CMS software framework adopts a modular approach in the sense that different components of the code can be plugged in and out upon the users needs. All desired modules are collected into one executable called cmsRun which will run on the RAW events producing the reconstructed or RECO events. Several types of plug-ins can be distinguished depending on their purpose. The EDProducers will read data in from the event, produce something, eg. an additional reconstruction step and write it back into the event. The EDFilter reads the data and produces a boolean allowing the software to adapt the chain of modules still to be executed depending on it's outcome. The EDAnalyzer, this module reads the data but is not allowed to add or affect the execution chain. It will typically be used for performing analysis and store histograms in a output file. The size of a RECO event is generally smaller than the RAW event and is of the order of 0.25 MB/event.

Analysis Object Data format (AOD)

After the reconstruction step the size of a RECO event is still rather large. Especially since millions of events will be produced by the CMS experiment. Only a subset of this information is relevant for most physics studies. For this reason the Analysis Object Data (AOD) format is available. The AOD will contain only the relevant objects for physics analyses and will, in contrary to the RECO events, not support additional re-reconstruction with different parameter settings. The event-size is reduced to the order of 50 kB/event.

Physics Analysis Toolkit (PAT)

On top of the AOD/RECO data format an extra layer of data format is built, namely the Physics Analysis Tools data format or shortly PAT. The PAT will extend the RECO objects by gathering all related information into a single self-contained object. Eg. the tracks associated to a jet will be added to the jet itself. This makes the associations between different objects in the event obsolete and can be removed and each object can be dealt

with independently. The event-size is further reduced to a few kb/event depending on the stored physics objects

Top Quark Analysis Framework (TQAF)

The aim of PAT events is to be generally applicable in most analyses in CMS. For the analysis in this thesis the events in PAT format are transformed into a top quark physics oriented structure. This happens in the Top Quark Analysis Framework or shortly TQAF [87]. The PAT objects are combined in the TQAF in different types of top quark final state topologies. The advantage of these topologies is that specific analysis tools can be developed more easily. It is on top the TQAF objects that the analysis in this thesis is built.

3.8 Overview of simulated event samples

An overview of the event samples used in this thesis is given here. In Table 3.3 an overview is given of the samples generated with a full GEANT4 detector simulation. The $t\bar{t}$ samples are generated with different event generators and with different top quark masses. The NLO cross section (cf. 3.13) is used for the $t\bar{t}$ sample generated with PYTHIA and MadGraph while the LO cross section is used for the ALPGEN samples. The background processes were generated using the MadGraph event generator. The single top samples are generated separately for the different production modes. For the s-channel and the t-channel only events are stored where the W boson decays in a lepton and a neutrino. The multi-jet samples are generated in $\widehat{p_T}$-bins corresponding to the transverse momentum of the hard interaction in its rest-frame. An alternative multi-jet sample, $pp \rightarrow \mu + X$ ($p_T^{\mu} > 15$ GeV/ c), was generated with additional cuts in the event generation procedure. Only events were at least one muon with a transverse energy greater than 15 GeV/ c were stored. The cross sections of the background processes are given in [88–93]

In table 3.4 the samples generated with a fast detector simulation are listed. Due to the small integrated luminosity of the W + jets and Z + jets sample with the full detector simulation additional samples are generated with fast simulation. To study the influence on the ISR/FSR uncertainty private $t\bar{t}$ samples were produced with altered settings for ISR/FSR, these parameters are discussed in Section 3.3.1.

Sample	generator	σ^{NLO} (pb)	# events	$\mathcal{L}$ (fb^{-1})
$t\bar{t}$ + jets ($m_{top} = 170.9$ GeV/c^2)	MadGraph	414	947 k	2.3
$t\bar{t}$ ($m_{top} = 172.4$ GeV/c^2)	PYTHIA	414	103 k	0.25
$t\bar{t}$ + 0 jets excl.	ALPGEN	118 (LO)	149 k	1.3
$t\bar{t}$ + 1 jets excl.	ALPGEN	61.8 (LO)	66 k	1.1
$t\bar{t}$ + 2 jets excl.	ALPGEN	20.6 (LO)	30 k	1.4
$t\bar{t}$ + 3 jets excl.	ALPGEN	5.2 (LO)	16 k	3.0
$t\bar{t}$ + 4 jets incl. ($m_{top} = 175$ GeV/c^2)	ALPGEN	1.6 (LO)	11 k	7.2
single top				
tW-channel	MadGraph	29	169 k	5.8
s-channel (only leptonic decay)	MadGraph	1.6	12 k	7.5
t-channel (only leptonic decay)	MadGraph	41.6	282 k	6.8
W + jets	MadGraph	45.6 10^3	9.7 M	0.2
Z + jets	MadGraph	4.2 10^3	1.3 M	0.3
W + c + jets	MadGraph	1.5 10^3	3.3 M	2.0
V + qq + jets (V=W/Z and q=b/c)	MadGraph	290	968 k	3.3
multi-jet $\widehat{p_T}$-bins				
100-250 GeV/c	MadGraph	15 10^6 (LO)	22 K	0.8 10^{-3}
250-500 GeV/c	MadGraph	400 10^3 (LO)	115 K	13 10^{-3}
500-1000 GeV/c	MadGraph	14 10^3 (LO)	448 K	0.3
1000-∞ GeV/c	MadGraph	370 (LO)	260 K	2.9
$pp \to \mu + X$ ($p_T^{\mu} > 15$ GeV/c)	PYTHIA	122 10^3 (LO)	6.3 M	49 10^{-3}

Table 3.3: Overview of the centrally produced full simulation samples and their event generator. The cross section of the process, the size of the sample and the corresponding integrated luminosity are given.

Sample	generator	σ^{NLO} (pb)	# events	$\mathcal{L}$ (fb^{-1})
$t\bar{t}$ + jets ($m_{top} = 170.9$ GeV/ c^2)	MadGraph	414	10.8 M	26.0
W + jets	MadGraph	45.6 10^3	94.2 M	2.0
Z + jets	MadGraph	4.2 10^3	7.8 M	1.8
ISR/FSR				
$t\bar{t}$ nominal	PYTHIA	414	4.0 M	9.7
$t\bar{t}$ more ISR/FSR	PYTHIA	414	2.0 M	4.9
$t\bar{t}$ less ISR/FSR	PYTHIA	414	2.1 M	5.1

Table 3.4: Overview of the centrally and privately produced fast simulation samples and their event generator. The cross section of the process, the size of the sample and the corresponding integrated luminosity are given.

Chapter 4

Object reconstruction and flavour identification

In this chapter an overview is given of the reconstruction algorithms to build the higher level physics objects used in the analysis in Chapter 6. These algorithms combine the signals from the various subdetectors to build complex physics objects representing the final state particles formed in the proton collisions. The analysis to estimate the b-tag efficiency is based on the semi-muonic $t\bar{t}$ decay channel, $t\bar{t} \rightarrow bqqb\mu\nu_\mu$, therefore the objects of interest are muons and jets and the identification of the flavour of jets. The same analysis could as well be applied in the semi-electronic $t\bar{t}$ decay channel.

In Section 4.1 the reconstruction of the muon is introduced. This reconstruction algorithm combines the information of the muon system and the tracking system to build the trajectory of the muon through the detector. In Section 4.2 the algorithms clustering the energy deposits in the electromagnetic and hadronic calorimeter into jet objects are described. These jets represent the energy and the direction of the partons before their hadronization. In the final Section 4.3 the complex task of identifying the flavour of the parton initiating the observed jet is introduced. The broad spectrum of bottom flavour identification algorithms or shortly b-tagging algorithms is described. At the end of each section the performance of the reconstruction algorithms is briefly discussed.

4.1 Muon reconstruction

The muon plays a key role in the discrimination of semi-muonic $t\bar{t}$ events from the enormous multi-jet background. Therefore a good reconstruction of the trajectory of the muon and the information about its isolation are prerequisites to successfully separate signal events from background. To provide a good determination of the trajectory of the muon through the detector, the information of the outer part of the CMS detector, the muon system, is combined with the most inner part of CMS, the tracker and pixel detector. The muon reconstruction in CMS is performed in three stages [35]. In the first stage, the local reconstruction, information of the muon sub-systems is combined into track segments which serve as regional seeds for further trajectory building. These track segments or seeds are then combined in the standalone reconstruction step (cf. Section 4.1.1) to build the muon trajectory in the muon system only. In the third and final step, the global muon recon-

struction step (cf. Section 4.1.2), the trajectories of the standalone muon are extrapolated towards the interaction point to search for compatible tracks in the tracking system (cf. Section 2.2.2). A last refit of the trajectory then provides the muon four-momentum and trajectory. After the reconstruction of the standalone and global muon the performance of the muon reconstruction is discussed in Section 4.1.3.

4.1.1 Standalone muon reconstruction

The standalone muon reconstruction procedure uses only information from the muon systems, namely the drift tubes (DT), the cathode strip chambers (CSC) and the resistive plate chambers (RPC). Although the much less precise spatial resolution of the RPCs, its information complements the other detectors by extending the geometrical coverage in the overlap region between barrel and endcap.

The track position, momentum and direction associated with the segments found in the innermost chambers will serve as seeds to reconstruct the muon trajectory using the Kalman-filter approach [37]. The trajectory building is performed inside out where using the DT segments in the barrel and the individual reconstructed hits in the CSC chambers are used due to the inhomogeneous magnetic field for the endcap. Reconstructed hits from the RPC chambers are included as well. For each inclusion of a new layer, going towards the outer layers, the predicted track parameters are compared to the measured parameters and the track parameters are updated. The propagation of one layer to the next one takes into account the muon energy loss in the material, the non-uniform magnetic field and the possibility of multiple scattering in the muon system. When the outermost layer is reached the procedure ends and an outside-in Kalman-filter is applied. This provides the track parameters at the innermost muon layer and an extrapolation to the interaction region.

4.1.2 Global muon reconstruction

To perform the global muon reconstruction the track from the standalone muon reconstruction is extrapolated inwards from the innermost layer of the muon system to the outermost tracker layer. The extrapolation takes into account the effects of energy losses in the trespassed detector material and the effect of multiple scattering to define a region of interest to perform regional track reconstruction. This region of interest is based on the parameters of the extrapolated track and its uncertainties assuming that the muon originated from the interaction point.

Inside the region of interest, two reconstructed hits from different tracker layers are combined to define the regional seed. Starting from this seed the Kalman-filter approach is used to reconstruct local tracks inside the region of interest. The resulting tracks are refitted in a final step combining the hits from the standalone muon and the tracker system. To determine the final global muon candidates a cleaning is performed based on the χ^2-value of the fit. In addition to this cut, the χ^2-value of a refit, using only tracker hits and the innermost muon layer, is compared to the χ^2-value of tracker-only tracks to detect significant energy losses due to multiple interactions in the material in between the tracker and the muon system.

4.1.3 Performance of the muon reconstruction

Figure 4.1 shows the total number of reconstructed global muons in semi-muonic $t\bar{t}$ events. The $t\bar{t}$ decay modes, such as the semi-muonic $t\bar{t}$ events, are classified based on the information from simulation. In about 65% of the events exactly one global muon is reconstructed. Also shown is the number of global muons in semi-muonic $t\bar{t}$ events that are not matched to the generated muon from the leptonic W boson decay. Reconstructed muons are considered matched to the generated muon from the W boson decay if the angular distance (in (η, ϕ)-space) is $\Delta R<0.2$. It can be verified that these unmatched muons come primarily from the decays of heavy hadrons produced in b quark and c quark jets. In about 30% of the events such a muon is reconstructed. In Figure 4.2 the transverse momentum and the pseudo-rapidity are shown for all muons in semi-muonic $t\bar{t}$ events and for muons not matched to the generated muons. Muons not produced in the W boson decay have a softer p_T spectrum and are less central compared to muons from the W boson. The properties of the muons in b quark jets will be exploited in the soft muon b-tagging algorithms, introduced in Section 4.3.3.

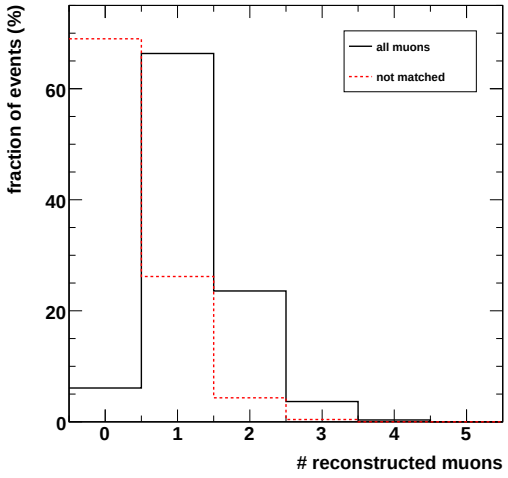

Figure 4.1: The total number of reconstructed muons in semi-muonic $t\bar{t}$ events and the number of reconstructed muons not matching the generated muon from the W boson decay.

The relative resolution on the transverse momentum, the polar angle θ and the azimuthal angle ϕ of the reconstructed muons, matching to the W boson muon, are displayed in Figure 4.3. The relative resolution is obtained from the difference between the reconstructed and the generated muon transverse momentum,

$$\frac{p_T^{rec} - p_T^{gen}}{p_T^{gen}}. \tag{4.1}$$

This distribution is fitted by a Gaussian function, the standard deviation σ of this function is the relative resolution. The relative resolution on the transverse momentum is found to

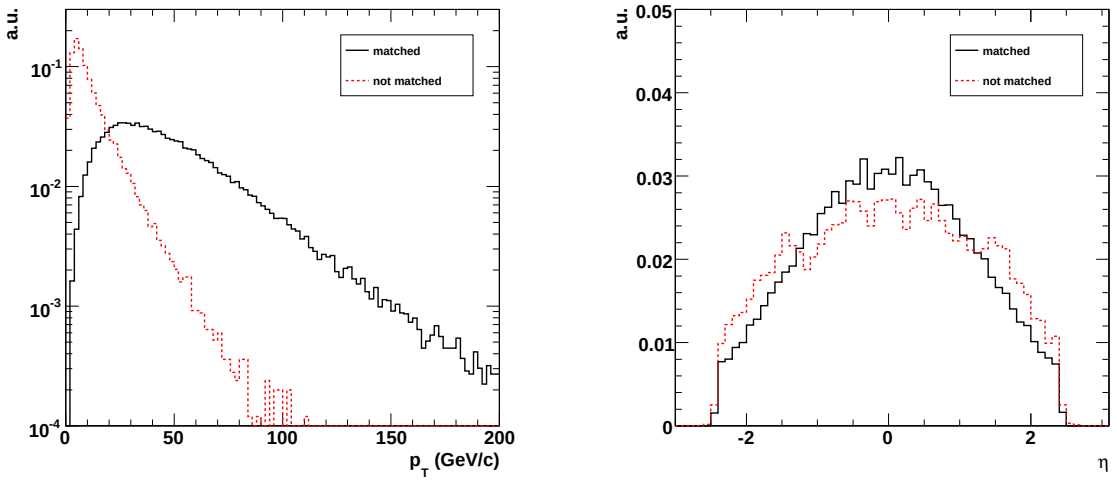

Figure 4.2: Transverse momentum (left) and pseudo-rapidity (right) of global muons in semi-muonic $t\bar{t}$ events. Reconstructed muons matching and not matching the generated muon from the W boson decay are shown separately.

become worse with increasing transverse momentum. The relative resolutions on the polar angle θ and azimuthal angle ϕ is found to improve with increasing transverse momentum.

4.2 Jet reconstruction

The quarks and gluons present in the proton collision cannot exist in free form and thus fragment into stable hadrons. This process results in a jet of particles, depositing energy in the CMS detector. In this section the algorithms are introduced to reconstruct the energy and the direction of these jets of particles.

In Section 4.2.1 two jet algorithms are discussed clustering the four-momenta of the calorimeter towers (cf. Section 2.2.3) into jets. Several detector effects influence the energy determination of the jets leading to a biased estimation of the energy. To account for this bias a factorized jet energy calibration approach is adopted as introduced in Section 4.2.2. In the last Section 4.2.3 the resolution of the reconstructed jets for the introduced algorithms is discussed.

4.2.1 Jet algorithms

A wide range of jet algorithms, like the iterative cone algorithm [94], midpoint cone algorithm [95] and the seedless infrared safe cone algorithm [96], exists and are extensively used in collider experiments. Cone-based algorithms generally use the $\Delta R = \sqrt{(y_1 - y_2)^2 + (\phi_1 - \phi_2)^2}$ metric[1] based on the rapidity y, while k_T-based algorithms, like the inclusive k_T algorithm [97], use a E_T-weighted ΔR metric based on the rapidity y.

[1] In this thesis the angular distance ΔR is generally computed in (η, ϕ)-space, $\Delta R = \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2}$, with exception of this section describing the jet algorithms where it is calculated in (y, ϕ)-space.

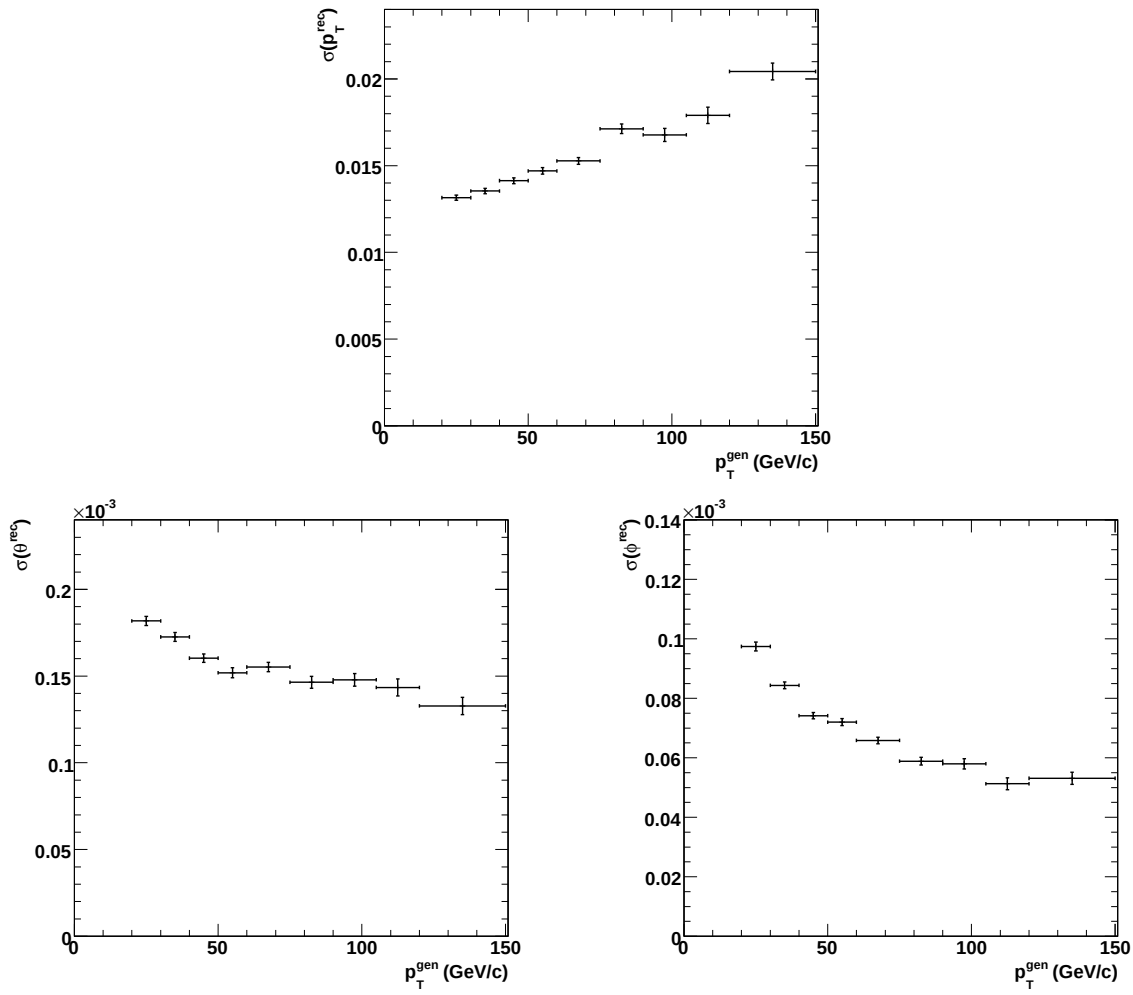

Figure 4.3: The relative resolution on the transverse momentum (up), the polar angle θ (left) and the azimuthal angle ϕ (right) of the reconstructed global muons in semi-muonic $t\bar{t}$ events. Only reconstructed muons matching to generated muon from the W boson decay are considered.

The input objects of jet clustering algorithms can be of different types for as long as they are represented by an energy deposit and a direction. In this thesis the energy deposits in the calorimeters and the direction of the calorimeters w.r.t. the interaction point will be used. More complex input objects can be used as well. Eg. tracks can be included to have a more accurate estimation of the direction of the jet and its energy. Ultimately jets can be reconstructed using so-called particle-flow objects. The particle flow reconstruction algorithms aim to reconstruct each individual particle in the detector making use of all CMS subdetectors.

When adding the four-momenta in the clustering process two possibilities are present to add these four-momenta. In the first recombination scheme the jet constituents are added as four-vectors, resulting in massive jets, this is the so-called E-scheme. In second scheme, the E_T-scheme, the transverse momentum of the jet is equated to the scalar sum of the transverse energy E_T of the jet constituent leading to massless jets. In this section

the seedless infrared safe algorithm and the inclusive k_T algorithm are introduced using the E-scheme based only on calorimeter information.

Seedless infrared safe cone algorithm

The seedless infrared safe cone algorithm (SC), in contrast to the other cone algorithms like the iterative cone algorithm, does not start from input objects above a certain E_T-threshold, so-called seeds. It however uses any input object to search for the stable cones in the event. To find the stable cones the four-momenta of the input objects are added in a cone and the direction of the cone is compared to the summed four-momentum of the input objects enclosed in this cone. A cone is considered stable if the direction of the cone agrees with the input objects, if it is found to be unstable the direction of the input objects is used to define a new cone. The stable cones are added to a list of proto-jets.

After all stable cones around a set of input objects are found, it is possible that input objects are shared among the proto-jets. A split-and-merge procedure is applied to resolve assignments of the same input objects to several proto-jets. All proto-jets are ordered in descending transverse momentum p_T and proto-jets not exceeding a lower threshold are discarded. The first jet of the p_T-ordered list is chosen and the highest p_T proto-jet that shares input objects with the first proto-jet is looked for. If no such overlapping proto-jet exists the proto-jet is removed from the list and it is added to the list of jets. If the overlapping fraction of scalar sum transverse momentum of the two jets exceeds a threshold f, the two proto-jets are merged, otherwise the proto-jets are split in two proto-jets by assigning the input objects to either one of the proto-jets. This process is repeated until no proto-jets remain. The parameters for this algorithm are the cone size, the allowed fraction of momentum sharing f and the minimal p_T requirement on the proto-jets.

Inclusive k_T algorithm

The inclusive k_T algorithm (KT) calculates for each input object i and each pair of input objects (i, j) two distances:

$$d_i = E_{T,i}^2 R^2 \tag{4.2}$$

$$d_{ij} = min\{E_{T,i}^2, E_{T,j}^2\} R_{ij}^2 \tag{4.3}$$

where R is a dimensionless parameter and R_{ij} is the angular distance in (y, ϕ)-space. The algorithms searches for the smallest value among d_i and d_{ij}. If a value of the type d_{ij} is minimal, the two input objects are merged to form one object. In the case d_i is minimal the corresponding object is removed from the list and added to the list of final jets. This procedure is repeated until no objects remain and all final jets are found. For the final jet the distance between them R_{ij} is always larger than R. The parameter R can be interpreted as a similar parameter as the cone size in the seedless infrared safe cone algorithm.

Properties of reconstructed jets in semi-muonic $t\bar{t}$ events

A comparison between the kinematic properties of the reconstructed jets are shown for semi-muonic $t\bar{t}$ events. The seedless infrared safe cone algorithm was used with an opening angle $R = 0.5$, while for the k_T algorithm the R parameter was fixed to 0.4. In Figure

4.4 the total number of jets reconstructed with the seedless infrared safe cone algorithm (SC) is compared to the total number of jets reconstructed with the inclusive k_T algorithm (KT), no selection criteria are applied on the transverse momentum or the pseudo-rapidity of the jets.

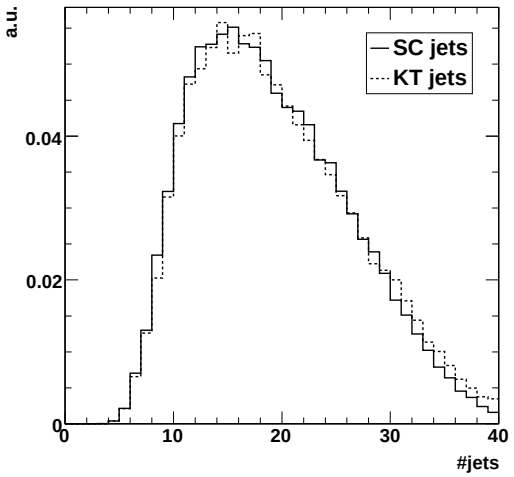

Figure 4.4: The total number of reconstructed jets in semi-muonic $t\bar{t}$ events.

In Figure 4.5 the uncalibrated transverse momentum p_T and the pseudo-rapidity η is shown for the jet with highest p_T and for the fourth jet, ordered in descending p_T. In general a good agreement is found between the two jet reconstruction algorithms.

A good estimation of the direction of a jet is important for the identification of b quark jets. The algorithms for flavour identification which will be introduced in Section 4.3 rely on the direction of the jet to reconstruct variables sensitive to the presence of heavy hadrons in the jet. In Figure 4.6 the angular distance between a jet reconstructed with the SC algorithm and the KT algorithm is shown as a function of respectively the transverse momentum and the pseudo-rapidity of the generated parton matching both jets. The generated partons are the decay products of the top quarks in the semi-muonic $t\bar{t}$ channel and match to a reconstructed jet if the angular distance is smaller than $\Delta R < 0.2$. The average distance between the two reconstructed jets is constant as a function of the pseudo-rapidity of the parton and is about 0.02. As a function of the transverse momentum a clear dependency is observed. For low p_T partons the angular distance is on average larger. This is due to a worse reconstruction of the direction of jets at low p_T (cf. Section 4.2.3).

4.2.2 Jet energy scale calibration

Many detector effects complicate the reconstruction of the energy of jets. These effects diffuse and bias the jet energy compared to the parton shower it is representing. The uncertainty on the scale of the reconstructed jet energy has an important impact on the systematic uncertainty of the analysis. In CMS jets are calibrated adopting a factorized approach [98]. In a fixed sequence, partial corrections are applied taking care of different detector and physics effects. The following levels of correction are available in CMS.

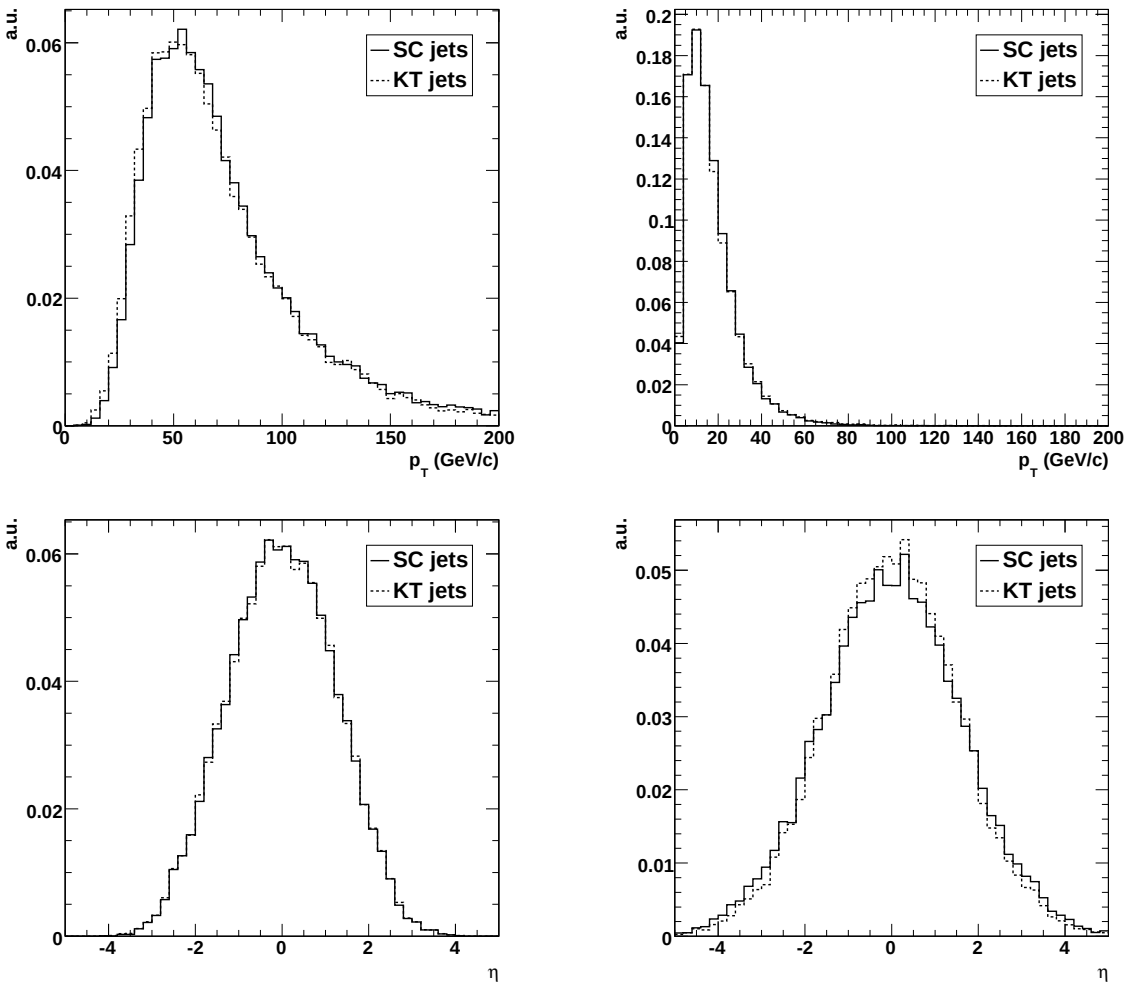

Figure 4.5: The uncalibrated transverse momentum (top) and pseudo-rapidity (bottom) of the first jet (left) and fourth jet (right) in semi-muonic $t\bar{t}$ events.

1. **Offset correction:** this correction is applied to account for pile-up and electronic noise in the calorimeter system.

2. **Relative η-dependent correction:** to correct for the variation in jet response as function of the pseudo-rapidity relative to a control region ($|\eta| < 1.3$).

3. **Absolute p_T-dependent correction:** to obtain the particle level energy scale a p_T-dependent energy correction is applied.

4. **EMF-correction:** the variation in jet energy response with the electromagnetic energy fraction (EMF) can be corrected with this factor.

5. **Flavour dependent correction:** this correction is to correct for the difference in response for different flavour types (gluon, light and heavy quark).

6. **Underlying event correction:** to correct for the underlying event response this correction factor can be applied.

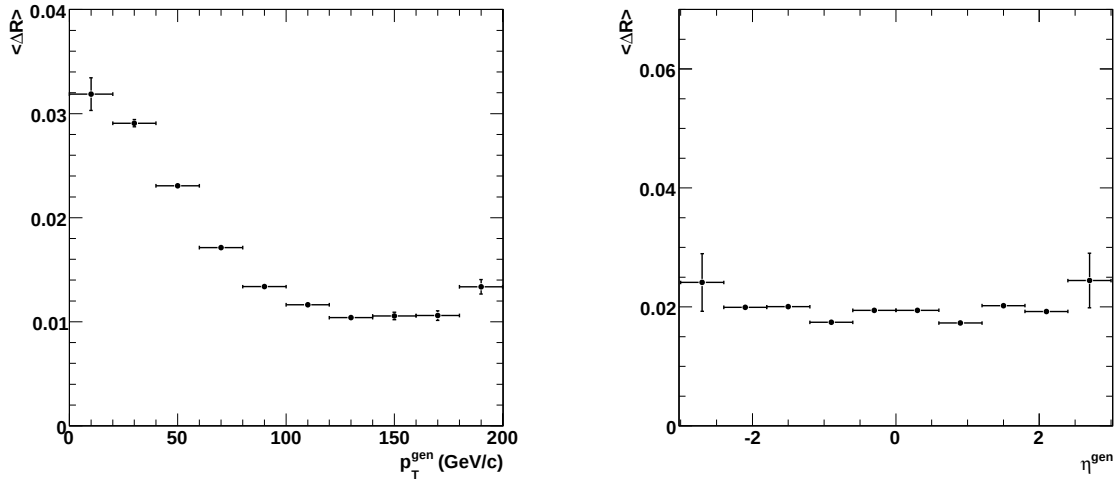

Figure 4.6: The average angular distance between SC jets and KT jets as a function of the transverse momentum (left) and pseudo-rapidity (right) of the generated parton, in semi-muonic $t\bar{t}$ events

7. **Parton correction:** this correction will change the energy scale of the reconstructed jet to the energy scale of the initial parton.

The main advantage of the factorized approach is that the various jet energy scale corrections can be determined quasi independently. In the first phase of the CMS experiment the corrections will be obtained from simulated events, while at a later stage, when more data becomes available, data-driven techniques will be used to determine some of the correction factors. In the analysis presented in this thesis the jets have been corrected with the level 2 and level 3 corrections, this to achieve a jet energy response which is flat in transverse momentum and pseudo-rapidity.

In Figure 4.7 the reconstructed W boson mass and top quark mass in, semi-muonic $t\bar{t}$ events, is displayed for both the SC jet algorithm and the KT jet algorithm. The reconstructed W boson mass peak and reconstructed top quark mass peak is shifted when comparing the jet algorithms. This is due to different calibration factors used in the jet energy correction scheme for both algorithms.

4.2.3 Jet resolutions

Analogue to the resolutions on the reconstructed muon properties, the resolution on the kinematic properties of the reconstructed jets can be obtained by matching them to the generated quarks of the semi-muonic $t\bar{t}$ events. A jet is considered matched to a quark if the angular distance, in (η, ϕ)-space, is smaller than $\Delta R < 0.2$. The resolution is obtained from the distribution of the relative difference (cf. Equation 4.1) between the generated reconstructed property of the calibrated jet. This distribution is fitted with a Gaussian function and the standard deviation of this function is the resolution. The resolution is differentiated between jets from b quarks and light quark jets from up, down, strange and charm quarks and gluons. In Figure 4.8 the resolution on the transverse momentum, the polar angle θ and the azimuthal angle ϕ of the reconstructed jet is shown as a function

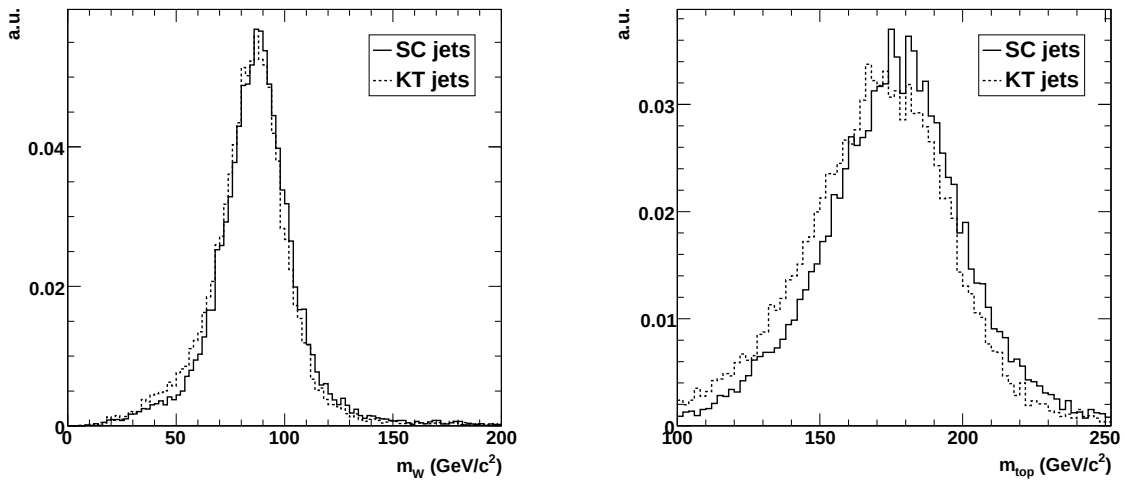

Figure 4.7: Reconstructed W boson mass (left) and top quark mass (right) with the SC algorithm and the KT algorithm in semi-muonic $t\bar{t}$ events.

of the transverse momentum of the matching generated parton. The resolution on the transverse momentum is found to be worse for low-p_T jets. This is due to the magnetic field strongly bending charged particles with a low momentum causing them to bend outside the jet cone. The same effect induces a worse resolution of the direction, parametrized by ϕ and θ, for jets matched to a parton with a low transverse momentum. No important difference is found between the b quark jets and the jets from other quarks as well as no significant difference is found between the different jet reconstruction algorithms. A more detailed comparison between the SC algorithm and the KT algorithm can be found in [99].

4.3 b-Tagging algorithms

Heavy hadrons are formed in the hadronization process of heavy quarks, such as b quarks and c quarks. As discussed in Section 3.4.2 these hadrons have a significantly long lifetime which can be exploited to identify the heavy quark jets. A very important property is the presence of significantly displaced tracks w.r.t. the primary vertex, coming from the decay of a heavy hadron in the jet. Another property of the heavy quarks is the potential presence of a soft lepton in the jet which is originating from the leptonic decay of a b or c quark. Both properties are used to construct algorithms, so-called b-tagging algorithms, to identify b quark jets. In Section 4.3.1 b-tagging algorithms exploiting the presence of displaced tracks are discussed while in Section 4.3.2 the presence of a fully reconstructed displaced secondary vertex is exploited. The b-tagging algorithm introduced in 4.3.3 exploits the leptons present inside the jets to distinguish them from light jets. Often the various b-tagging algorithms are based on very similar information and are thus strongly correlated, Section 4.3.4 elaborates on this correlation. Finally Section 4.3.5 is dedicated to the performance of the b-tagging algorithms based on simulated events and a description is given of the methods to estimate their efficiency using data.

When reconstructing simulated jets, their flavour can be assigned according to two definitions. Given the list of partons close to the reconstructed jets (ΔR <0.3) the physics

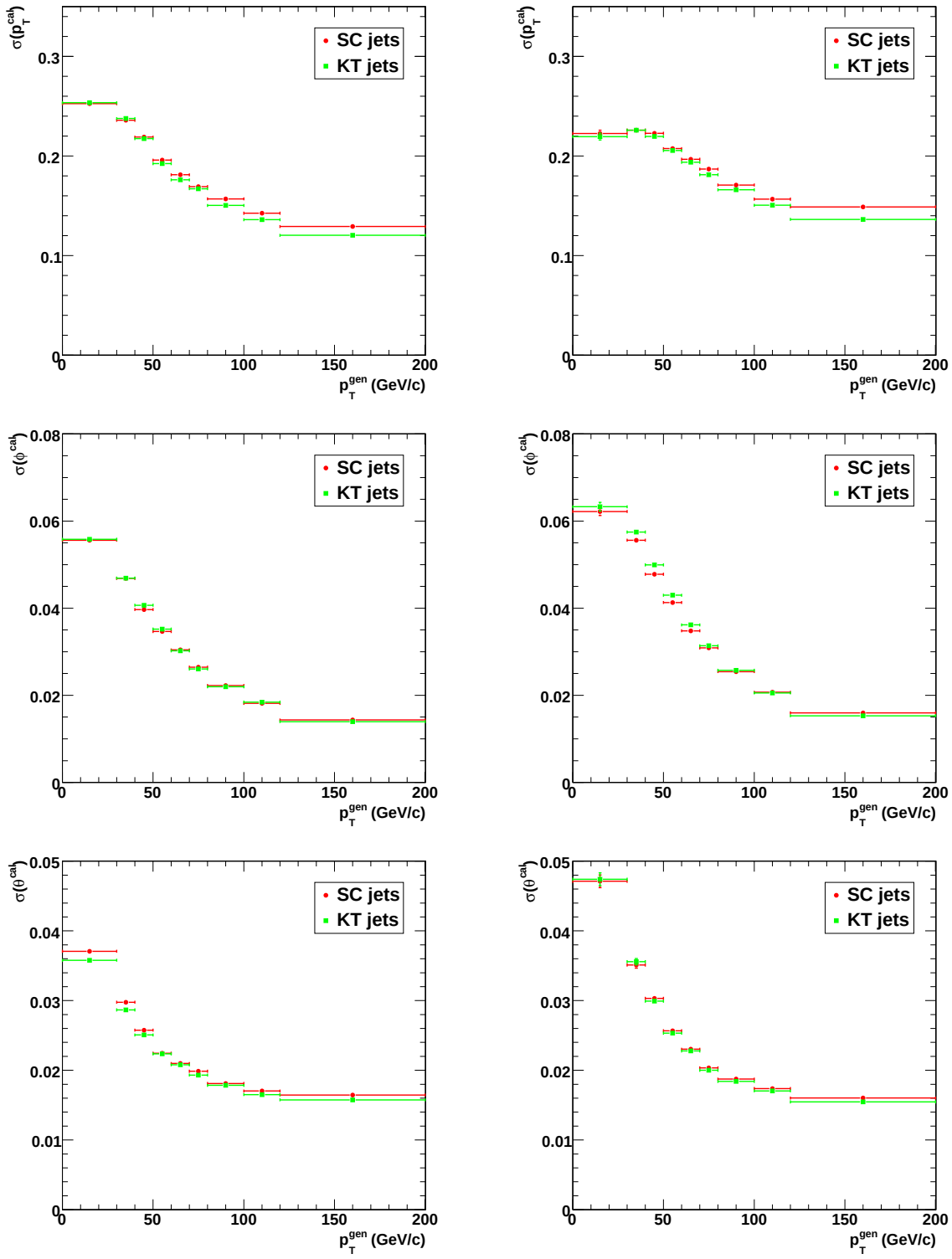

Figure 4.8: The resolution on the transverse momentum (upper), the azimuthal angle (middle) and the polar angle (lower) for non-b quark jets (left) and b quark jets (right) in semi-muonic $t\bar{t}$ events.

definition assigns a flavour to the jet according to the flavour of the initial parton. The algorithmic definition assigns the flavour to the jet according to the heaviest parton close to the jet. The main difference arises in the treatment of the jets from gluons where a $c\bar{c}$ or a $b\bar{b}$ pair can be formed (cf. Section 3.3.2). The algorithmic definition would then label the jet as a c quark jet or a b quark jet while the physics definition would label the jet as a gluon jet. In the following section and throughout this thesis the physics definition is used. Reconstructed jets not associated to any parton using this algorithm ($\mathcal{O}(1\%)$) are discarded for the performance studies of the b-tagging algorithms in this section. Only jets reconstructed with the seedless infrared safe cone algorithm are considered if their transverse momentum is greater than $p_T > 20$ GeV/c and their pseudo-rapidity is within $|\eta| < 2.4$.

4.3.1 Impact parameter based b-tagging algorithm

Tracks in the event are reconstructed using the combinatorial Kalman filter algorithm and are required to fulfill some basic quality requirements to reduce fake and mis-reconstructed tracks. The tracks must have sufficient hits in both the silicon strip detector as the pixel detector to ensure precise track extrapolation in the vicinity of the primary vertex. The tracks need to have a minimal transverse momentum of 1 GeV/c and a good fit quality by constraining the χ^2 of the tracks. The association of tracks to the jets is based on a simple ΔR criterion, tracks with an angular distance to a jet smaller than $\Delta R 0.5$ (in (η, ϕ)-space) are considered associated to that jet.

The impact parameter of a track is quantified by the distance of the track trajectory to the primary vertex as illustrated in Figure 4.9. The calculation is done either in the transverse plane or in three dimensions taking into account the longitudinal position of the track. For both the transverse and the three-dimensional impact parameter the track parameters at the innermost measurement point are used. The closest point of approach of the track to the jet direction is extracted. From this point the tangent of the track is determined and the impact parameter is the distance of the primary vertex to the extrapolated tangent of the track.

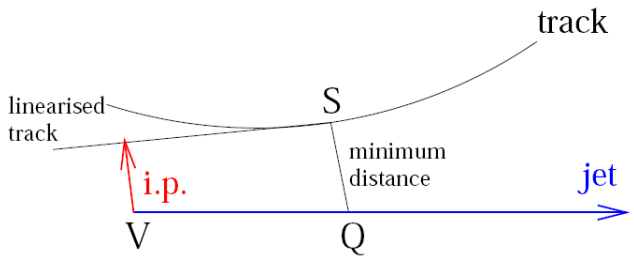

Figure 4.9: Representation of the impact parameter of a track.

The impact parameter is given a positive sign if the angle, measured at the impact point, between the impact parameter and the jet direction is smaller than 90° and a negative sign if the angle is greater than 90°. Since the flight path of the B hadron in a b quark decay is approximately in the same direction as the b quark or the jet, the impact parameter of

the B hadron is expected to be positive. The impact parameter can be negative due to a badly reconstructed jet direction or a badly reconstructed track.

To take into account the experimental resolution of the reconstructed tracks, the impact parameter is divided by its uncertainty, obtaining the impact parameter significance $S = IP/\sigma_{IP}$. Two b flavour identification algorithms, based in the impact parameter significance of tracks, are used in the CMS collaboration. These methods, the track counting method and the jet probability method [100] are described here.

Track counting b-tagging algorithm

To discriminate bottom jets from other jets, a discriminating value, or discriminator, is calculated. According to the value of this discriminator the probability of a jet to originate from a bottom quark can be calculated. The discriminator value for the track counting b-tagging algorithm is defined as the three-dimensional impact parameter significance of the n-th track[2], where the tracks are ordered in descending impact parameter significance. Different options for n are available depending on the needs. If $n = 2$ the algorithm yields a higher efficiency for selecting bottom jets while if $n = 3$ the lower efficiency is compensated by a higher purity of the bottom jet selection. Jets not containing sufficient good quality tracks are not given a discriminator value and are assumed as non-taggable. The normalized distributions of the algorithm with $n = 2$ and $n = 3$ are shown in Figure 4.10 for jets reconstructed with the seedless infrared safe cone algorithm in $t\bar{t}$ events. For the bottom jets a long positive tail is observed giving discrimination power to the algorithm.

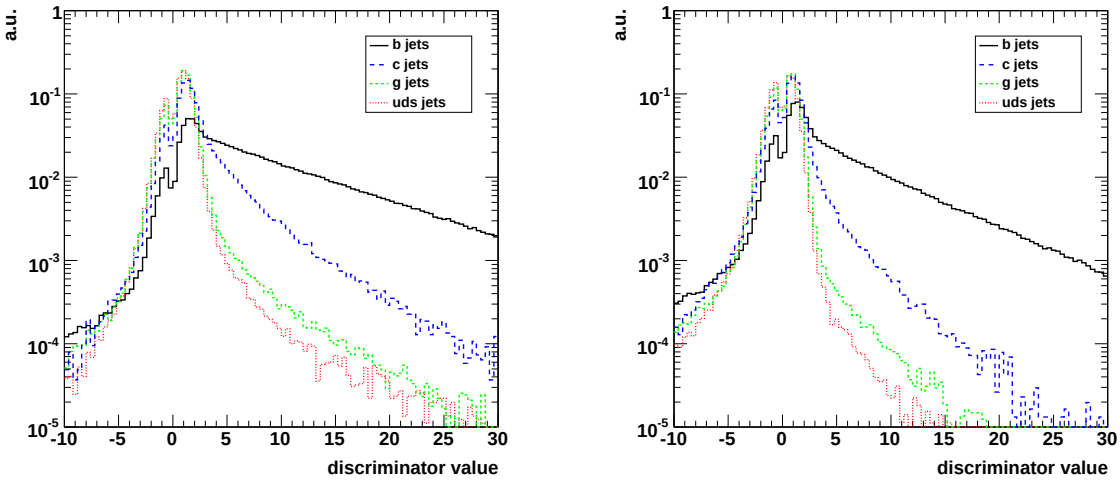

Figure 4.10: The discriminator value of the high efficiency ($n = 2$) track counting b-tagging algorithm (left) and the high purity ($n = 3$) track counting b-tagging algorithm (right) for jets reconstructed in $t\bar{t}$ events.

Jet probability b-tagging algorithm

The jet probability b-tagging algorithm calculates for each jet the probability of the set of tracks associated to the jet to have originated from the primary vertex using the impact

[2]Tracks are associated to the jet if they are closer then $\Delta R<0.5$.

parameter significance of the tracks. This is an extension of the impact parameter significance b-tagging algorithm taking into account not only the first n tracks but rather all tracks associated to the jet. The first step in the algorithm is to calculate the probability $P_{tr}(S)$ for each associated track, with impact parameter significance S, to come from the primary vertex. This is calculated as the following confidence level

$$P_{tr}(S) = sng(S) \int_{|S|}^{\infty} R(x)dx. \tag{4.4}$$

The resolution function $R(x)$, the normalized impact parameter significance distribution, is obtained from the tracks in data with negative impact parameter significance since they are mainly coming from the primary vertex or fake tracks and assuming their distribution to be the same as positive signed tracks. This probability is nearly uniform for tracks coming from the primary vertex while it is peaked at 0 for tracks coming from a displaced vertex. In order to enhance the discrimination power of the method, the tracks are divided in several categories for which different resolution functions are obtained. The categories are the p_T and η of the track, the χ^2 and the number of hits of the track.

The second step in the jet probability b-tagging algorithm combines the probabilities of the tracks associated to the jet. The jet probability for a given jet with N associated tracks is defined as the confidence level that any set of N tracks, presumed to come from the primary vertex, would give the observed track probability or any less likely value. This is given by

$$P_{jet} = \Pi \sum_{j=0}^{N-1} \frac{(-ln\Pi)^j}{j!} \tag{4.5}$$

where

$$\Pi = \prod_{i=1}^{N} \widehat{P}_{tr}(i) \tag{4.6}$$

and where $\widehat{P}_{tr} = P_{tr}/2$ for tracks with positive impact parameter significance S and $\widehat{P}_{tr} = 1 + P_{tr}/2$ for negative S, so that the track probability is always positive. The discriminator for the jet probability tagger is defined as $-log(P_{jet})/4$ and is shown in Figure 4.11. A more performant b-tagging algorithm, called jet B probability tagger, is defined by the following discriminator, $-log(P_{jet})/4 - log(P_{jet}^{4tks})/4$. The variable P_{jets}^{4tks} is the jet probability computed using the four tracks with lowest track probabilities with positive impact parameter significance. This discriminator value is also shown in Figure 4.11. The peaks in the distributions can be understood from the cut-off introduced in the track probabilities P_{tr}. Tracks with a probability value lower than $P_{tr} < 0.005$ are assigned the fixed minimum value of $P_{tr} = 0.005$ leading to peaks at fixed values.

4.3.2 Secondary vertex based b-tagging algorithm

The natural extension of the lifetime based b-tagging algorithms is to go beyond the individual tracks[3] and use the properties of the reconstructed secondary vertex. Due to the very good reconstruction of tracks in the CMS tracking detector displaced vertices can be identified and reconstructed. The reconstruction of displaced vertices additional to the

[3]Tracks are now considered associated to a jet if there angular distance is smaller than ΔR <0.3.

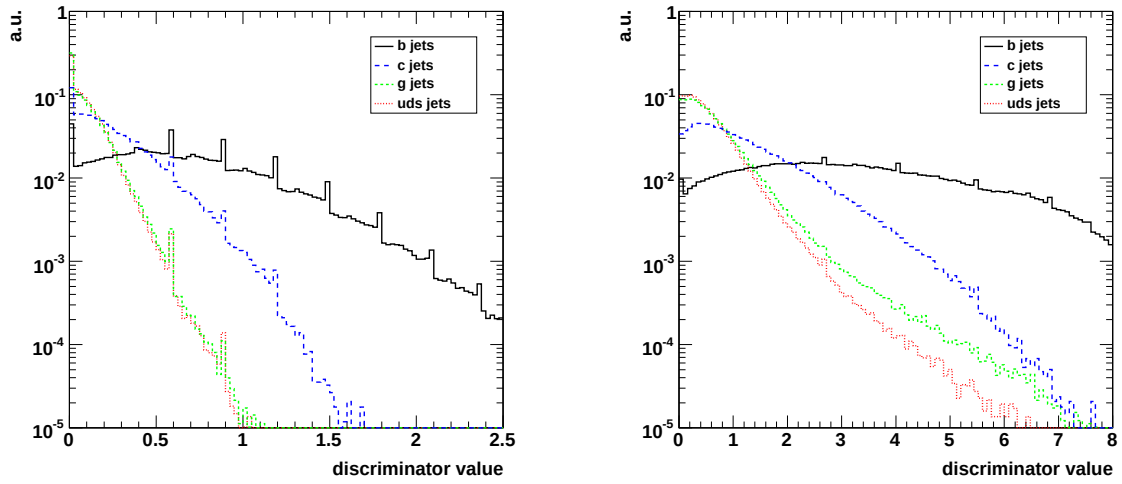

Figure 4.11: The discriminator value of the jet probability b-tagging algorithm (left) and the jet B probability b-tagging algorithm (right) for jets reconstructed in $t\bar{t}$ events.

primary vertex of the event is performed using the Adaptive Vertex Fitter [39], which iteratively determines additional vertices. Each track in the event is given a weight according to the distance of the track to the vertex candidate. This weight can be interpreted as the probability of the track to originate from this vertex candidate. A fit is performed, incorporating these weights, to determine the properties of the vertex candidate. Tracks with a low weight w.r.t. the vertex are considered un-associated and are subject to a new vertex fit. This process is repeated iteratively until no further vertices are found. The resulting list of vertices is then cleaned to reduce the number of poorly reconstructed vertices and to increase the purity of the vertex candidates.

Several vertex candidate categories are identified. In the first place this are the so-called *RecoVertex* candidates which are well reconstructed secondary vertices that pass all cleaning cuts. A second category are the *PseudoVertex* candidates and consists of vertices where no fit could be performed but two tracks with a transverse impact parameter significance greater than 2 are available. Those tracks are then assigned to the vertex allowing the calculation of several vertex related variables. In the simple secondary vertex b-tagging algorithm introduced in the next section only RecoVertex candidates are used whereas the combined secondary vertex b-tagging algorithm uses the PseudoVertex candidates as well.

Simple secondary vertex b-tagging algorithm

A simple secondary vertex b-tagging algorithm is defined from the significance of the flight distance D/σ_D. The flight distance D is the distance between the reconstructed primary vertex and the secondary vertex, calculated in the transverse plane or in three dimensions. The discriminator is defined as $log(1+D/\sigma_D)$ and is shown for the three-dimensional flight distance in Figure 4.12. This vertex based b-tagging algorithm, although using complex high level objects such as secondary vertices, is rather robust to a possible mis-alignment of the tracking system [101] and is favorable for analysis in the start-up phase of the CMS experiment.

Combined secondary vertex b-tagging algorithm

The combined secondary vertex b-tagging algorithm [102] exploits a wide range of lifetime-sensitive variables in a jet to yield a good discriminating, high efficiency b-tagging discriminator. Although this algorithm makes explicitly use of the secondary vertex related variables it is not limited to it. The combination of track related and secondary vertex related variables allows this tagger to yield a discriminator value even if no good secondary vertex was reconstructed, in contrary to the simple secondary vertex b-tagging algorithm. The following list of variables is combined in a multivariate analysis technique, by default the Likelihood Ratio technique is used. According to the presence of a vertex in a certain vertex category, all or only a subset of variables is calculated. Specific probability density functions are used for the different vertex categories and the different p_T and η categories of the jet.

- In the case a vertex candidate in the RecoVertex category is reconstructed, the following variables are used to calculate the likelihood discriminator:
  - **Transverse flight significance:** this is the variable used in the simple secondary vertex b-tagging algorithm and is the significance of the flight distance calculated in the transverse plane.
  - **Angle between the vertex flight direction and the jet axis:** this angle is calculated in the (η, ϕ)-space and is related to the energy carried by the bottom quark. The more energetic it is, the more collinear it will be with the jet direction.
- If a vertex candidate is reconstructed in the PseudoVertex category or in the RecoVertex category, the following variables serve as input to the likelihood discriminator:
  - **Vertex mass:** this variable is invariant mass of the charged particles associated to the secondary vertex. Secondary vertices originating from B-hadron decays are significantly more massive than e.g. those from D-hadrons.
  - **Number of tracks associated with the vertex:** jets with B-hadron decays show a significantly higher track multiplicity than e.g. those with D-hadron decays.
  - η_{rel} **of all tracks from the vertex:** for each track at the secondary vertex its pseudo-rapidity with respect to the jet axis η_{rel} is computed.
  - **Transverse impact parameter significance of the first track above the charm threshold:** all the tracks in the jets are ordered in decreasing transverse impact parameter significance and are added one by one to calculate the mass. The track exceeding the charm mass, 1.5 GeV$/c^2$, is selected and its impact parameter significance is used. This discriminator is particularly relevant for discrimination between bottom and charm quarks.
- Finally in the case there is no secondary vertex or there is a vertex in the PseudoVertex category or in the RecoVertex category, the following variables are calculated:

- **Track multiplicity:** the number of all selected tracks in the jet.
- **Three-dimensional impact parameter significance of all selected tracks:** the first three tracks are evaluated using dedicated pdf's whereas the additional tracks are evaluated using a generic pdf.

The likelihood ratio variable is computed by combining the individual variables

$$LR_{b\ vs.\ c,udsg} = \frac{\mathcal{L}_b}{\mathcal{L}_b + \mathcal{L}_{c,udsg}} \tag{4.7}$$

with

$$\mathcal{L}_{b,c,udsg} = \prod_i p_{b,c,udsg}(x_i) \tag{4.8}$$

where $p(x)$ is the probability density function. The likelihood ratio is evaluated twice, once with c jets as background and once with udsg jets as background. The relative composition of the three vertex categories is accounted for by transforming the likelihood ratio variable to

$$d_{b\ vs.\ c,udsg} = \frac{\widehat{\mathcal{L}}_b}{\widehat{\mathcal{L}}_b + \widehat{\mathcal{L}}_{c,udsg}} \tag{4.9}$$

with

$$\widehat{\mathcal{L}}_{b,c,udsg} = p_{b,c,udsg}(\text{category}) \cdot p_{b,c,udsg}(LR_{b\ vs.\ c,udsg}) \tag{4.10}$$

Finally the discriminator is calculated as $(0.75 d_{b\ vs.\ udsg} + 0.25 d_{b\ vs.\ c})$ which reflects the flavour composition of hadronic W boson decays in $t\bar{t}$ events. The distribution of the discriminator is displayed in Figure 4.12.

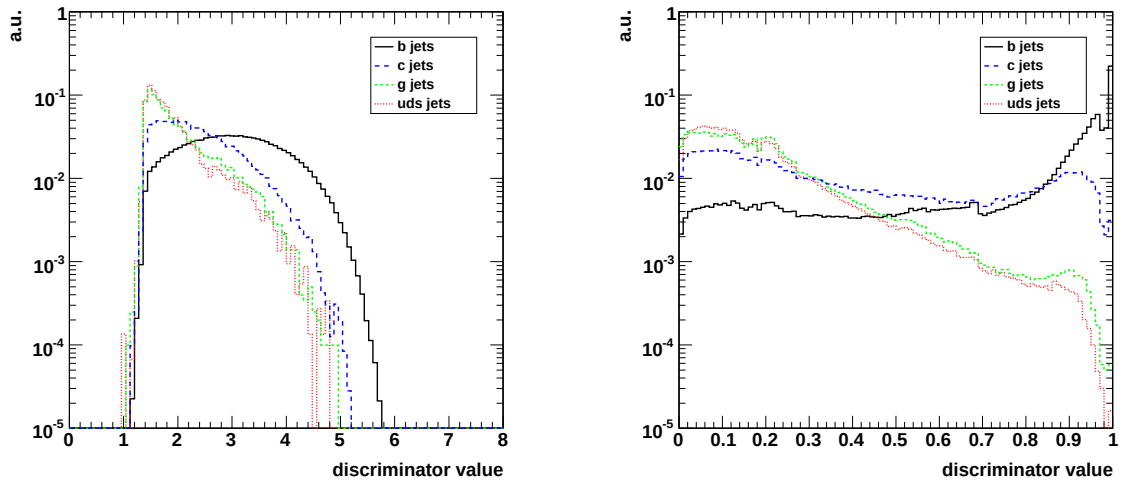

Figure 4.12: The discriminator value of the simple secondary vertex b-tagging algorithm (left) and the combined secondary vertex b-tagging algorithm (right) for $t\bar{t}$ events.

4.3.3 Soft lepton based b-tagging algorithm

The branching ratio for direct and cascade decays of B hadrons to electrons and muons is about 20 % as discussed in Section 3.4.2. This property is exploited in the soft lepton b-tagging algorithms [103] where the presence of soft electrons or muons is searched for in jets. For the electron based b-tagging algorithm, electrons are selected based on certain identification variables[4]. The most common sources of objects faking electrons are coming from charged hadrons with significant loss of energy in the electromagnetic calorimeter, neutral pions and photon conversions. For the soft muon based b-tagging algorithms a global muon candidate, as introduced in Section 4.1 used to select muon candidates. The main limitation for an efficient reconstruction of these low p_T muons is due to the strong bending of the muon in the magnetic field. The most important source of mis-identification of light quark jets as b quark jets comes from decays of $\pi^{\pm}$ and $K^{\pm}$ in light jets resulting in real muons inside a light quark jet.

Four variables differentiating soft leptons in b quark jets from observed leptons in other jets are exploited to construct the soft lepton b-tagging algorithm:

- $\mathbf{p_T^{rel}}$**:** the relative transverse momentum of the lepton w.r.t. the jet axis
- **The impact parameter significance:** the significance of the three-dimensional distance of closest approach of the lepton track to the primary vertex.
- **The distance** ΔR**:** the angular distance in (η, ϕ)-space between the lepton direction and the jet axis.
- **The momentum ratio:** the ratio between the momentum of the reconstructed track and the energy of the jet.

To build a discriminator these variables are combined with a neural network. The neural network is trained for b quark jets from $t\bar{t}$ events and light quark jets from QCD di-jet events. The distributions of the b-tag discriminator for the soft muon b-tagging algorithm and for the soft electron b-tagging algorithm are shown in Figure 4.13 for $t\bar{t}$ events.

4.3.4 Correlation between b-tagging algorithms

In Table 4.1 the difference between different b-tagging algorithms is shown. The difference is quantified as the fraction, in percent, of b quark jets that is tagged by b-tagging algorithm A (row) but not by b-tagging algorithm B (column). The fixed b-tag efficiency for this comparison was chosen to be 50% for each lifetime based b-tagging algorithm while for the soft muon b-tagging algorithm a b-tag efficiency of 15% was chosen. The four lifetime based b-tagging algorithms show differences of less than 20%, this indicates that they share a very large fraction of tagged b jets as expected since they exploit the similar information. The soft muon b-tagging algorithm has a much smaller overlap, the last row indicates that 50% of the jets that are tagged by the soft muon b-tagging algorithm are not tagged by the lifetime based b-tagging algorithms.

In Figure 4.14 the correlation for b quark jets in $t\bar{t}$ events is shown comparing following discriminator values: the high efficiency track counting variable, the simple secondary vertex

[4] More details can be found in [103].

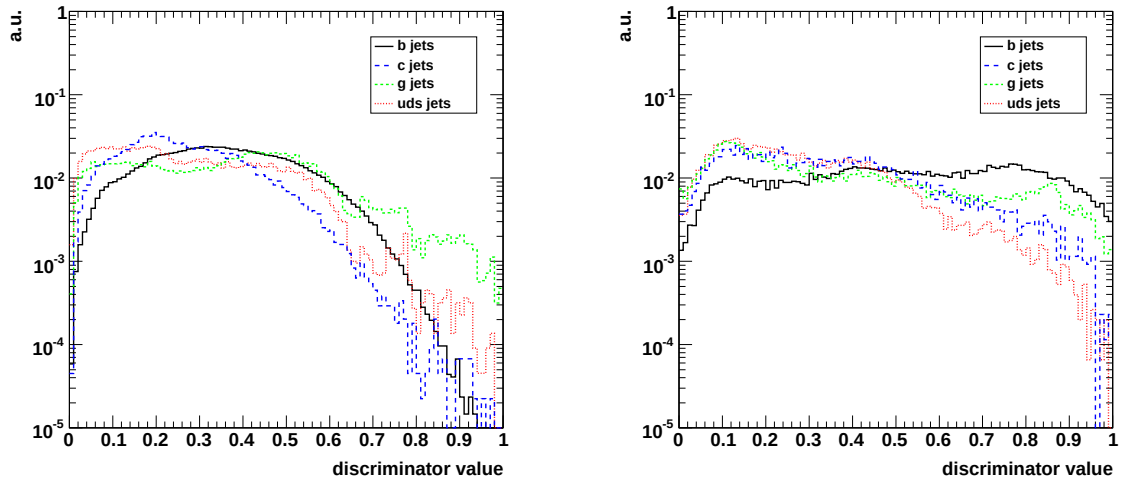

Figure 4.13: The discriminator value of the soft muon b-tagging algorithm (left) and the soft electron b-tagging algorithm (right) for $t\bar{t}$ events.

A B → ↓	Track counting	Track prob.	Simple sec. vtx	Combined sec. vtx.	Soft μ
Track counting	-	12.0	17.8	18.2	85.0
Track probability	14.0	-	18.1	16.2	85.0
Simple sec. vtx.	19.7	18.2	-	15.7	84.3
Combined sec. vtx.	17.4	13.4	12.8	-	84.4
Soft μ	48.8	50.0	45.0	47.3	-

Table 4.1: Difference in b-tagging algorithm yield for the track counting, track probability, simple secondary vertex, combined secondary vertex and soft muon b-tagging algorithms.

variable, the combined secondary vertex variable and the soft muon variable. A large correlation is found between the lifetime based b-tagging algorithms as expected from the observations in Table 4.1. The linear correlation coefficient between the high efficiency track counting b-tag variable and the soft muon b-tag variable is 0.20 reflecting the small correlation between the lifetime based algorithms and the soft lepton based algorithms. This low correlation can be exploited by combining several b-tagging algorithms to construct more performant b-tagging algorithms.

4.3.5 Performance of b-tagging algorithms

b-tag efficiency vs. light mistag rate

The output of each b-tagging algorithm is a single discriminator value for a given jet in the event. This discriminator value is used to test the hypothesis that the given jet is indeed coming from a b quark. In the hypothesis test, two types of mistakes can be made. First of all, the jet can be identified as a light jet while it was originating from a b quark. The second mistake is made when a jet is identified as a b jet while is was not originating from a b quark but rather from a charm quark or a lighter quark. The first mistake defines the efficiency of

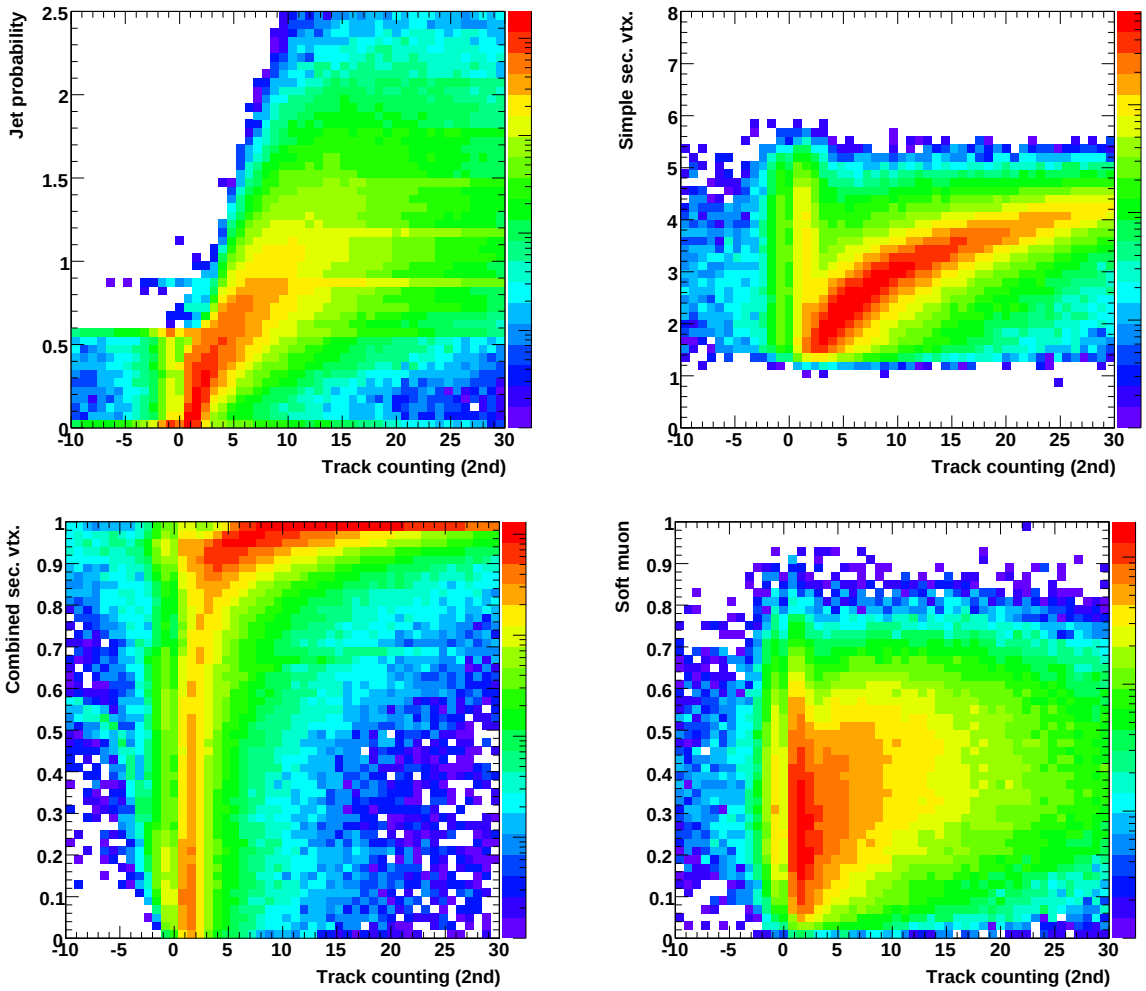

Figure 4.14: Correlation between the track counting variable and the jet probability variable (upper left), the simple secondary vertex variable (upper right), the combined secondary vertex variable (lower left) and the soft muon variable (lower right). Only b quark jets in $t\bar{t}$ events have been considered.

the b-tagging algorithm while the second mistake defines the mis-identification or mistag rate of the algorithm. To apply the hypothesis test a cut on the b-tag discriminator needs to be chosen defining a working point. Working points can, broadly speaking, be categorized as *loose*, when a loose cut is applied on the discriminator, to obtain a high efficiency by not rejecting too much b jets but as well yielding a relative high mistag rate. A *tight* working point can be adopted when a low mistag rate is required, therefore requiring a hard cut on the discriminator leading thus to a lower efficiency.

The b-tagging performance is exploited by scanning a cut over the full range of the b-tag discriminator. In Figure 4.15 the performance of the b-tag efficiency is shown as function of the uds-, gluon- and c-mistag rate, for the high efficiency track counting b-tagging algorithm, the simple secondary vertex b-tagging algorithm, the combined secondary vertex b-tagging algorithm and the soft muon b-tagging algorithm. The performance is obtained in $t\bar{t}$ events and the flavour was assigned to the jets according to the physics definition. The difference between the mistag rate of jets originating from uds quarks and the jets from

gluons comes from the flavour definition. The performance for gluon jets is worse since gluon-splitting into $b\bar{b}$-pairs of $c\bar{c}$-pairs might occur in these jets. The simple secondary vertex b-tagging algorithm has a maximal b-tag efficiency of about 70% reflecting the efficiency to reconstruct the secondary vertex in the jets. The maximum b-tag efficiency of about 20% for the soft muon b-tagging algorithm is due to the branching fraction of the B-mesons in soft muons.

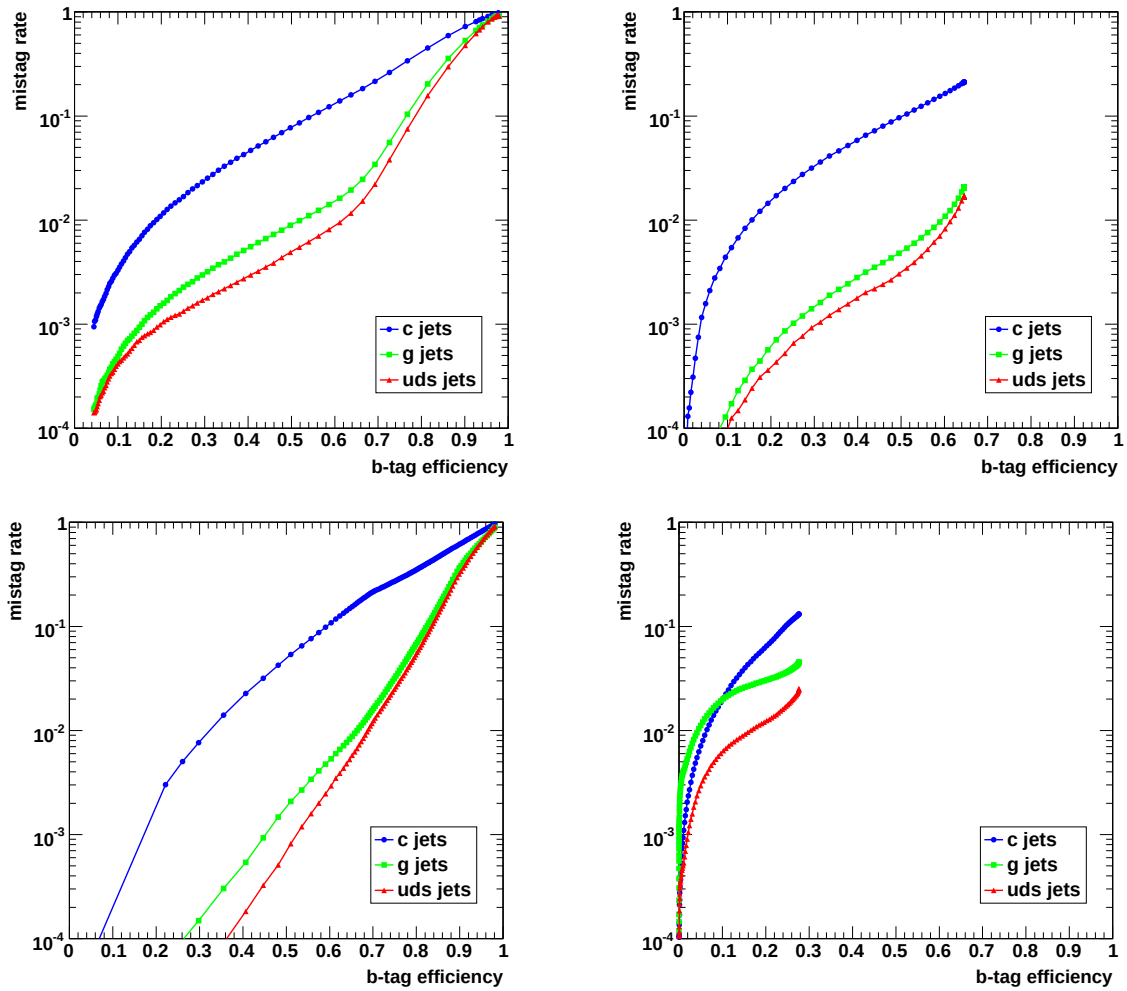

Figure 4.15: Performance of the high efficiency track counting b-tagging algorithm (upper left), the simple secondary vertex b-tagging algorithm (upper right), the combined secondary vertex b-tagging algorithm (lower left) and the soft muon b-tagging algorithm (lower right). A separate mistag rate is calculated for c quark jets, uds quark jets and gluon jets in $t\bar{t}$ events

In Figure 4.16 the mistag rate for respectively uds quark jets and c quark jets as a function of the b-tag efficiency is shown for all b-tagging algorithms introduced in the previous sections in $t\bar{t}$ events. An increasing performance is observed when going from impact parameter significance algorithms towards more complex algorithms like the combined secondary vertex algorithm which combines the information of displaced tracks and the reconstructed secondary vertex. The soft lepton algorithms exhibit a lower performance but

they are based on different aspects of b jets which makes them particularly interesting to combine with the lifetime based algorithms. In the performance plots listed here a perfectly aligned detector was assumed, in [101] the performance degradation was studied in multiple non-ideal detector scenarios reflecting several phases of the CMS experiment.

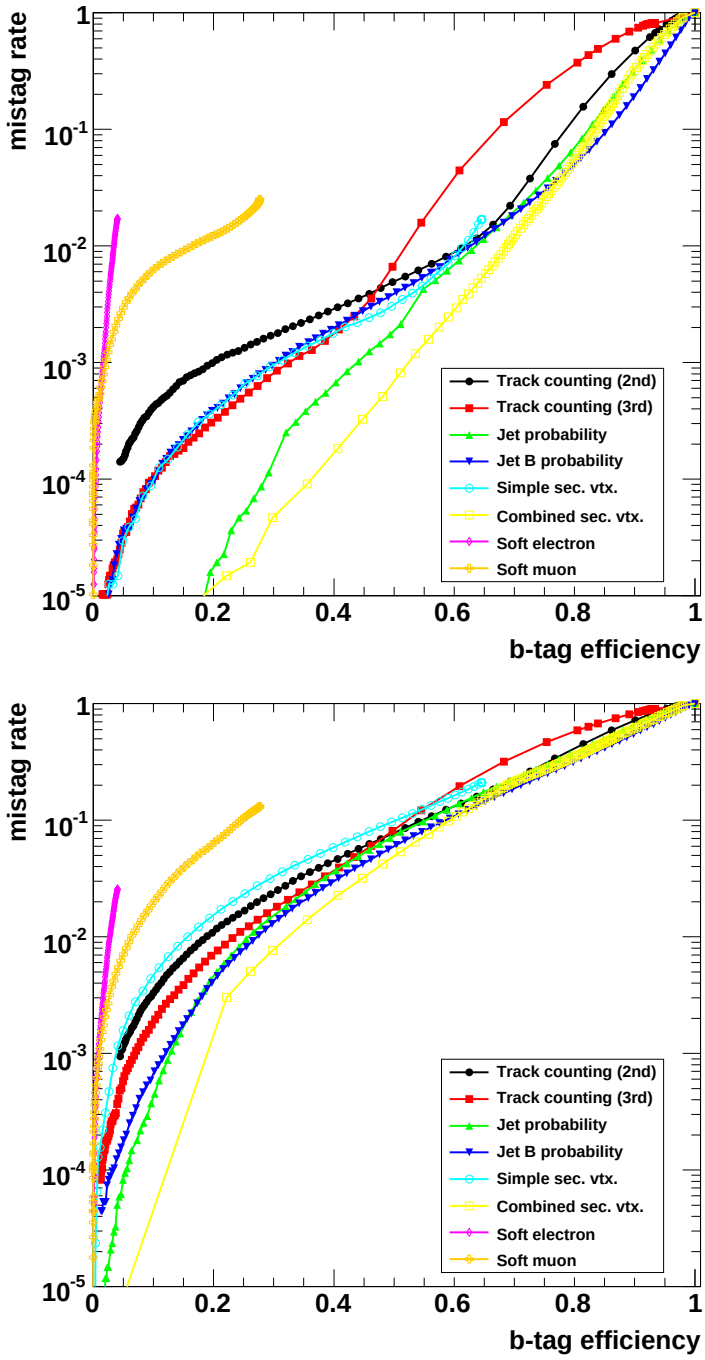

Figure 4.16: uds-mistag rate (upper) and c-mistag rate (lower) as a function of the b-tag efficiency for all b-tagging algorithms, evaluated in $t\bar{t}$ events.

Performance dependency on the kinematic properties of a jet

The performance of b-tagging algorithms depends on the kinematic properties of the jets. In Figure 4.17 the dependency of the performance on the transverse momentum and the pseudo-rapidity of the jet is shown for $t\bar{t}$ events. The b-tag efficiency in both figures is obtained for the track counting high efficiency algorithm at a fixed cut on the discriminator value yielding an average b-tag efficiency of 50% for all jets. The decrease in b-tag efficiency for jets with low transverse momentum is mainly due to multiple scattering leading to a degraded resolution of the impact parameters of the tracks and a worse reconstruction of the direction of the jet. The decrease in efficiency for jets with a high transverse momentum is due to a higher track multiplicity in these jets. This leads to a higher occupancy in the tracker, resulting in a degraded performance of the trajectory reconstruction. The b-tag efficiency degrades with increasing pseudo-rapidity due to a large material budget leading to a worse resolution on the impact parameter of the tracks. The performance dependency of the b-tag efficiency on the kinematic properties of the jets can be obtained with the method developed in this thesis.

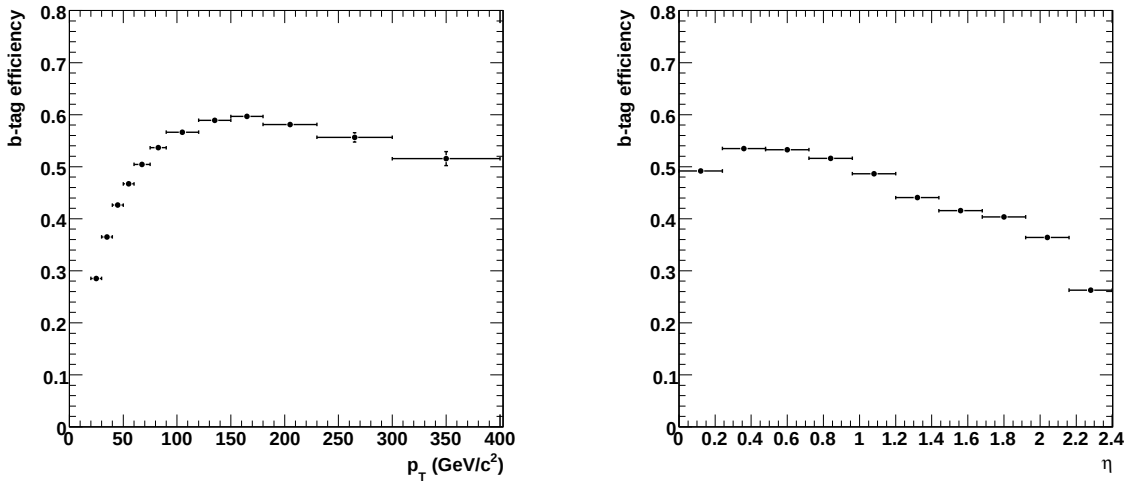

Figure 4.17: The dependency of b-tag efficiency on the transverse momentum (left) and the pseudo-rapidity (right) of b quark jets at an average b-tag efficiency of 50% for the track counting high efficiency b-tagging algorithm.

Performance dependency on the jet reconstruction algorithm

The default jet clustering algorithm used in this thesis is the seedless infrared safe cone algorithm. To test the dependency of the b-tag efficiency on the considered jet reconstruction algorithm, the b-tag performance was compared between the default algorithm and the inclusive k_T jet reconstruction algorithm. For the SC jets an opening angle of R=0.5 was used while for the KT algorithm the R-parameter was set to 0.4. Figure 4.18 shows the difference in efficiency versus mistag rate for the track counting high efficiency b-tagging algorithm respectively for uds quark jets and for c quark jets. The relative difference between the mistag rate for the SC algorithm and the KT algorithm is shown as well and only small performance differences are observed.

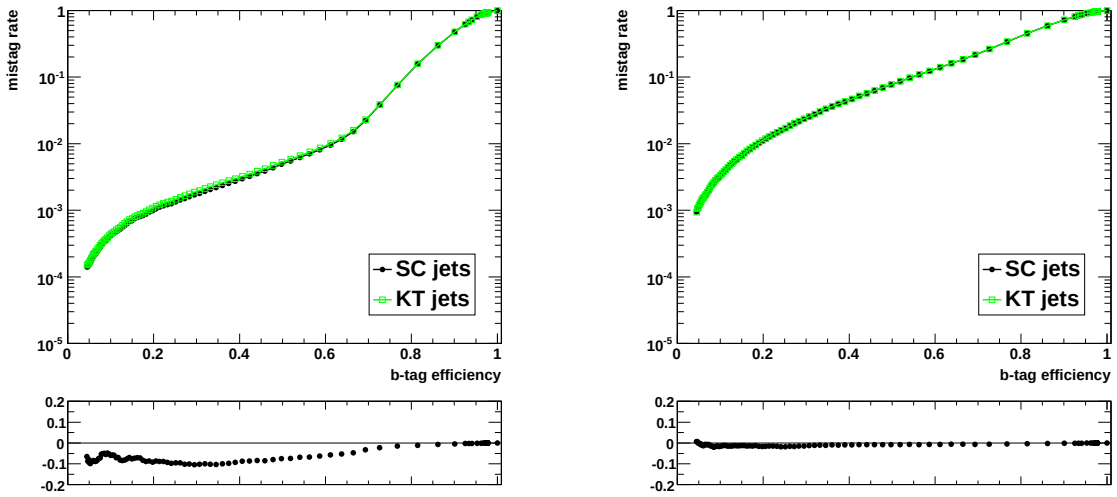

Figure 4.18: uds-mistag rate (left) and c-mistag rate (right) as a function of the b-tag efficiency for the track counting high efficiency b-tagging algorithm, evaluated in $t\bar{t}$ events. The relative difference between the two jet algorithms is shown as well.

Performance measurements using data

Several methods to measure the b-tag efficiency or the mistag rate with data collected by the CMS detector have been developed. These methods are complementary and can be compared and used for cross checks.

- **Mistag rate using negative tags.** The mistag rate in multi-jet events can be determined using tracks with negative impact parameters [104] as follows

$$\epsilon_{data}^{mistag} = \epsilon_{data}^{-} \cdot R_{light} \ , \tag{4.11}$$

  where ϵ_{data}^{-} is the negative tag rate in multi-jet data and $R_{light} = \epsilon_{MC}^{mistag}/\epsilon_{MC}^{-}$ is the ratio between the mistag efficiency of udsg jets and the negative tag rate of all jets in the simulated events. The measurement of the mistag rate is sensitive to the fractions of c and b quark jets in the jet sample with negative tag. A positive tag veto can be applied to significantly reduce these fractions; jets with a negative tag can be rejected if they have any track with impact parameter significance exceeding $IP/\sigma_{IP} > 4$. The udsg-mistag rates for a jet, depending on its transverse momentum and pseudo-rapidity can be obtained in data from,

$$\epsilon_{data}^{mistag}(p_T, \eta) = \frac{\epsilon_{data}^{-}(p_T, \eta)}{\epsilon_{MC}^{-}(p_T, \eta)} \epsilon_{MC}^{mistag}(p_T, \eta). \tag{4.12}$$

- **b-Tag efficiency using jets containing muons.** Three methods based on multi-jet events that have at least two reconstructed jets and a non-isolated muon close to one of the jets have been developed [105]. The non-isolated muon is required to be closer than $\Delta R(\mu, jet) < 0.4$ to the jet. Two of these methods rely on a template fit to the distribution of the p_{Trel} variable that is defined as the transverse momentum of the muon relative to the direction of the total muon-jet momentum vector. A third method is not based on p_{Trel} templates but tries to solve a system of 8 equations.

  - The p_{Trel} *Method* relies directly on a fit to the p_{Trel} distribution of the muon before and after tagging the muon-jet. The p_{Trel} distribution of the muons is fitted with a linear combination of the b quark jet template and the other jet template. The process is repeated after tagging the muon-jet. The b-tagging efficiency is calculated as the ratio between the number of b jets before and after tagging, as determined by the p_{Trel} fit.
  - The *Counting Method* also relies on the p_{Trel} fit but uses additional information derived from data. It assumes that the jets not containing the muon in the sample are dominated by light jets, and that the average probability of tagging them can be estimated from light jets data sample with negative impact parameter.
  - The *System8 Method* does not rely on the p_{Trel} fit to extract the b quark jet content of the samples. It consists of solving a system of eight equations constructed from the total number of events in two samples with different b quark jet content, before and after tagging with two different b-tagging algorithms.

- **b-Tag efficiency from top quark events.** A method based on a jet sample with a highly enriched b quark jet content was developed in [27]. Both the semi-leptonic and di-leptonic $t\bar{t}$ decay channels are used in this method to measure the b-tag efficiency. After the event selection, the discrimination between good jet associations and combinatorial background is performed based on a likelihood ratio technique. The jet combination in the event with the highest combined likelihood ratio discriminator is chosen. To extract a b-enriched jet sample from the selected events in the semi-leptonic $t\bar{t}$ channel hypothesis, the b quark jet coming from the leptonically decaying top quark is chosen because the b quark jet from the hadronically decaying top quark is required to be tagged. A selection cut on the combined likelihood ratio is applied to purify the jet sample in b quark jets. A b-tagging algorithm is applied on the selected jet sample and the b-tag efficiency ϵ_b is determined using

$$\epsilon_b = \frac{1}{x_b}\left[x_{tag} - \epsilon_o(1 - x_b)\right], \tag{4.13}$$

  where x_{tag} is the number of jets which are tagged in the selected jet sample, ϵ_o the mistag rate and x_b the expected fraction of b quark jets in the selected jet sample. The purity of the selected jet sample is therefore to be estimated from simulation, while all other parameters can be obtained from true data collisions.

Chapter 5

Event selection and topology reconstruction

In the previous chapter the reconstruction algorithms for the physics objects in the final state of a proton collision, observed by the CMS detector, were introduced. In the first section of this chapter the event selection carried out by applying cuts on the properties of these objects is discussed. The event selection has a dual purpose. The first goal is to select the $pp \rightarrow t\bar{t}$ events among the enormous number of background events. In Section 5.1.1 an overview is given of the selection criteria applied in this thesis, while in Section 5.1.2 the efficiencies of the different event selection criteria are listed for the relevant samples.

The second aim of the event selection is to select the correct objects to reconstruct the semi-muonic $t\bar{t}$ final state topology. The final state particles of the semi-muonic $t\bar{t}$ final state topology, $t\bar{t} \rightarrow bqqb\mu\nu_\mu$, are two b quarks from the electroweak decay of the two top quarks together with two light quarks from the decay of one of the two W bosons and a muon and a neutrino from the decay of the other W boson. Once four jets have been selected to represent the four quarks in the semi-muonic $t\bar{t}$ events an assignment needs to be made to link each observed jet to its initiating quark. In Section 5.2.1 the efficiency to select the four jets originating from the four quarks, among all observed jets in the final state, is discussed. In Section 5.2.2 an algorithm is introduced to assign the jets to the partons and in Section 5.2.3 the corresponding matching efficiency is studied.

5.1 Selection of the semi-muonic $t\bar{t}$ events

The event selection criteria, introduced in this section, are applied on the objects reconstructed in each event. The selected muons are required to be global muons while the jets are reconstructed with the Seedless Infrared Safe Cone jet algorithm, with a cone opening angle of R=0.5. The jets are calibrated with the relative and absolute (L2+L3) calibration factors. The detector simulation, for most of the samples used in this section, was performed using the fast simulation of the CMS detector. The efficiencies of the selection cuts are listed at the end of this section.

5.1.1 Selection criteria

The selection criteria are applied on the kinematic properties, the quality and the number of reconstructed objects in the events. In this section the event selection criteria for selecting the $t\bar{t}$ events in this thesis are applied sequentially on the available event samples. In a real data taking environment a proton collision event detected by the CMS detector is only stored when the event passes the trigger requirements. This trigger reduces the large rate of proton collisions to a manageable rate but it might happen that the trigger rejects potentially interesting events. It is therefore important to study the influence of the trigger on the event selection, this is done by applying the offline trigger after the event selection.

The initial skimming of the event samples requires at least 4 jets with $p_T > 15$ GeV/c and $|\eta| < 2.4$ and at least one global muon with $p_T > 15$ GeV/c and $|\eta| < 2.5$

To analyze the simulated samples PAT objects are constructed from the reconstructed objects which have been centrally produced by the CMS collaboration. In this process a preliminary event selection or skim is applied. This initial skim reduces significantly the disk space needed to store the event samples and, since only potential semi-muonic $t\bar{t}$ event candidates are stored, the time needed to analyze the remaining event samples is drastically reduced. This results in an important gain in time and computing resources when analyzing very large samples and is a technique that will also be applied on the enormous real data samples.

The event should contain at least one reconstructed, good quality, global muon with $p_T > 30$ GeV/c and $|\eta| < 2.1$

The semi-muonic $t\bar{t}$ events contain one muon with a large transverse momentum produced in the decay of a W boson. The collection of muons used to select the events is required to pass some quality cuts to reduce the number of fake muons. A muon is required to have a χ^2-value, normalized to the number of degrees of freedom, of the global muon fit smaller than 10. Additional to this criterion the muon is required to have more than 10 hits in the silicon tracker detector and a transverse impact parameter smaller than 2 mm with respect to the interaction point. All selected good quality muons in an event are ordered in descending transverse momentum. Figure 5.1 shows the transverse momentum distribution of the leading muon in this list. The semi-muonic $t\bar{t}$ events are denoted as '$t\bar{t}$ signal' while the events from other $t\bar{t}$ decay channels are denoted as '$t\bar{t}$ other'. The contribution of multi-jet ($pp \rightarrow X + \mu$ ($p_T > 15$ GeV/c)) background events is dominating therefore the transverse momentum distribution is shown again without this contribution. Muons in multi-jet events are mainly produced in the decay of heavy hadrons, therefore the spectrum of the transverse momentum of these muons is much softer than for muons produced from the decay of a W boson or a Z boson. The cut on the transverse momentum of the muon, $p_T > 30$ GeV/c, will reduce the multi-jet background strongly but additional criteria are needed to further reduce the contribution of background processes. The muons are required to fulfill $|\eta| < 2.1$ to be within the level-1 trigger acceptance.

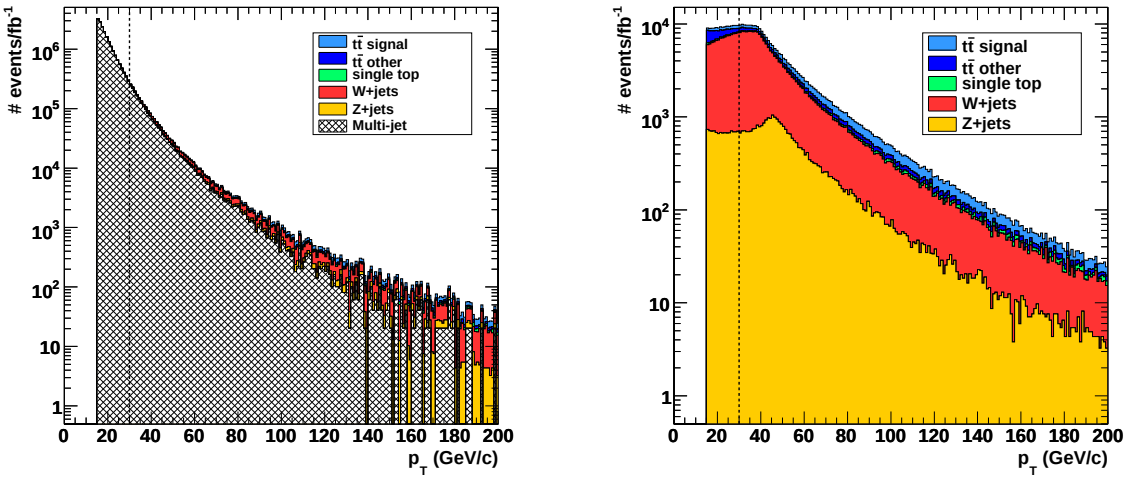

Figure 5.1: The transverse momentum of the leading muon in the event with (left) and without (right) the multi-jet background.

The event should contain at least four calibrated jets with $p_T > 30$ GeV/c and $|\eta| < 2.4$

To fully reconstruct the semi-muonic $t\bar{t}$ events at least four jets need to be present. In Figure 5.2 the distribution of the fourth jet in the event, ordered in descending p_T, is shown for various background processes and again the multi-jet background is overwhelming. Most of the background processes, such as W+jets and Z+jets have a softer p_T spectrum for the fourth jet. Therefore a cut on the transverse momentum, $p_T > 30$ GeV/c, of this jet will reduce their contribution. The jets are required to fulfill $|\eta| < 2.4$ to be within the tracker acceptance.

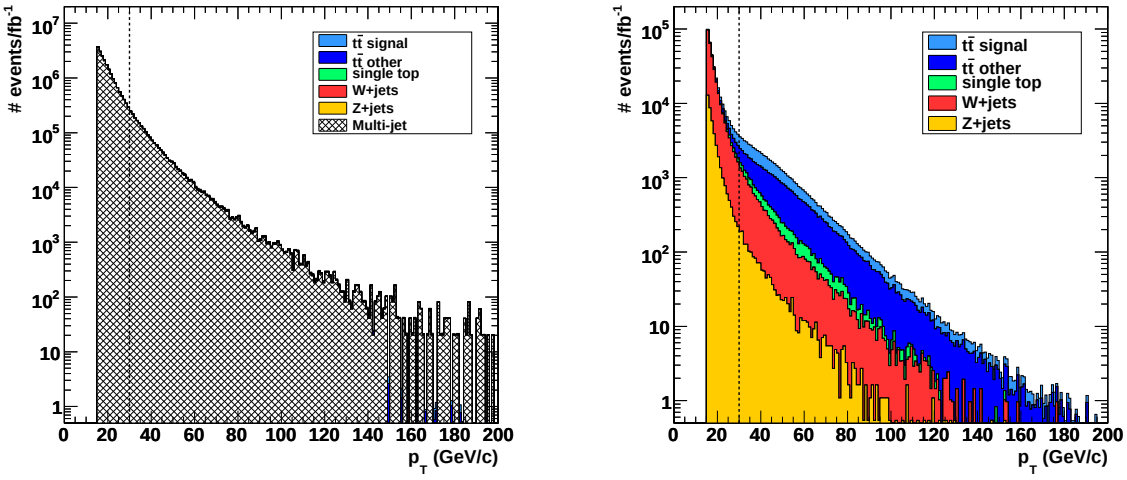

Figure 5.2: The transverse momentum of the fourth leading jet in the event, with (left) and without (right) the multi-jet background.

The event should contain exactly one isolated muon among the selected good quality muons

Muons originating from the decay of heavy hadrons, such as the muons in the large multi-jet background, are produced in the fragmentation process of the partons. Therefore they are contained inside a jet while muons coming from the decay of a Z boson or a W boson are in general isolated. To reduce further the enormous multi-jet background the isolation properties of the muon are exploited.

The first variable to distinguish isolated from non-isolated muons is the so-called relative isolation variable. The relative isolation of a muon is quantified by constructing an isolation cone with opening angle $\Delta R = 0.3$ around the muon direction at the vertex. The energy deposited in this cone in the electromagnetic and hadronic calorimeter is added, neglecting the energy deposits in a smaller cone around the muon to veto the energy of the muon itself. This veto cone has an opening angle of $\Delta R = 0.07$ in the electromagnetic calorimeter and $\Delta R = 0.1$ in the hadronic calorimeter. The veto cone is reconstructed around the direction of the muon at the corresponding calorimeter surface. Both cones are depicted in Figure 5.3, where the outer cone is the isolation cone and the inner cone is the veto cone.

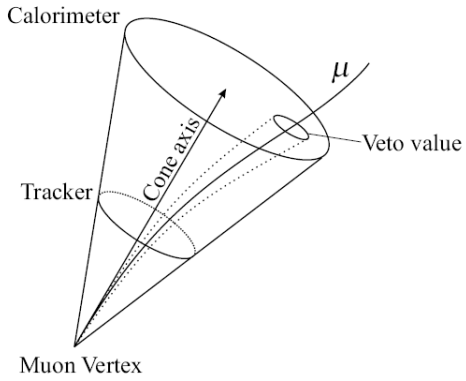

Figure 5.3: Isolation and veto cone around the reconstructed muon to calculate the relative isolation variable.

Based on the definition of the isolation cone and the veto cone the relative isolation of a muon is defined by the ratio

$$\frac{p_T^{\mu}}{p_T^{\mu} + \sum p_T^{tracks} + \sum E_T^{ECAL} + \sum E_T^{HCAL}}, \tag{5.1}$$

where p_T^{μ} is the transverse momentum of the muon and $\sum p_T^{tracks}$ is the sum of the transverse momenta of the tracks in the isolation cone leaving out the momentum of the muon track. The quantities $\sum E_T^{ECAL}$ and $\sum E_T^{HCAL}$ are the sum of the transverse energies in respectively the electromagnetic and hadronic calorimeters in the isolation cone leaving out the deposits in the veto cone.

Figure 5.4 shows the distribution of the relative isolation of the selected muons in the different channels with and without the multi-jet contribution. The difference between the non isolated muons in the multi-jet background and the, mainly, isolated muons in the other

processes suggests the requirement that the relative isolation is larger than 0.95 to reduce the multi-jet background. Besides the relative isolation, two additional variables are used to further reduce the multi-jet background, being the energy deposited in the electromagnetic calorimeter in the isolation cone around the muon and the energy deposited in the hadronic calorimeter in the isolation cone around the muon. These variables are shown as well in Figures 5.4. By requiring the events to have an energy deposit in an isolation cone in the electromagnetic calorimeter to be less than 4 GeV and an energy deposit in a veto cone in the hadronic calorimeter to be less than 6 GeV additional events with non-isolated muons are removed.

In semi-muonic $t\bar{t}$ events only one isolated muon is expected, while in some background processes, such as Z+jets two isolated muons can be present. To reduce events with more than one isolated muon, a cut is applied on the number of isolated muons in the list of selected good quality, muons. Events are only accepted if exactly one muon passes the isolation criteria.

The event should be accepted by the HLT trigger

The soft muon HLT trigger bit `HLT_Mu9` is used in this thesis. It requires a level-3 muon with a transverse momentum exceeding $p_T > 9$ GeV/c. Also the transverse impact parameter between the muon and the beamspot is required to be less than 2 cm. The HLT trigger bit `HLT_Mu9` is based on the seed of the level-1 trigger bit `L1_SingleMu7` which requires a muon with a transverse momentum exceeding $p_T > 7$ GeV/c reconstructed in one of the subdetectors of the muon system. The level-1 trigger is found to reject an additional amount of about 6% of the semi-muonic $t\bar{t}$ events that pass all the event selection criteria.

5.1.2 Efficiency of the event selection

The performance of the event selection is quantified as a function of its efficiency to retain the signal events with respect to the obtained signal-to-background ratio. The result of the sequential event selection is summarized in Table 5.1 and Table 5.2. The first column displays the number of events expected for an integrated luminosity of 1 fb^{-1}, the cross section for each sample is given in Table 3.3 and Table 3.4 and are all NLO except for the single top and multi-jet samples where the LO cross section is used instead.

The largest reduction of the Z+jets and W+jets events comes from the requirement of at least four jets in the event. Both in the initial skim and by applying the requirement to have at least four jets with $p_T > 30$ GeV/c a large reduction is obtained while the fraction of semi-muonic $t\bar{t}$ events decreases much less. The largest background after the event selection comes from the W+jets events. The W+c+jets events and the V+qq+jets events have only a small fraction of events left after the event selection. These samples are not taken into account in the analysis since they contain events that are also contained in the W+jets and Z+jets samples and would thus be double counted.

Also shown in Table 5.2 is that the multi-jet samples are strongly reduced by the isolation criteria on the muon. Several of the QCD $\widehat{p}_T$ bins have no events left after applying all selection criteria, therefore an upper limit is given for the number of events remaining after applying all selection criteria. The number of selected events in the QCD $\widehat{p}_T$ bins and the $pp \rightarrow \mu + X$ multi-jet sample are comparable within their uncertainties.

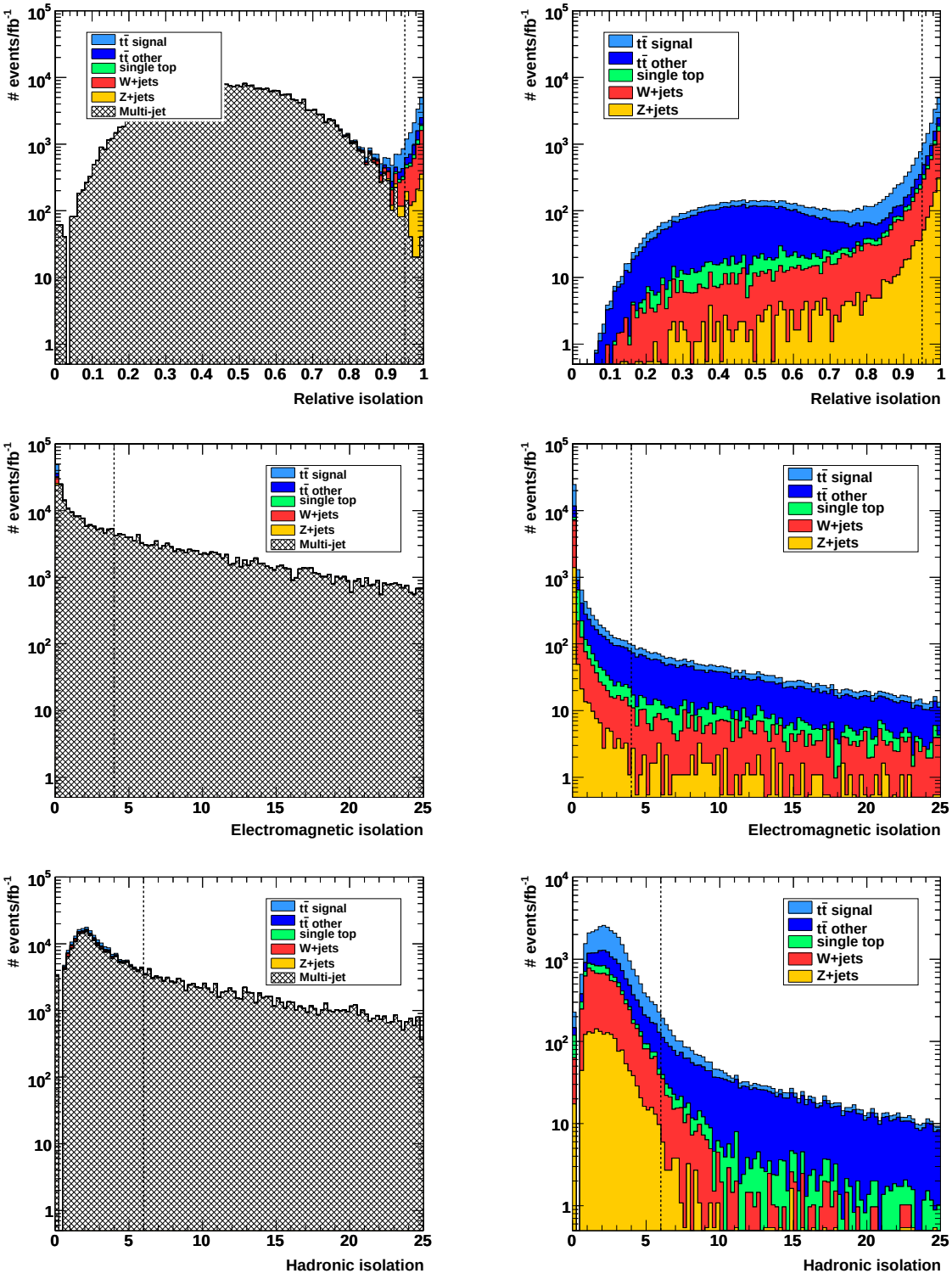

Figure 5.4: The relative isolation variable for the selected muons with (upper left) and without (upper right) the multi-jet background. The electromagnetic isolation variable for the selected muons with (upper left) and without (upper right) the multi-jet background. The hadronic isolation variable for the selected muons with (upper left) and without (upper right) the multi-jet background.

After applying the described event selection the expected number of semi-muonic $t\bar{t}$ events for an integrated luminosity of 1 fb^{-1} is 10249 $\pm$ 20 while the number of background events is 8566 $\pm$ 73 taking into account the $pp \rightarrow \mu + X$ multi-jet sample. The corresponding signal-to-background ratio is expected to be S/B =1.2. The multi-jet sample contributes very little to the background processes (about 1.4%) and is not considered further in this thesis.

	$t\bar{t}$ semi-μ	$t\bar{t}$ other	single top	Z + jets	W + jets
#evts.	61 k	352 k	72.2 k	4.2 M	45.6 M
skim	65.7 10^{-2}	11.5 10^{-2}	11.4 10^{-2}	10.3 10^{-3}	59.8 10^{-4}
muon p_T	46.2 10^{-2}	40.6 10^{-3}	56.8 10^{-3}	70.5 10^{-4}	33.7 10^{-4}
$\geq$ 4 jets	24.3 10^{-2}	21.0 10^{-3}	16.3 10^{-3}	29.3 10^{-5}	15.2 10^{-5}
muon iso.	17.8 10^{-2}	69.6 10^{-4}	91.8 10^{-4}	18.2 10^{-5}	11.0 10^{-5}
HLT	16.8 10^{-2}	66.0 10^{-4}	83.7 10^{-4}	17.5 10^{-5}	10.4 10^{-5}
sel. #evts.	10249 $\pm$ 20	2323 $\pm$ 9	663 $\pm$ 10	736 $\pm$ 20	4722 $\pm$ 48

Table 5.1: Overview of the cumulative event selection efficiencies after different steps of the event selection. The number of events before and after the event selection are normalized to represent an integrated luminosity of 1 fb^{-1}. The original size and cross section of each sample can be found in Table 3.3 and Table 3.4.

5.2 Reconstruction of the event topology

In the previous section the criteria to select the semi-muonic $t\bar{t}$ events are introduced. The cuts on the four leading jets and the leading isolated muon reduce the large multi-jet, W+jets and Z+jets background to a level that is manageable for performing the estimation of the b-tag efficiency as will be argumented in the next chapters.

The four quarks present in semi-muonic decaying $t\bar{t}$ events, $t\bar{t} \rightarrow bqqb\mu\nu_\mu$, are two b quarks originating from the decay of the two top quarks and two light quarks originating from the decay of one of the two W bosons. The four jets with highest p_T remaining after the event selection are assumed to represent these final state quarks. The aim of this section is to assign, in the selected events, the four leading jets to these four quarks. The jet associated to the b quark originating from the top quark with a hadronic decaying W boson is called the hadronic b quark jet, whereas the two jets associated to the light quarks originating from the hadronic decaying W boson are called the hadronic light quark jets. The jet associated to the b quark originating from the top quark with a leptonic decaying W boson is called the leptonic b quark jet.

In the following section the probability that the four leading jets are originating from the four quarks is discussed. This is followed by the introduction of an algorithm to match

	W+c+jets	V+qq+jets	$pp \to \mu + X$	QCD $\widehat{p}_T$ 100-250	QCD $\widehat{p}_T$ 250-500	QCD $\widehat{p}_T$ 500-1000	QCD $\widehat{p}_T$ 1000-∞
#evts.	1.5 M	290 k	122 M	15000 M	400 M	14 M	370k
skim	$11.5\ 10^{-3}$	$12.5\ 10^{-3}$	$18.2\ 10^{-2}$	$17.9\ 10^{-4}$	$22.6\ 10^{-3}$	$97.9\ 10^{-3}$	$24.6\ 10^{-2}$
muon p_T	$71.2\ 10^{-4}$	$77.2\ 10^{-4}$	$13.9\ 10^{-3}$	$45.9\ 10^{-6}$	$10.9\ 10^{-4}$	$49.7\ 10^{-4}$	$11.0\ 10^{-3}$
$\geq$ 4 jets	$18.8\ 10^{-5}$	$20.4\ 10^{-4}$	$28.1\ 10^{-4}$	$31.7\ 10^{-7}$	$37.8\ 10^{-5}$	$30.2\ 10^{-4}$	$87.0\ 10^{-4}$
muon iso.	$12.8\ 10^{-5}$	$12.0\ 10^{-5}$	$99.8\ 10^{-8}$	0	0	$65.5\ 10^{-8}$	0
HLT	$11.8\ 10^{-5}$	$11.3\ 10^{-5}$	$99.8\ 10^{-8}$	0	0	$43.7\ 10^{-8}$	0
sel. #evts.	177 ± 10	33 ± 3	122 ± 50	< 1220	< 79	6 ± 4	< 0.4

Table 5.2: Overview of the cumulative event selection efficiencies after different steps of the event selection. The number of event before and after the event selection are normalized to represent an integrated luminosity of 1 fb^{-1}. The original size and cross section of each sample can be found in Table 3.3 and Table 3.4.

each jet to a quark. This jet-quark matching algorithm is based on the top quark mass and W boson mass constraints in the observed $\text{t}\bar{\text{t}}$ events. Finally the performance of the jet-quark matching algorithm is discussed.

5.2.1 Selection performance of the four leading jets

Each event, passing the event selection criteria, contains at least four jets with a transverse momentum exceeding 30 GeV/ c. These four jets are now used to reconstruct the semi-muonic $\text{t}\bar{\text{t}}$ event topology. To study the performance of the choice of the four leading jets and, later on, the performance of the jet-quark matching algorithm, the true origin of these jets should be determined from the underlying information in the simulation. The originating quark of a jet is determined as its closest quark from the semi-muonic $\text{t}\bar{\text{t}}$ topology. To associate the four selected jets to the four quarks, they are first ordered in decreasing transverse momentum. The first jet is taken and the ΔR angle in (η, ϕ)-space between this jet and all four quarks is calculated. If this angle is smaller than ΔR<0.3, the jet is defined as originating from this quark. If more than one quark is closer than 0.3 to this jet, the closest quark is used. Once a quark is used to define the origin of a jet, it is removed from the list of available quarks and the next jet is selected. This continues until no jets remain in the list of four selected jets. It is possible that one or more jets have no associated quark. Jets not originating from any of the quarks are likely to have originated from radiation in the initial or final state. Initial state radiation jets are present among the four selected jets if they have a transverse momentum greater than one of the jets originating from the quarks. If final state radiation occurred with a radiated parton that leads to a sufficiently hard additional jet the two jets will not have the same direction as the initiating quark. It

might as well happen that one of the jets is not in the acceptance region of the detector, therefore the jet is not selected and a softer jet originating from radiation replaces the missing jet.

After applying the event selection and by taking the four leading jets the probability to have selected the four jets originating from the four quarks from the semi-muonic $t\bar{t}$ events is 25.3%. Thus in nearly three out of four events at least one of the four leading jets is originating from other sources than the $t\bar{t}$ decay. These jets are mainly coming from hard radiation in the initial and final state but could as well be originating from the underlying event. In the events where at least one jet is not matching to any of the quarks about $2/3$ has exactly one such a jet while about $1/3$ has two or more jets not originating from the $t\bar{t}$ decay. The amount of jets not originating from the quarks produced in the semi-muonic $t\bar{t}$-channel strongly limits the possibility to completely match the four leading jets to the quarks. The probability that at least one of the quarks from the semi-muonic $t\bar{t}$ decay is produced with a pseudo-rapidity $|\eta| > 2.4$ is 27.5 %.

Figure 5.5 shows the distributions of the transverse momentum and the pseudo-rapidity of the jets matching the hadronic light quarks, the hadronic b quark and the leptonic b quark in semi-muonic $t\bar{t}$ events. The events have been selected with the cuts introduced in the previous section leaving out the cut on the transverse momentum of the jets. The distributions of jets due to radiation in semi-muonic $t\bar{t}$ events are displayed as well. The tail of the distribution of the transverse momentum of radiation jets is longer than for the light quark jets and the b quark jets. The b quark jets have a slightly harder transverse momentum spectrum than the light quark jets while the radiation jets are in general less central than all jets originating from the quarks from the top quark decays.

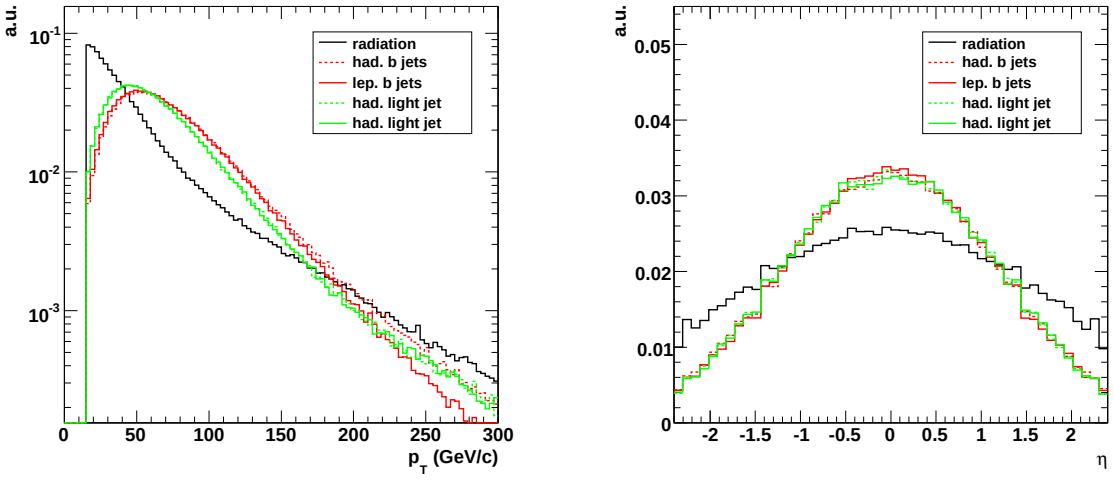

Figure 5.5: The transverse momentum (left) and pseudo-rapidity (right) of jets originating from the quarks from the semi-muonic $t\bar{t}$ events and from radiation.

The probability to select the four jets originating from the quarks in the semi-muonic $t\bar{t}$ events, when selecting the four jets with highest transverse momentum in the event, is shown in Figure 5.6 as a function of the p_T threshold applied on the four selected jets. The most optimal threshold is found to be around 30 GeV/c, supporting the current selection cut. Increasing the threshold on the selected jets results in a lower probability to select the

four jets originating from the quarks in the semi-muonic $t\bar{t}$ events due to the longer tail in the transverse momentum distribution of the radiation jets.

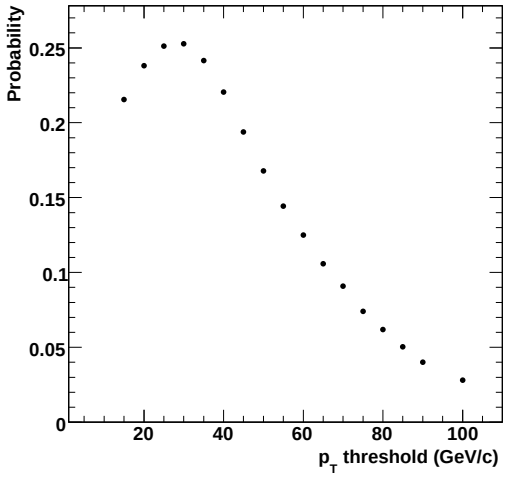

Figure 5.6: The probability to select the correct four jets originating from the quarks in semi-muonic $t\bar{t}$ events, when selecting the four jets with highest transverse momentum in the event.

5.2.2 Jet-quark matching algorithm

To perform the matching of the four selected jets to the quarks in semi-muonic $t\bar{t}$ events, the top quark mass and W boson mass constraints are used. The jet-quark matching algorithm searches for the jet configuration that maximizes the probability that the jets fulfill these mass constraints. There are a priori 24 possible configurations to match the jets to the quarks. This number of configurations is reduced to 12 since the interchange of the two jets from the W boson is not relevant in the analysis performed in this thesis. For each of the 12 possible configurations a χ^2-variable is calculated,

$$\chi^2 = \frac{(m_{j_1 j_2} - m_W)^2}{\sigma(m_W)^2} + \frac{(m_{j_1 j_2 j_3} - m_{top})^2}{\sigma(m_{top})^2}, \tag{5.2}$$

where m_W and $\sigma(m_W)$ are the W boson mass and its corresponding resolution and m_{top} and $\sigma(m_{top})$ are the top quark mass and its resolution. The combination with the minimal χ^2-value χ^2_{min} is in general the jet combination that most likely represents the semi-muonic $t\bar{t}$ event topology. The two jets, j_1 and j_2, are the hadronic light quark jets, jet j_3 is the hadronic b quark jet and the remaining jet j_4 in the collection of four selected jets is the leptonic b quark jet.

The W boson mass and the top quark mass are obtained using the hadronic light quark jets and the hadronic b quark jet from the selected events where the four jets are explicitly matching the four quarks. Figure 5.7 shows the reconstructed top quark mass and W boson mass. The value of the masses and the resolutions used in Equation 5.2 are obtained by

fitting a Gaussian function to the distributions. The parameters of the fit used in this thesis are indicated as well. Due to the dependence of the mass on the jet energy scale these values are preferred rather than the world average for the top quark mass and W boson mass. In this thesis jets have been corrected with the relative and absolute jet energy correction factors (cf. Section 4.2.2) and are found to not return the world average value of the top quark and W boson mass. In data the world average will be used and any possible deviation is absorbed in the systematic uncertainty due to the jet energy scale corrections.

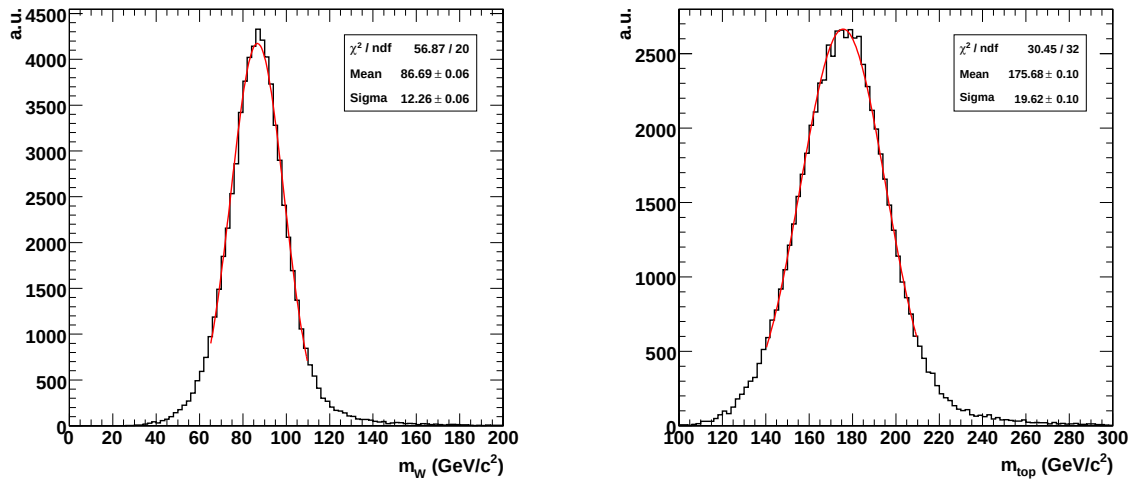

Figure 5.7: The W boson mass (left) and top quark mass (right) for events where the four selected jets are matched better than $\Delta R<0.3$ to the four quarks in the semi-muonic $t\bar{t}$ events. The distributions are fitted with a Gaussian function and the relevant parameters are indicated.

5.2.3 Performance of the jet-quark matching algorithm

To study the performance of the jet-quark matching algorithm several subsets of the simulated events are considered. In the first place the performance is studied for selected semi-muonic $t\bar{t}$ events, only if the four leading jets are originating from the four quarks in the semi-muonic $t\bar{t}$ event topology. This gives a view on the intrinsic performance of the algorithm. In the second place the effect of radiation jets in semi-muonic $t\bar{t}$ events is studied by considering all selected semi-muonic $t\bar{t}$ events. Finally, to obtain the performance of the jet quark matching algorithm, all selected events from the $t\bar{t}$ sample and the various background samples are considered.

Performance of the jet-quark matching algorithm for semi-muonic $t\bar{t}$ events where the four leading jets match the four quarks

In Section 5.2.1 it was found that the probability that the four leading jets in the selected semi-muonic $t\bar{t}$ events are originating from the four quarks is 25.3%, therefore only in these cases the jet-quark matching can return the correct configuration. Now, in the case a correct association exists, selecting the configuration with minimal χ^2_{min}-value returns in 44.5% of the events the correct jet-quark association, if an interchange of the two hadronic

light quark jets is considered irrelevant. The configuration with minimal χ^2-value is chosen and will be denoted as the best jet combination. The distribution of the χ^2_{min}-value of this best jet combination is shown in Figure 5.8 for events where the jets are correctly assigned to the quarks and for events where a wrong assignment is chosen by the jet-quark matching algorithm. In Figure 5.9 the probability to assign correctly the four jets is shown as a function of a cut on the χ^2_{min}-value. By requiring only events with a χ^2_{min}-value smaller than a threshold, a slight but negligible increase of the matching probability is found.

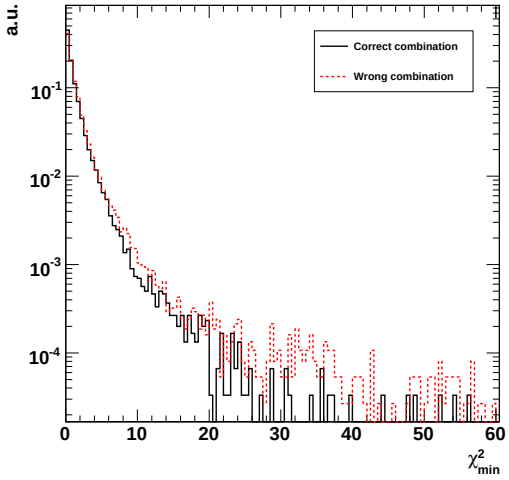

Figure 5.8: The χ^2_{min}-value of the chosen jet combination for semi-muonic $t\bar{t}$ events where the jet combination is correct and wrong.

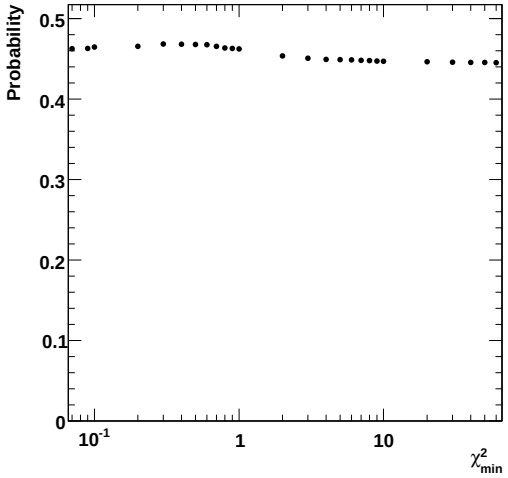

Figure 5.9: The probability that the jet-quark matching returns the correct jet configuration as function of a cut on the χ^2_{min}-value.

In the method to estimate the b-tag efficiency, introduced in the following chapter, a jet sample is selected with a sufficient fraction of jets originating from b quarks. This jet sample is obtained by selecting the jet labeled as the leptonic b quark jet in the best jet combination. Although the probability to match all four jets correctly is rather low, the probability that the leptonic b quark jet is correctly matched to its quark is higher and found to be 69.6%. In Figure 5.10 the probability that the leptonic b quark jet is correctly matched is shown as a function of a cut on χ^2_{min}. Although the probability to have assigned all four jets correctly increases by imposing a stronger threshold on the χ^2_{min}-value, a slight decrease in the probability to assign the leptonic b quark jet correctly is found. This is mainly due to the decreasing probability that events are present where the leptonic b quark jet is indeed correctly assigned but the other jets are wrongly assigned. Due to the presence of two b quarks in the semi-muonic $t\bar{t}$ event topology the chosen leptonic b quark jet could be wrongly assigned to the hadronic b quark returning a matching with a b quark as well. The probability that the leptonic b quark jet is matched to a b quark is 79.3 % and is shown as a function of a cut on the χ^2_{min}-value in Figure 5.10[1].

The probability that the matching algorithm assigns the leptonic b quark jet correctly to the leptonic b quark as a function of its transverse momentum and pseudo-rapidity is shown

[1]For the remainder of the thesis a b quark jet is defined as a jet that is matched to a b quark better than $\Delta R<0.3$.

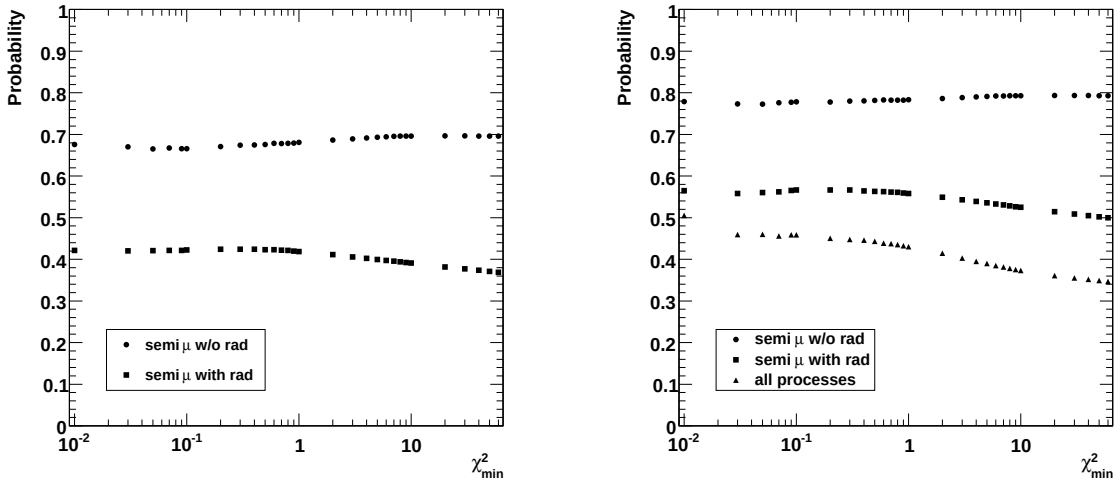

Figure 5.10: The probability that the leptonic b quark jet is correctly assigned to the leptonic b quark (left) or assigned to a generic b quark (right) as a function of a maximal threshold on the χ^2_{min}-value of the best jet combination. The probabilities are indicated for semi-muonic $t\bar{t}$ events where the four leading jets are originating from the four quarks, for all semi-muonic $t\bar{t}$ events, thus including events with at least one radiation jet among the four leading jets and all events including $t\bar{t}$ events and the various background processes.

in Figure 5.11. The decrease of the probability for jets with a small transverse momentum is due to the worse reconstruction of jets at low transverse momentum. In Figure 4.2.3 the resolution on the reconstructed transverse momentum of a jet is shown and a worse resolution is seen for low p_T jets. The probability to find the correct jet association does not depend on the pseudo-rapidity of the jet. The probability for the leptonic b quark jet to be matched to a b quark as a function of its transverse momentum and pseudo-rapidity is shown in Figure 5.12. The same tendency is observed as in Figure 5.11 but at a higher average probability.

Performance of the jet-quark matching algorithm for all semi-muonic $t\bar{t}$ events

Often one of the four leading jets is originating from radiation, therefore a full jet-quark matching is not possible. In Figure 5.13 the minimal χ^2-value over all combinations in the event is shown for the semi-muonic $t\bar{t}$ events where no radiation is present among the leading jets compared to semi-muonic $t\bar{t}$ events where at least one of the four leading jets is due to radiation.

Due to the presence of radiation jets among the four leading jets, the probability to correctly assign the leptonic b quark jet decreases to 35.4% while the probability that the leptonic b quark jet is originating from a b quark is now 48.3%. In Figure 5.10 the probability that the selected leptonic b quark jet is correctly matched and thus originating from the leptonic b quark or that the selected leptonic b quark jet is originating from a b quark is shown as a function of a cut on the χ^2_{min}-value. In both cases only a slight increase is found by imposing a more stringent threshold on the χ^2_{min}-value of the best jet combination.

In Figure 5.11 the probability that the leptonic b quark jet is originating from the leptonic b quark is given as a function of the transverse momentum and the pseudo-rapidity

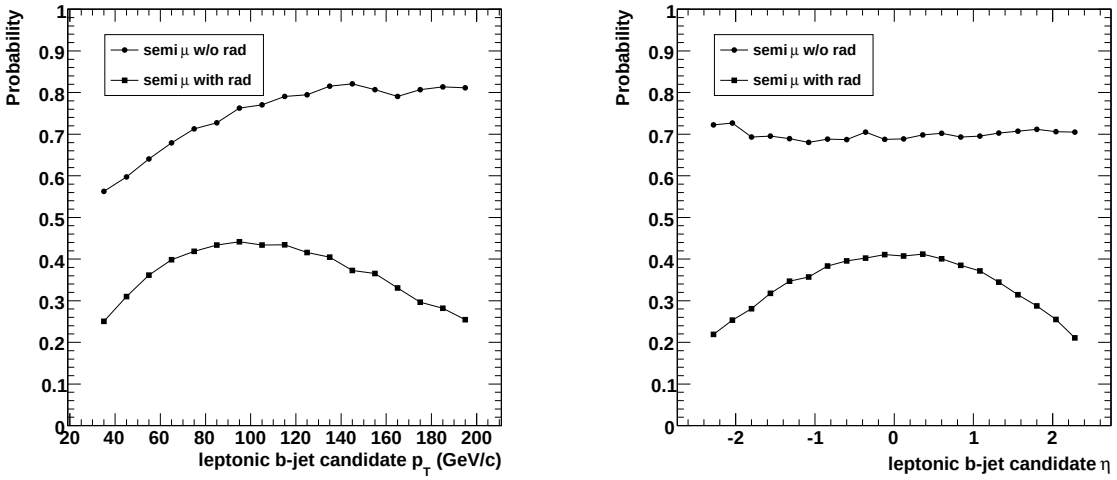

Figure 5.11: The probability that the leptonic b quark jet is originating from the leptonic b quark as a function of the transverse momentum (left) and pseudo-rapidity (right) of the leptonic b quark jet. The probabilities are indicated for semi-muonic $t\bar{t}$ events where the four leading jets are originating from the four quarks, for all semi-muonic $t\bar{t}$ events, thus including events with at least one radiation jet among the four leading jets.

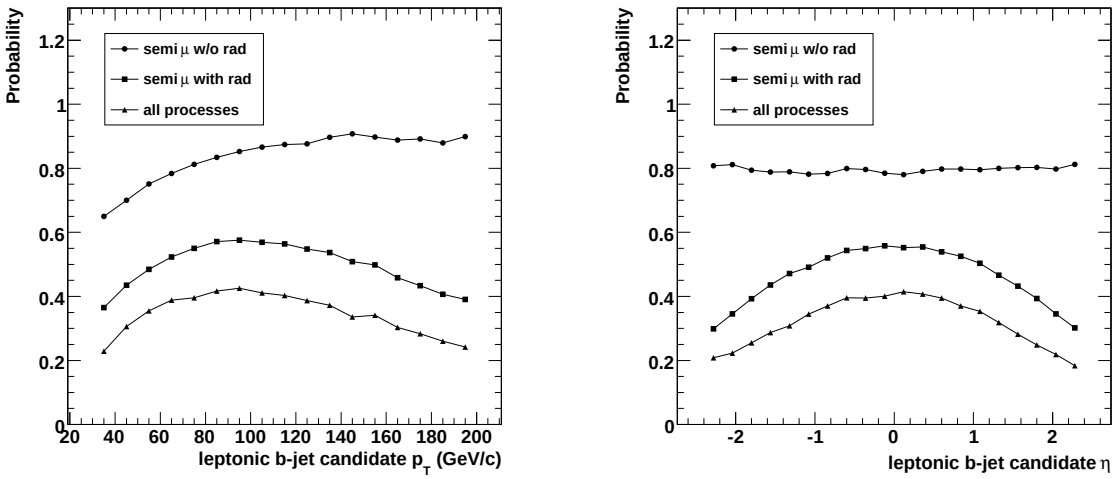

Figure 5.12: The probability that the leptonic b quark jet is originating from a generic b quark as a function of the transverse momentum (left) and pseudo-rapidity (right) of the leptonic b quark jet. The probabilities are indicated for semi-muonic $t\bar{t}$ events where the four leading jets are originating from the four quarks, for all semi-muonic $t\bar{t}$ events, thus including events with at least one radiation jet among the four leading jets and all events including $t\bar{t}$ events and the various background processes.

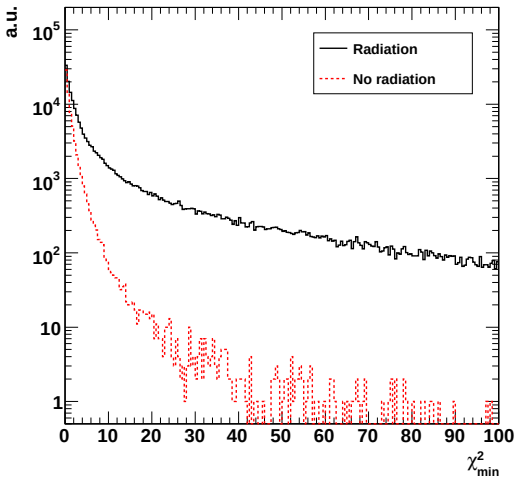

Figure 5.13: The χ^2_{min}-value of semi-muonic $t\bar{t}$ events where no radiation is present among the four leading jets and events where at least one jet is due to radiation.

of the leptonic b quark jet. An additional decrease at high transverse momentum is found w.r.t. the behavior for semi-muonic $t\bar{t}$ events where no radiation is present among the four leading jets. This is due to the longer tail in the transverse momentum distribution of jets originating from radiation (cf. Figure 5.5). In this case also the probability decreases at higher pseudo-rapidity values of the jet. This is due to the increased probability that a jet is originating from radiation at higher pseudo-rapidity (cf. Figure 5.5). In Figure 5.12 the probability is shown if the chosen leptonic b quark jet is originating from a b quark. The same shape is observed for the transverse momentum as well as for the pseudo-rapidity, but at a higher overall probability due to the mis-identification of the hadronic b quark.

Performance of the jet-quark matching algorithm for all events

The performance of the jet-quark matching algorithm has been studied up to now only including the semi-muonic $t\bar{t}$ events. To obtain a complete view on the performance of the algorithm in the calculation of the purity of the selected b quark jet sample, also the background processes need to be included. In this section the other $t\bar{t}$ channels, the single top processes, the W+jets process and the Z+jets process have been included as will be done to study the estimation of the b-tag efficiency. In Figure 5.14 the χ^2_{min}-value of the chosen jet combination is shown in semi-muonic $t\bar{t}$ events on one hand and in other $t\bar{t}$ events, single top quark events, W+jets events and Z+jets events on the other hand. The probability that the selected event is a semi-muonic $t\bar{t}$ event as a function of a cut on χ^2_{min} is shown in Figure 5.15 and an increasing probability is found for harder χ^2_{min} cuts as expected.

The probability that the leptonic b quark jet, selected to obtain a jet sample with sufficient b quark jets, is effectively originating from a b-quark is about 33% considering also the background events. In Figure 5.10 this probability as function of a cut on χ^2_{min} is shown. In Figure 5.12 the probability that the chosen leptonic b quark jet is originating

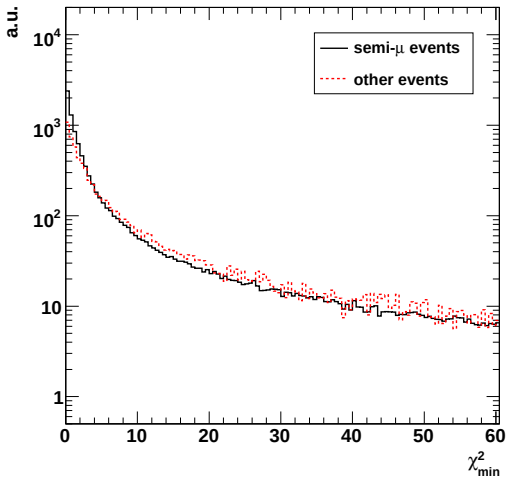

Figure 5.14: The distribution of the χ^2_{min}-value of the chosen jet combination for semi-muonic $t\bar{t}$ events and for all other background processes.

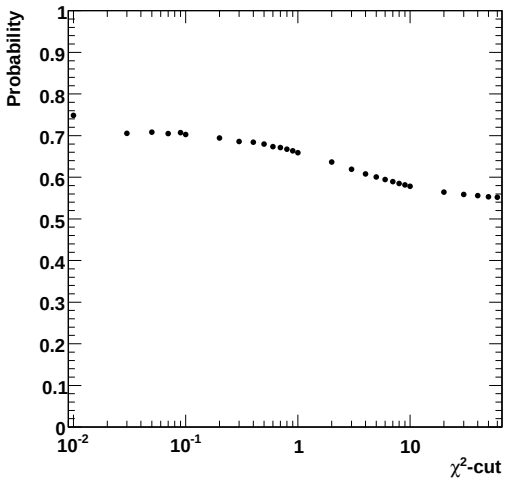

Figure 5.15: The probability that the selected event is a semi-muonic $t\bar{t}$ event w.r.t. all events as a function of a cut on the χ^2_{min}-value.

from a b quark is shown as a function of the transverse momentum and the pseudo-rapidity of the leptonic b quark jet. The shape of the distributions is comparable to the shape observed for all semi-muonic $t\bar{t}$ events.

Chapter 6

Inclusive estimation of the b-tag efficiency

The b-tagging algorithms, implemented in CMS, have been introduced in Section 4.3. Their performance, the b-tag efficiency versus the light jet mistag rate, has been studied on the basis of simulated events. In this chapter a method is developed to estimate the b-tag efficiency from semi-muonic $t\bar{t}$ events.

In Section 6.1 the method to estimate the b-tag efficiency, based on a jet sample with increased b quark jet content, is described. The effect of the contribution of non-b quark jets to this jet sample is addressed based on information from simulated events. In Section 6.2 the effect of this contribution is handled in a data driven way by selecting a control sample with a large fraction of non-b quark jets. Section 6.3 overviews the potential results when applying the method on an early set of data events. Section 6.4 describes in detail the treatment of the systematic uncertainties on the estimated b-tag efficiency. The result for various b-tagging algorithms is shown in Section 6.5.

6.1 Method to estimate the b-tag efficiency

In Chapter 5 the selection of $t\bar{t}$ events, in a data sample dominated by the enormous multi-jet background, is described. For the events retained after the event selection a kinematic algorithm to associate the four jets to the four quarks in the semi-muonic $t\bar{t}$ event topology is applied. The jet identified as the leptonic b quark jet is selected to form the b quark jet candidate sample. This sample will then be used to select two subsamples, one with an enriched b quark jet content and one with a depleted b quark jet content. With these samples the estimation of the b-tag efficiency is performed by explicitly reconstructing the b-tag discriminant distribution. The method developed in this section is exploited for the track counting high efficiency b-tagging algorithm and potential results are given for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV.

6.1.1 Selection of the b quark jet candidate sample

It was found in Section 5.1.2 that, after applying the selection cuts on the properties of the muon and the jets and after applying the trigger, the expected number of semi-muonic $t\bar{t}$

events is 10249 $\pm$ 20 while the expected number of background events is 8444 $\pm$ 53, for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV. On these remaining events a jet-quark matching algorithm, based on the mass of the hadronic top quark and the hadronic W boson, was applied. The combination with the minimal χ^2_{min}-value was chosen to assign the four leading jets to the four initial partons. Over all events, the jets assumed to originate from the b quark from the top quark with a leptonic decaying W boson, shortly the leptonic b quark jets, are now selected. This jet sample consisting of leptonic b quark jets is denoted as the b quark jet candidate sample or shortly the b candidate sample and is used to perform the estimation of the b-tag efficiency. It was found in Section 5.2.3 that the probability for the jets in this sample to effectively originate from a b quark is about 33%, setting the b quark purity of the b candidate sample.

To estimate the b-tag efficiency, based on the b candidate sample, two disjunct subsamples are created. One subsample needs to have an increased b quark purity while the other subsample is required to have a depleted b quark purity. To obtain these two subsamples a discriminating variable is defined. This discriminating variable is related to the mass of the leptonic top quark and is defined as the mass of the combined system of the b quark jet candidate and the selected muon. This variable is denoted as $m_{\mu j}$, and will be called the jet-muon mass. In Figure 6.1 the distribution of the jet-muon mass is shown for the various processes. The distribution of pseudo-data points, by randomly selecting events reflecting an integrated luminosity of 1 fb^{-1}, is indicated as well. The distribution is less peaked for background events such as W+jets and Z+jets than for $t\bar{t}$ events due to the absence of a top quark in the former two processes.

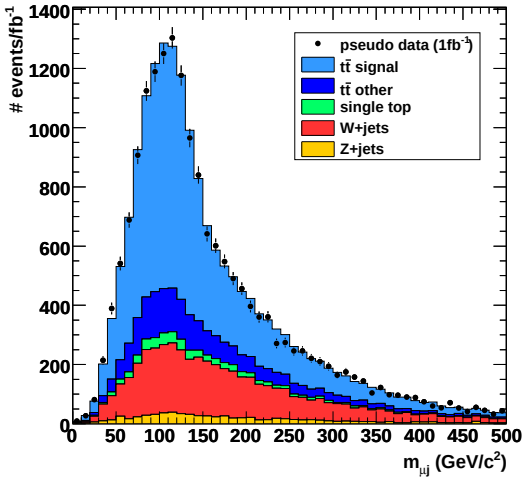

Figure 6.1: The distribution of the jet-muon mass, $m_{\mu j}$, for all processes.

The distribution of the jet-muon mass for both b quark jets and non-b quark jets to demonstrate the discrimination power of the jet-muon mass is shown in Figure 6.2. In the case that the candidate jet is originating from a b quark, the main contribution is coming from the semi-muonic $t\bar{t}$ events. The two other important contributions come from the other $t\bar{t}$ events and from the single top quark events. This is due to the presence of b quarks produced in the top quark decay in these samples. In the case that the b quark jet candidate is not originating from a b quark, additional contributions of W+jets events and, to lesser extent, of Z+jets events are present.

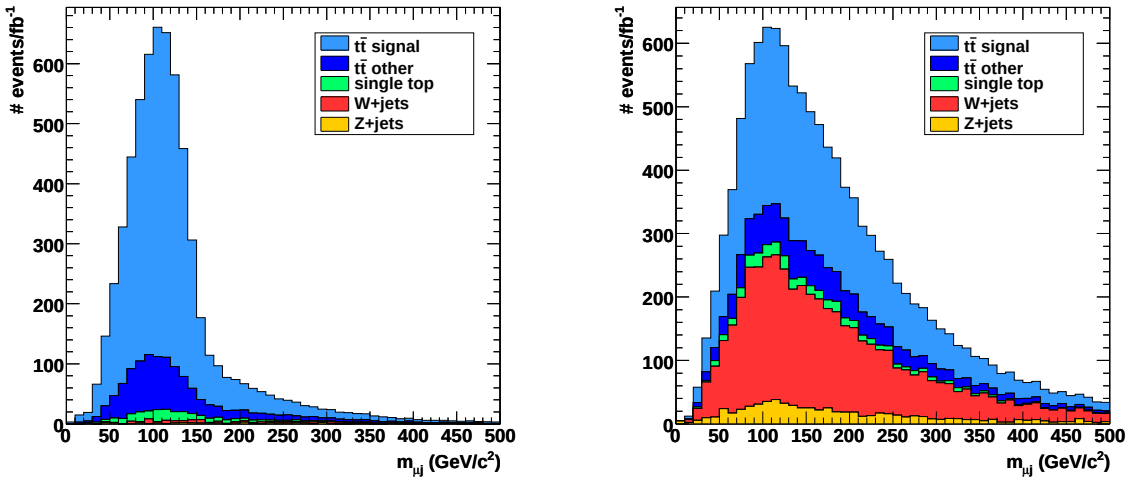

Figure 6.2: The distribution of the jet-muon mass in the case that the leptonic b quark jet is effectively originating from a b quark (left) and in the case that the leptonic b quark jet is not originating from a b quark (right).

A more detailed view on the distribution of the jet-muon mass is given in Figure 6.3. In this figure only the leptonic b quark jets, from the b candidate sample, effectively originating from a b quark in semi-muonic $t\bar{t}$ events are shown. The events where the leptonic b quark jet is matched incorrectly and is thus originating from the hadronic b quark are shown separately from the events where the jet is correctly matched and thus originates from the leptonic b quark. The fraction of b quark jets originating from the leptonic b quark is about 70% for semi-muonic $t\bar{t}$ events. The jets which are effectively originating from the leptonic b quark show a sharp cut-off in the $m_{\mu j}$ distribution while this cut-off is not present for jets originating from the hadronic b quark. This cut-off is around a mass of 160 GeV/c^2 as expected from Figure 3.5 where the $m_{\mu j}$-variable is shown for simulated leptonic b quarks.

In Figure 6.3 the distribution of $m_{\mu j}$ is also shown in the case where the b quark jet candidate does not originate from a b quark in semi-muonic $t\bar{t}$ events. The contribution of radiation jets is shown separately from the light quark jets originating from the hadronically decaying W boson. The shape of both distributions is rather similar except that for radiation jets a longer tail is observed reflecting the harder transverse momentum spectrum of radiation jets (cf. Section 5.2.1). The fraction of non-b quark jets from radiation is about 67% while the remaining fraction originates from W boson decays, $W \rightarrow q\bar{q}$, without significant final state radiation.

The distinct peak in the distribution of $m_{\mu j}$ for b quark jet candidates effectively originating from b quarks is exploited to select a b-enriched and a b-depleted subsample. The purity P_i as a function of the jet-muon mass is, for each bin i, calculated as

$$P_i = \frac{m^i_{\mu j}(\mathrm{b})}{m^i_{\mu j}(\mathrm{b}) + m^i_{\mu j}(\mathrm{non-b})} \tag{6.1}$$

where $m^i_{\mu j}(\mathrm{b})$ and $m^i_{\mu j}(\mathrm{non-b})$ are the jet muon mass for respectively jets originating from b quarks and jets not originating from b quarks.

In Figure 6.4 the bin-by-bin purity is shown and two regimes can be distinguished. The first regime is defined by jets in the b candidate sample which yield a jet-muon mass smaller

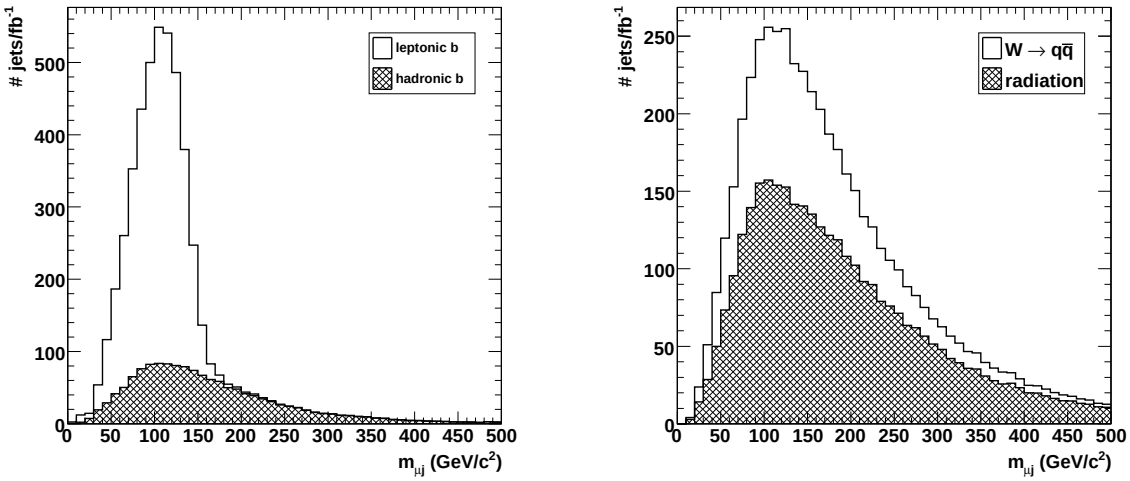

Figure 6.3: The distribution of the jet-muon mass for semi-muonic $t\bar{t}$ events where the leptonic b quark jets is originating from a leptonic or a hadronic b quark (left) or alternatively when the jet is not originating from any b quark but rather from the hadronic quarks from the W boson decay or from radiation (right).

than $m_{\mu j}$<160 GeV/ c^2. In this regime the b quark jet purity is higher than the average of the b candidate sample. The second regime is defined by jets which yield a jet-muon mass greater than $m_{\mu j}$>160 GeV/ c^2. The b quark jet purity in this regime is about 15%, significantly lower than the average of the b candidate sample. The presence of b quark jets in this high mass-regime is due to the combinatorial background in semi-muonic $t\bar{t}$ events and due to the b quark jets from other $t\bar{t}$ and single top quark events.

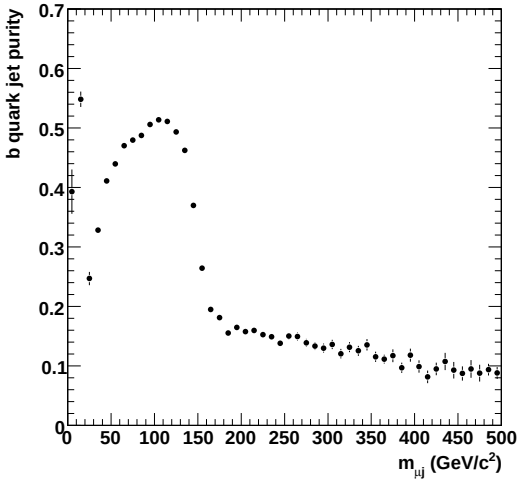

Figure 6.4: The b quark jet purity for all jets in the b quark jet candidate sample as a function of the jet-muon mass.

Based on the observed difference in b quark jet purity for jets in the two mass-regimes, a b-enriched and a b-depleted subsample can be constructed. These subsamples will be

used to obtain the b-tag discriminator distribution for pure b quark jets. To acquire an unbiased b-tag discriminator distribution, the b-tag discriminator should be uncorrelated to the jet-muon mass. In Figure 6.5 the mean of the b-tag discriminator is shown as a function of the jet muon mass for b quark jets and non-b quark jets. The mean was calculated from the positive b-tag discriminator values only. A correlation is seen between the b-tag discriminator and the jet-muon mass for non-b quark jets, due to the correlation between the mass and the transverse momentum, p_T, of the jet. The correlation between the p_T of b quark jets and the jet-muon mass is less due to the top quark mass constraint for the leptonic b quark jets.

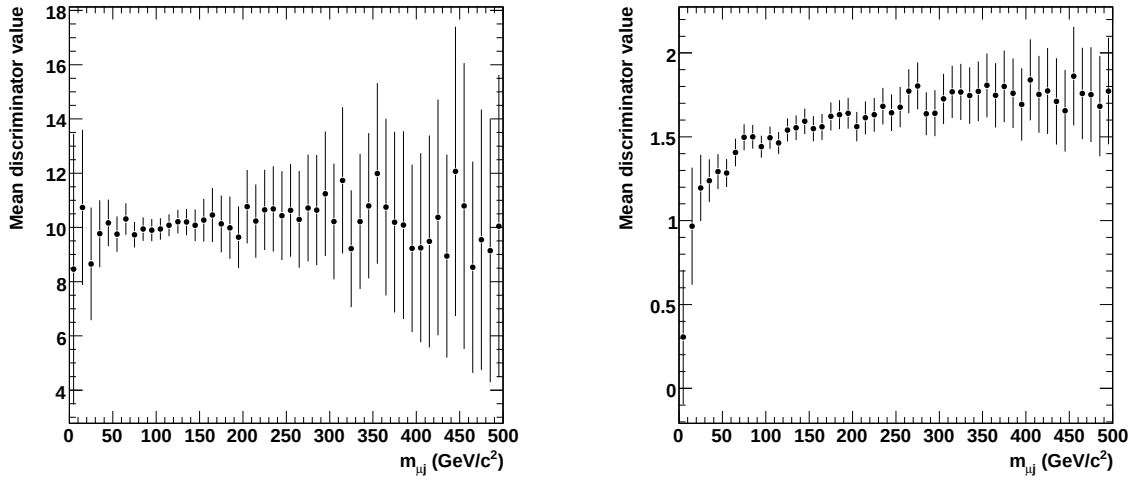

Figure 6.5: The mean of the positive b-tag discriminator distribution as a function of the jet-muon mass $m_{\mu j}$ for b quark jets (left) and non-b quark jets (right).

The b-enriched subsample is now defined by all the jets in the b quark jet candidate sample that satisfy the condition 90 GeV/ $c^2 < m_{\mu j} <$ 160 GeV/ c^2. The lower limit is set to reject the region with a strong correlation between the b-tag discriminator and $m_{\mu j}$ for the non-b quark jets. The upper limit of the b-enriched subsample is based on the cut-off value of the $m_{\mu j}$ peak for leptonic b quarks. The b-depleted subsample is constructed from the jets that satisfy the condition 160 GeV/ $c^2 < m_{\mu j} <$ 300 GeV/ c^2. The lower limit is set to the same value as the upper limit of the b-enriched sample whereas the upper limit of the b-depleted subsample is set rather arbitrary. The upper limit is defined to have approximately the same numbers of non-b quark jets in the b-depleted subsample as in the b-enriched subsample. In Table 6.1 the expected number of b quark jets and non-b quark jets in the b-enriched and b-depleted subsamples are given for an integrated luminosity of 1 fb^{-1}. The uncertainties on the expected number of events for 1 fb^{-1} reflect the limited number of events in the simulated samples. The b-enriched sample has a b quark jet purity of about 46% while in the b-depleted sample the b quark jet purity is about 16%.

6.1.2 Principle of the method

The method developed in this thesis aims to estimate the b-tag discriminator distribution of the b quark jets in the b-enriched subsample, denoted as $\widehat{\Delta}_b^{enr}$. Based on this estimated

	b-enriched subsample	b-depleted subsample
non-b quark jets	3994 ± 30	4161 ± 31
b quark jets	3451 ± 13	790 ± 8
purity	46 %	16 %

Table 6.1: The expected number of b quark jets and non-b quark jets in the b-enriched and b-depleted subsamples for an integrated luminosity of 1 fb^{-1}.

b-tag discriminator distribution, an estimation of the b-tag efficiency $\widehat{\epsilon}_b^{enr}$ at any given working point (w.p.) can be obtained.

The b-tag discriminator distribution of all b quark jets in the b-enriched subsample, Δ_b^{enr}, is assumed to represent the expected b-tag discriminator distribution of all b quark jets in the b quark jet candidate sample, Δ_b^{tot}, whether the jets are in the b-enriched or b-depleted region of $m_{\mu j}$. Figure 6.6 shows the b-tag efficiency ϵ_b^{tot} and ϵ_b^{enr}, corresponding to the expected b-tag discriminator distributions, as a function of a cut on the b-tag discriminator. The relative bias between the b-tag efficiencies in the b-enriched subsample ϵ_b^{enr} and the b-tag efficiency of all the jets in the b candidate sample ϵ_b^{tot}, is calculated as

$$\frac{\epsilon_b^{enr} - \epsilon_b^{tot}}{\epsilon_b^{tot}}. \tag{6.2}$$

The relative bias is found to be small and compatible with zero given the uncertainties reflecting the limited size of the simulated samples.

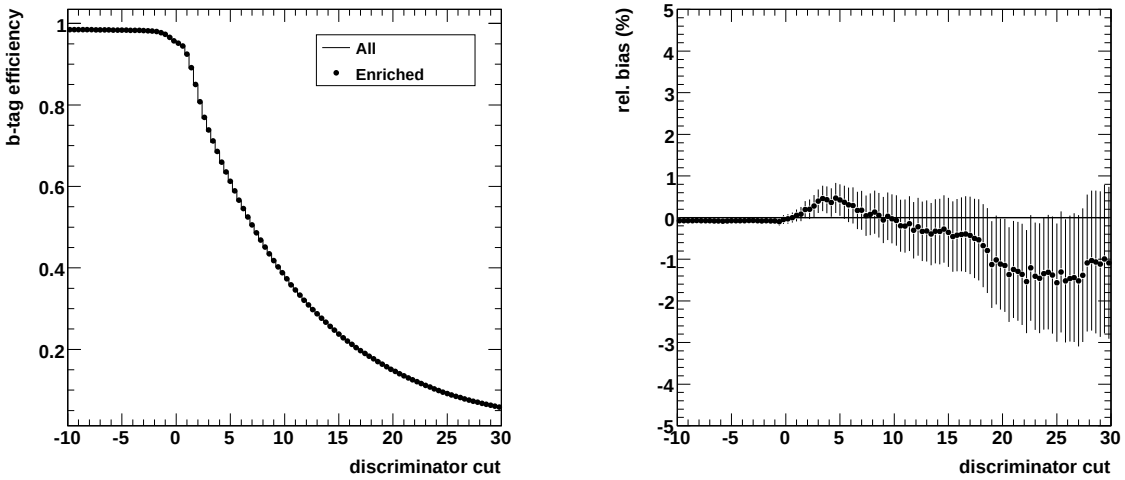

Figure 6.6: The distribution of the b-tag efficiency as a function of a cut on the b-tag discriminator value for all jets in the b quark jet candidate sample and for all jets in the b-enriched subsample (left) and the relative bias (right) between both b-tag efficiency distributions.

To estimate the b-tag discriminator distribution of b quark jets in the b-enriched subsample, $\widehat{\Delta}_b^{enr}$, the b-tag discriminator distribution of all jets reconstructed in the b-depleted subsample is subtracted bin-by-bin from the b-tag discriminator distribution of all jets in the b-enriched sample. This subtraction is carefully balanced to cancel the contribution of non-b quark jets to the b-tag discriminator distribution in the b-enriched subsample. The equation to obtain the content of each bin i of the estimated b-tag discriminator distribution is defined as

$$\widehat{\Delta}_b^{enr,i} = \Delta_{obs}^{enr,i} - F^{exp}.\Delta_{obs}^{depl,i}, \tag{6.3}$$

where Δ_{obs}^{enr} and Δ_{obs}^{dep} are respectively the observed b-tag discriminator distributions of all jets in the b-enriched and the b-depleted subsample. Both distributions are shown in Figure 6.7 and pseudo-data points corresponding to an integrated luminosity of 1 fb^{-1} are indicated. The contribution of b quark jets is shown separately and contributes for 46% in the b-enriched sample and for 16% in the b-depleted sample.

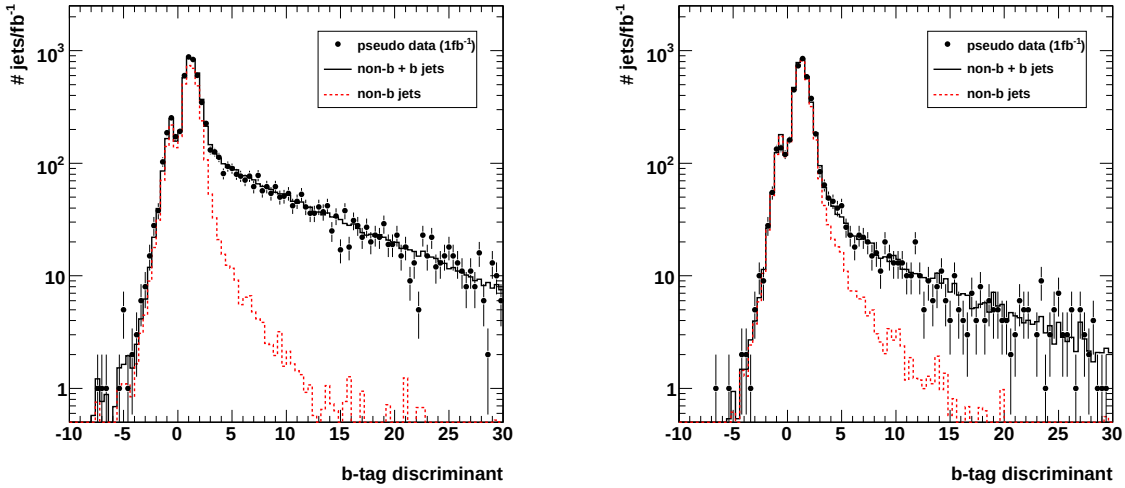

Figure 6.7: The distribution of the b-tag discriminator for all jets and b quark jets in the b-enriched (left) and the b-depleted (right) subsamples.

The subtraction factor or scale factor F in Equation 6.3 is defined in such a way that contribution of non-b quark jets in the b-enriched sample is canceled and is calculated as

$$F^{exp} = \frac{N_{non-b}^{enr}}{N_{non-b}^{depl}}. \tag{6.4}$$

For the b-enriched and b-depleted subsamples, defined in the previous section, the expected scale factor F^{exp}, obtained from the simulation, is equal to 0.960 $\pm$ 0.010 where the uncertainty is reflecting the limited size of the simulated samples. When now the b-depleted b-tag discriminator distribution Δ_{obs}^{dep} is subtracted from the b-enriched b-tag discriminator distribution Δ_{obs}^{enr}, the total amount of b quark jets in the b-enriched subsample is decreased due to the fraction of b quark jets present in the b-depleted subsample. The number of b quark jets expected in the b-enriched subsample after subtraction is expected to be 2693 $\pm$ 17 for an integrated luminosity of 1 fb^{-1}, compared to 3451 $\pm$ 13 before subtraction. In Figure 6.8 the resulting b-tag discriminator distribution $\widehat{\Delta}_b^{enr}$ is shown. The result is

compared to the expected b-tag discriminator distribution Δ_b^{enr} of the b quark jets in the b-enriched subsample. A reasonable agreement is found between both b-tag discriminator distributions except for a discrepancy at low b-tag discriminator values.

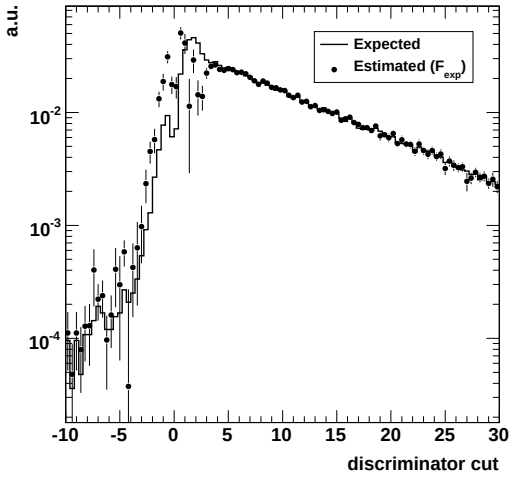

Figure 6.8: The estimated b-tag discriminator distribution for b quark jets compared to the expected b-tag discriminator distribution for all b quark jets in the b-enriched subsample.

The estimated b-tag discriminator distribution $\widehat{\Delta}_b^{enr}$ provides an estimation of the b-tag efficiency $\widehat{\epsilon}_b{}^{enr}$ at any given working point, defined by a cut on the b-tag discriminator value. In Figure 6.9 the estimated b-tag efficiency is shown as a function of any given cut on the b-tag discriminator value. Typically a working point for the track counting high efficiency b-tagging algorithm is chosen to be larger than zero. The estimated efficiency is compared to the expected efficiency Δ_b^{enr} obtained from all b quark jets in the b-enriched subsample and a reasonable agreement is found except for loose cuts on the b-tag discriminator. A more detailed view on the discrepancy between the estimated efficiency and the expected efficiency is given in Figure 6.9 where the relative bias, calculated as

$$\frac{\widehat{\epsilon}_b{}^{enr} - \epsilon_b^{enr}}{\epsilon_b^{enr}}, \tag{6.5}$$

between the estimated b-tag efficiency and the expected b-tag efficiency is shown. The uncertainties reflect the limited size of the simulated samples. In the next section it will be shown that the observed discrepancy is due to the correlation between the jet-muon mass and the b-tag discriminator.

6.1.3 Improvement of the b-tag efficiency estimation

A discrepancy is observed between the estimated and the expected b-tag efficiency for loose cuts on the b-tag discriminator. The main reason for the disagreement between the shapes of the b-tag discriminator distributions at low values is due to the correlation

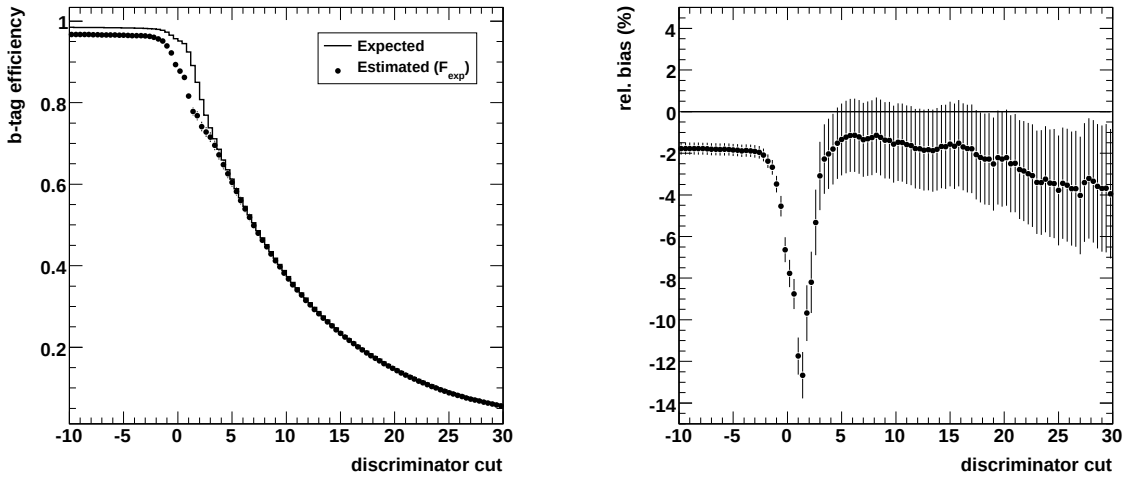

Figure 6.9: The distribution of the expected and estimated b-tag efficiency (left) and the corresponding relative bias (right) as a function of a cut on the b-tag discriminator value.

between the b-tag discriminator and the jet-muon mass (cf. Figure 6.5). The reason for this correlation is that the jet-muon mass is correlated with the p_T of the jet. In Section 4.3.5 it was shown that the b-tag discriminator is also correlated with the p_T of the jet. Due to these correlations, the b-tag discriminator values tend to be higher for higher jet-muon masses, therefore the shape of the b-tag discriminator distributions does not fully agree when comparing them in the b-enriched and b-depleted subsamples. In Figure 6.10 this is demonstrated for the non-b quark jets since the strongest dependency is observed there. Only the region around the peak is displayed and clearly shows an on average higher value for the b-tag discriminator for jets in the b-depleted sample w.r.t. jets in the b-enriched subsample.

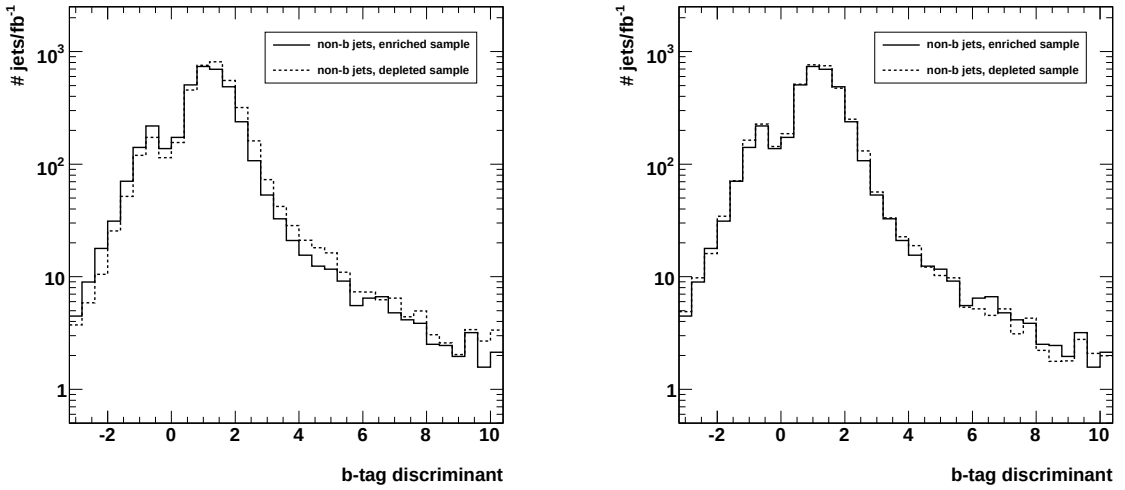

Figure 6.10: The distribution of the b-tag discriminator for non-b quark jets in the b-enriched and the b-depleted subsamples before (left) and after (right) reweighting of the jets in the b-depleted subsample.

To improve the agreement between the b-tag discriminator distribution of the b-enriched

and b-depleted subsamples, the spectrum of the p_T of the jets in the b-depleted subsample is reweighted to agree with the p_T spectrum of the jets in the b-enriched subsample. This is done by applying a weight to each jet in the b-depleted subsample. This weight is calculated as the bin-by-bin ratio between the p_T spectra of the b-enriched subsample and b-depleted subsample and is therefore completely data-driven. The reweighting function is obtained from the normalized p_T distributions, shown in Figure 6.11, indicating as well pseudo-data reflecting 1 fb^{-1}. The b-enriched subsample contains only jets satisfying 90 GeV/c^2 $< m_{\mu j} <$ 160 GeV/c^2 and has thus a softer p_T-spectrum than jets in the b-depleted subsample which satisfy 160 GeV/c^2 $< m_{\mu j} <$ 300 GeV/c^2. Figure 6.11 also shows the bin-by-bin ratio between the normalized p_T distributions and the corresponding exponential fit. The uncertainties are due to the limited size of the simulated samples. Now, to reweight the b-tag discriminator distribution of the b-depleted sample Δ_{obs}^{depl} to a new distribution $\Delta_{obs,rew}^{depl}$, to each individual jet in the b-depleted subsample a weight is applied based on the fitted reweighting function evaluated at the p_T-value of the jet. The uncertainty on the weight factor is expected to contribute only to second order and is thus not accounted for. In Figure 6.10 it is demonstrated that the b-tag discriminator distribution for non-b quark jets now agrees when comparing the b-enriched subsample with the reweighted b-depleted subsample, a similar but smaller effect has been observed for b quark jets.

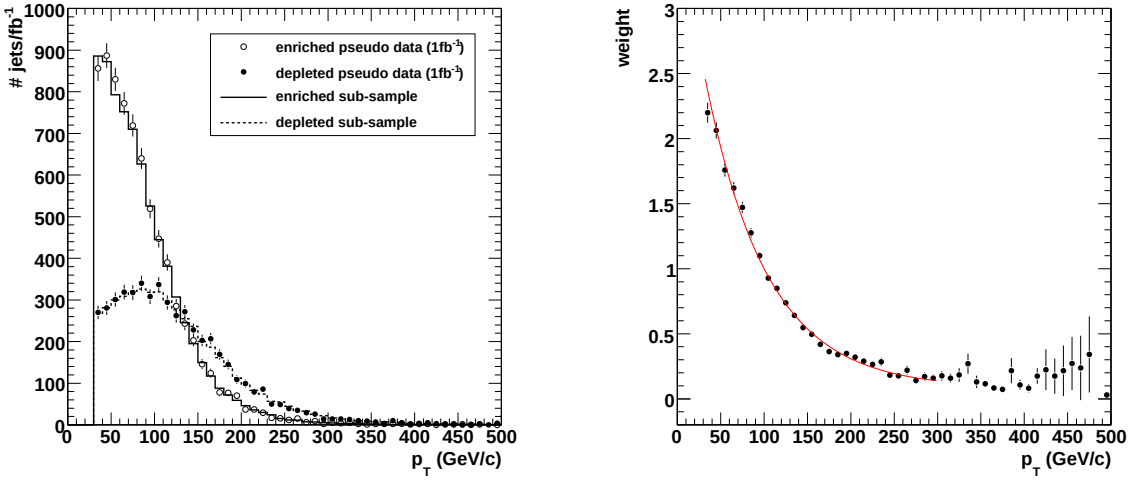

Figure 6.11: The distribution of the transverse momentum of all jets in the b-enriched and b-depleted subsample (left). The pseudo-data reflects the expectations for an integrated luminosity of 1 fb^{-1}. The ratio between the normalized p_T distributions and the fit of the exponential reweighting function (right).

After the correlation between the b-tag discriminator and the jet-muon mass is resolved by reweighting the jets in the b-depleted subsample, the corresponding reweighted b-tag discriminator distribution $\Delta_{obs,rew}^{depl}$ is subtracted from the b-tag discriminator distribution obtained in the b-enriched subsample according to,

$$\widehat{\Delta}_b^{enr,rew,i} = \Delta_{obs}^{enr,i} - F^{exp}.\Delta_{obs,rew}^{depl,i}. \tag{6.6}$$

Figure 6.12 shows the estimation of the b-tag discriminator distribution $\widehat{\Delta}_b^{enr,rew}$ where the uncertainties reflect the limited size of the simulated samples. When comparing Figure

6.12 with Figure 6.8 it is observed that the estimated b-tag distribution after reweighting agrees better with the expected b-tag discriminator distribution for low values of the b-tag discriminator.

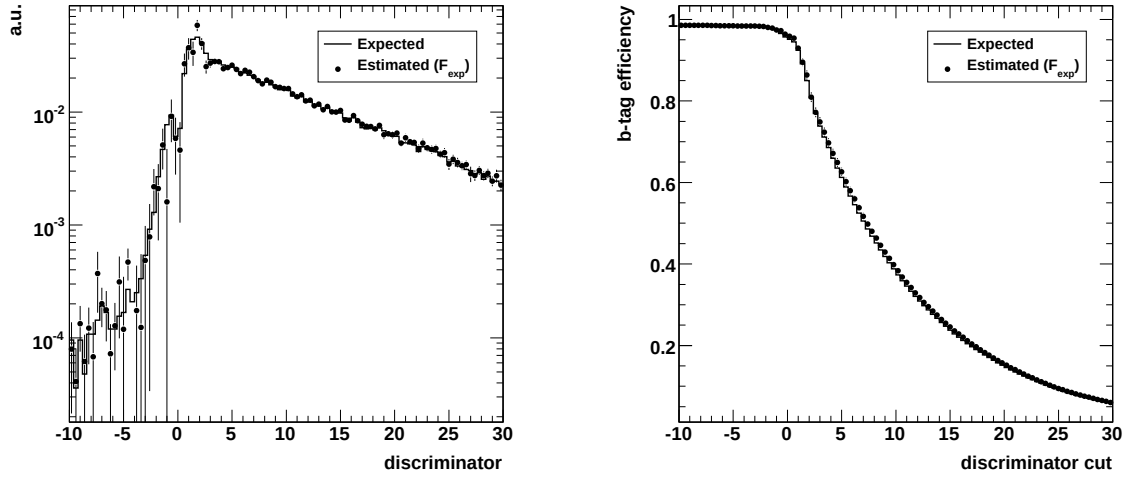

Figure 6.12: The estimated and expected b-tag discriminator distribution for b quark jets in the b-enriched subsample (left) and the corresponding estimated and expected b-tag efficiency as a function on the cut on the b-tag discriminator (right).

In Figure 6.12 the b-tag efficiency obtained from the estimated b-tag discriminator distribution is shown and is compared to the expected b-tag efficiency for b quark jets in the b-enriched subsample. In Figure 6.13 the relative and absolute bias between the estimated b-tag efficiency and the expected b-tag efficiency are shown. The estimated b-tag efficiency does not have a bias anymore for low cuts on the b-tag discriminator distribution. Although the bias at low b-tag discriminator cuts has disappeared an overall relative bias of 2-3% is introduced for harder cuts on the b-tag discriminator. The uncertainties reflect the limited size of the simulated samples.

In Table 6.2 an overview is given of the b-tag efficiency at three working points. The working points have been chosen to yield a b-tag efficiency of about 75% for the loose working point, a b-tag efficiency of about 50% for the medium working point and a b-tag efficiency of about 25% for the tight working point[1]. For each working point the expected b-tag efficiency ϵ_b^{tot} for all b quark jets in the b quark jet candidate sample is quoted together with the expected b-tag efficiency ϵ_b^{enr} for the b quark jets in the b-enriched subsample. The estimated b-tag efficiency $\widehat{\epsilon}_b{}^{enr}$, based on the estimated b-tag discriminator distribution $\widehat{\Delta}_b^{enr,rew}$, is compared to the expected b-tag efficiency in the b-enriched subsample. The relative bias with respect to the expected b-tag efficiency in the b-enriched subsample is quoted as well, the uncertainties are due to the limited size of the simulated samples. In the last column the relative and absolute expected statistical uncertainty on the estimated b-tag efficiency are shown for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV.

[1]For the high efficiency track counting algorithm these working points correspond to a cut on the b-tag discriminator of 2.72, 7.19 and 14.69 for the loose, medium and tight working point respectively.

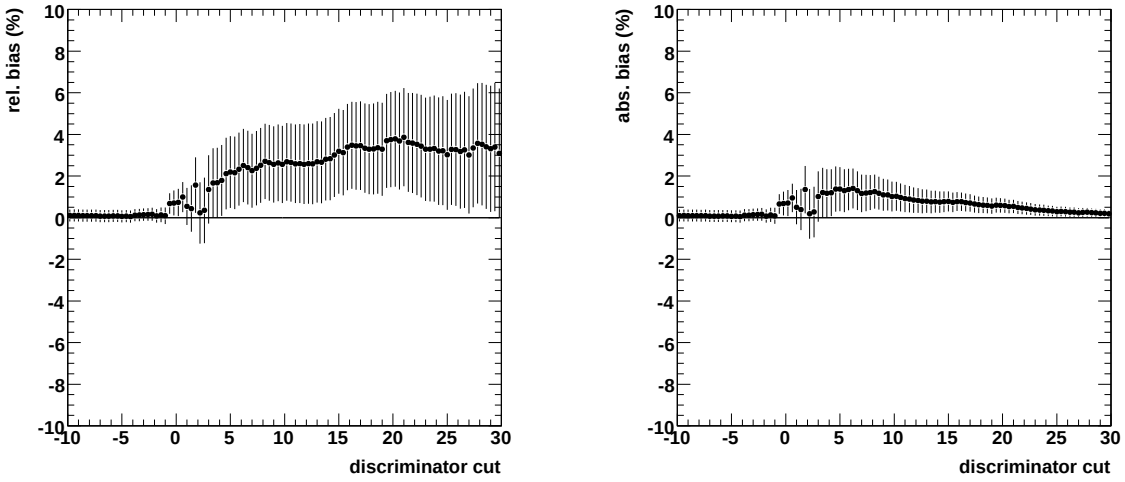

Figure 6.13: The relative bias (left) and absolute bias (right) between the expected and estimated b-tag efficiency as a function of a cut on the b-tag discriminator value.

working point	ϵ_b^{tot}	ϵ_b^{enr}	$\widehat{\epsilon}_b^{enr,rew}$	rel. bias	stat. uncertainty rel.	abs.
loose	74.1 ± 0.1	74.4 ± 0.2	75.2 ± 1.1	$(1.1 \pm 1.6)\%$	3.5%	2.6
medium	48.6 ± 0.1	48.7 ± 0.2	49.9 ± 0.8	$(2.4 \pm 1.7)\%$	4.0%	2.0
tight	24.1 ± 0.1	24.0 ± 0.2	24.8 ± 0.4	$(3.1 \pm 2.0)\%$	5.2%	1.3

Table 6.2: Estimated and expected b-tag efficiency for three working points. The next to last column shows the relative bias between the estimated and expected b-tag efficiency. The last column indicates the expected relative and absolute statistical uncertainty on the estimated b-tag efficiency for an integrated luminosity of 1 fb^{-1} at 10 TeV.

6.1.4 Statistical properties of the estimator

To study the statistical properties of the b-tag efficiency estimator, resampling techniques are used. In total N samples are constructed by randomly selecting a number of events from the total simulated sample in order to reflect a certain integrated luminosity. For each sample the event selection and the method are applied leading to N results for the estimation of the b-tag efficiency. The pull of the estimated b-tag efficiency for each pseudo-experiment i with respect to the mean obtained over all estimated b-tag efficiencies is calculated as

$$\frac{\widehat{\epsilon}_b^{\,i} - \overline{\epsilon}_b}{\delta\widehat{\epsilon}_b^{\,i}} \tag{6.7}$$

where $\widehat{\epsilon}_b^{\,i}$ and $\delta\widehat{\epsilon}_b^{\,i}$ are the estimated b-tag efficiency and its corresponding uncertainty for a given pseudo-experiment i and $\overline{\epsilon}_b$ is the average b-tag efficiency over all pseudo-experiments.

If the uncertainties $\delta\widehat{\epsilon}_b^{\,i}$ and the residuals $(\widehat{\epsilon}_b^{\,i} - \overline{\epsilon}_b)$ are well estimated, the distribution of the pull is expected to have a unit width.

The distribution of the pull for the estimation of the b-tag efficiency at the medium working point with and without reweighting is obtained for a total number of 500 pseudo-experiments each reflecting an integrated luminosity of 1 fb^{-1} and is shown in Figure 6.14. The distribution is fitted with a Gaussian function resulting in a width equal to 0.91 ± 0.03 and 0.92 ± 0.03 for respectively the estimation without and with reweighting. This indicates that the uncertainty on the b-tag efficiency is reliable and that the reweighting of the b-tag discriminator distribution has a negligible effect on the estimation of the uncertainty. The mean value of the fit of the pull distributions are found to be compatible with zero.

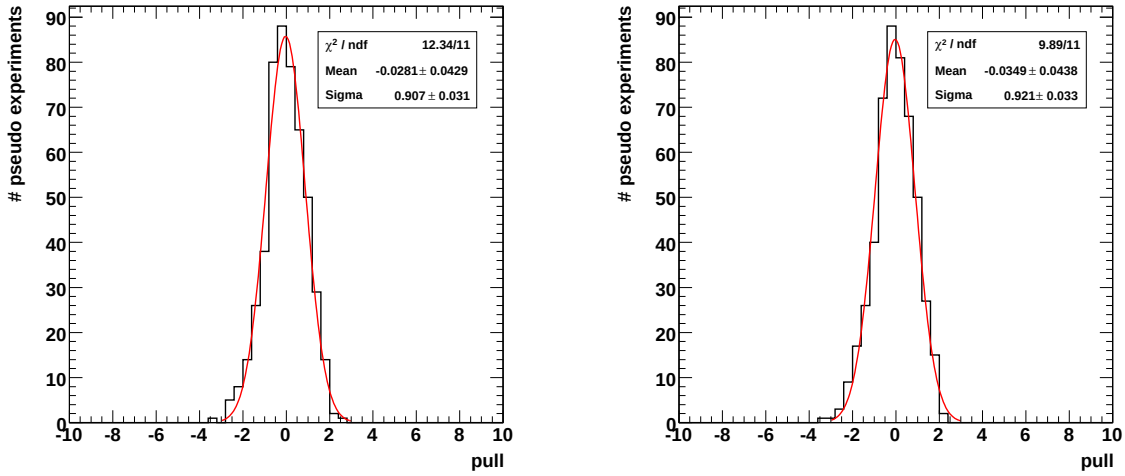

Figure 6.14: The distribution of the pull when applying the method in 500 pseudo-experiments, each reflecting an integrated luminosity of 1 fb^{-1}, without (left) and with (right) the reweighting of the b-tag discriminator distribution of the jets in the b-depleted subsample.

6.2 Data-driven scale factor estimation

In the previous section the method to estimate the b-tag efficiency from a b-enriched and a b-depleted jet sample is introduced. After reweighting the b-tag discriminator distribution the correlation between the jet-muon mass and the b-tag discriminator is resolved. An important aspect to obtain an unbiased estimation of the b-tag efficiency is a correct scale factor F to cancel the contribution of non-b quark jets in the b-enriched subsample. Up to now this ratio between the number of non-b quark jets in the b-enriched subsample and the number of non-b quark jets in the b-depleted subsample was assumed from the simulation. The distribution of the jet-muon mass for non-b quark jets strongly depends on the calibration of the jet energies. Uncertainties on the jet energy scale in the analysis of real proton collision data make the scale factor F^{exp} from simulations unreliable. In Table 6.3 the bias on the scale factor F after a variation of $\pm 10\%$ on the jet energy scale is

shown[2]. A relative bias on the scale factor, defined as,

$$(\widehat{F} - F^{exp})/F^{exp} \tag{6.8}$$

of 4-7% is observed when varying the jet energy scale with $\pm$10%.

	-10%	nominal	+10%
F^{exp}	1.025 $\pm$ 0.011	0.960 $\pm$ 0.010	0.918 $\pm$ 0.010
rel. bias	6.9 $\pm$ 1.5 %	-	-4.4 $\pm$ 1.4 %

Table 6.3: The relative bias on the scale factor F^{exp} due to the uncertainty on the jet energy scale.

A positive bias on the scale factor F would lead to an over-correction of the b-tag discriminator distribution in the b-enriched subsample while a negative bias would lead to an under-correction leaving a residual fraction of non-b quark jets in the estimated b-tag discriminator distribution. In either case the shape of the estimated b-tag discriminator distribution $\widehat{\Delta}_b^{enr,rew}$ does not represent the expected distribution and a bias is introduced on the estimated b-tag efficiency. In Figure 6.15 the relative bias on the estimated b-tag efficiency is given as a function of the relative bias on the scale factor F^{exp}. A relative bias of about 5% on the scale factor leads to a relative bias of 7-9% on the estimated b-tag efficiency. This clearly demands for a better control of the scale factor F to account for these variations.

6.2.1 Selection of the control sample

To demonstrate the principle of the method the scale factor F^{exp} was obtained from simulations by taking the ratio of the number of non-b quark jets in the b-enriched subsample and the number of non-b quark jets in the b-depleted subsample. The aim of the control sample is to select a reasonably pure sample of non-b quark jets where the jet-muon mass distribution can be used to estimate F. In the case of semi-leptonic $t\bar{t}$ events, the non-b quark jets in the b quark jet candidate sample come from radiation jets or from the hadronic decay of the W boson. Therefore the control sample is constructed from the two jets identified as the jets from the hadronic W-decay. Table 6.4 shows the number of b quark jets and non-b quark jets in this control sample obtained from all processes. Selecting these jets, a sample is obtained with a non-b quark jet purity of about 76%. This number is increased to about 91% when applying a cut on the b-tag discriminator of the jets. Only jets with a b-tag discriminator value lower than 3, for the track counting high efficiency b-tagging algorithm, are retained.

The estimated scale factor $\widehat{F}$ from the control sample, is obtained by reconstructing the shape of the jet-muon mass in the control sample. This distribution should reflect the shape

[2]A definition of the jet energy scale variation is given in Section 6.4.1.

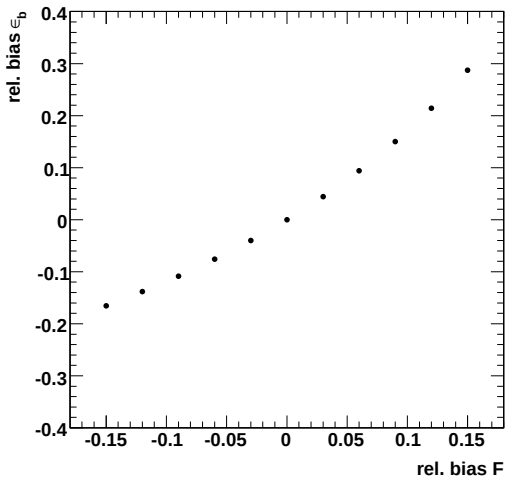

Figure 6.15: The relative bias on the estimated b-tag efficiency as a function of the relative bias on the scale factor.

	control sample	control sample + anti-b-tag cut
non-b quark jets	28414 $\pm$ 80	27127 $\pm$ 76
b quark jets	8853 $\pm$ 42	2645 $\pm$ 13
non-b purity	76%	91%

Table 6.4: The expected number of b quark jets and non-b quark jets in the control sample obtained from all processes, without and with applying an anti-b-tag cut on the selected jets.

of the jet-muon mass of non b-quark jets in the b candidate sample. In Figure 6.16 the distribution of the jet-muon mass in the control sample is shown for all processes together with pseudo-data reflecting an integrated luminosity of 1 fb^{-1}. Also shown is the jet-muon mass distribution for all jets in the control sample and for the non-b quark jets in the b candidate sample. It is found that the shape of the jet-muon mass in the control sample not reflects the expected jet-muon mass shape of the non-b quark jets in the b candidate sample.

6.2.2 Reweighting of the control sample

A disagreement was found in the previous section between the jet-muon mass reconstructed with the jets from the b candidate sample and the jet-muon mass reconstructed with the jets from the control sample. This difference is mainly originating from the different transverse

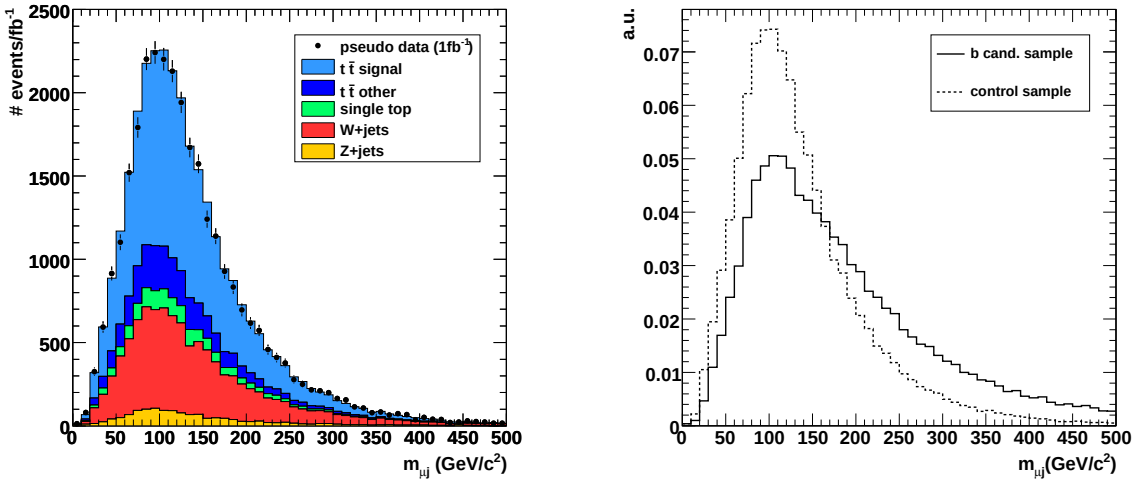

Figure 6.16: The distribution of the jet-muon mass reconstructed with the jets from the control sample for the various processes (left), compared to the mass distribution reconstructed with the non-b quark jets from the b candidate sample (right).

momentum and pseudo-rapidity spectra of jets in the control sample and the anti-b-tagged[3] jets in the b candidate sample as shown in Figure 6.17. It is observed that the jets in the control sample have in general a lower transverse momentum which is expected from the observations in Section 5.2.1.

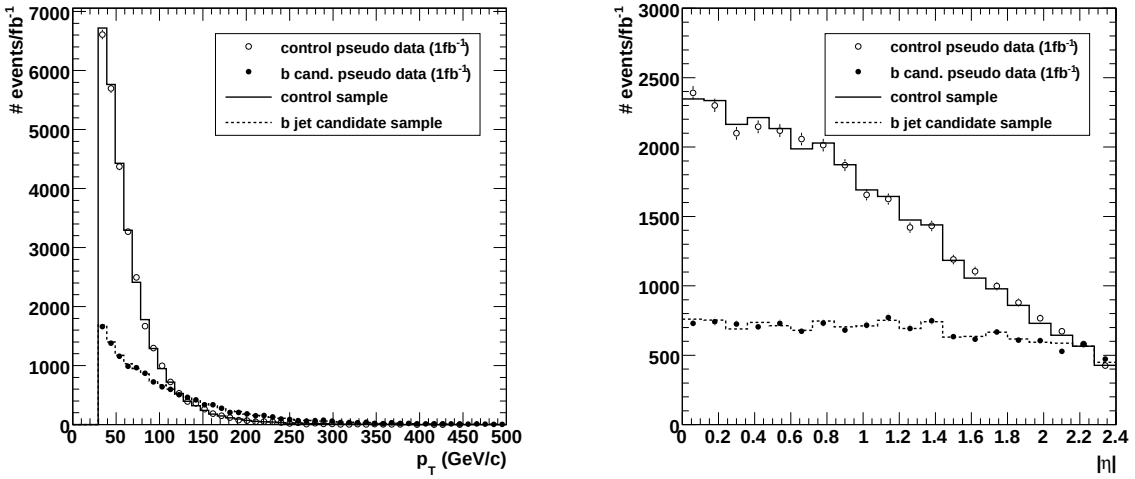

Figure 6.17: The distribution of the transverse momentum (left) and the pseudo-rapidity (right) of the anti-b-tagged jets in the b quark jet candidate sample and the jets in the control sample.

To account for the difference between both samples, the jets in the control sample will be reweighted in order to force the transverse momentum and pseudo-rapidity spectra to agree between the control sample and the b candidate sample. The reweighting factor

[3]The same anti-b-tag cut was applied to compare the non-b quark jet content of the b candidate sample with the jets in the control sample.

is calculated from the ratio of the normalized two-dimensional spectra of the transverse momentum and the pseudo-rapidity of the jets in the control sample and the anti-b-tagged jets in the b candidate sample. This reweighting can be performed in a data-driven way. In Figure 6.18 the bin-by-bin reweighting factors, as a function of the transverse momentum and the pseudo-rapidity of the jet, are shown.

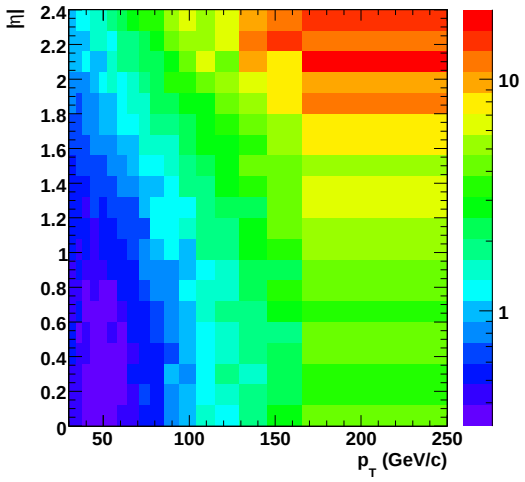

Figure 6.18: The distribution of the reweighting factors as a function of the transverse momentum and the pseudo-rapidity of the jets.

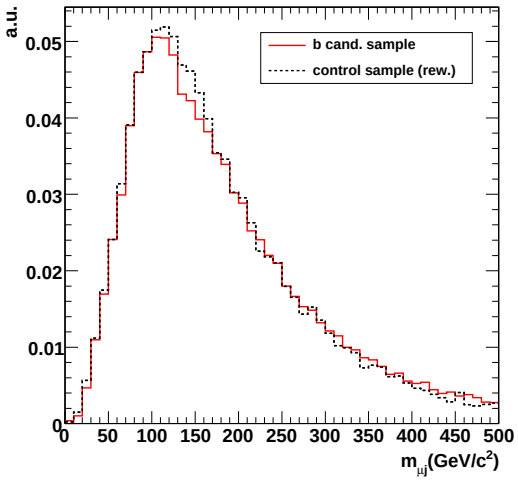

Figure 6.19: The reweighted distribution of the jet-muon mass for the jets in the control sample and the distribution of the expected jet-muon mass in the signal sample.

After applying a weight to each jet in the control sample the reweighted $m_{\mu j}$ distribution is constructed. This distribution is shown in Figure 6.19 and is compared to the expected mass distribution from the non-b quark jets in the b candidate sample. The estimated scale factor $\widehat{F}^{rew}$ is found to be 1.000 $\pm$ 0.006 while the expected scale factor F^{exp} was equal to 0.960 $\pm$ 0.010. The relative bias between the expected and estimated scale factor is 4.1 $\pm$ 1.1%. The uncertainty on the estimated scale factor is due to the limited size of the simulated samples. The uncertainties on the reweighting factors have not been propagated through the analysis. The effect on the uncertainty of the estimated of the b-tag efficiency is expected to be small.

6.2.3 Constraining the control sample

A relative bias on the scale factor F of about 4% is observed. The main reason for this discrepancy is coming from events where the best jet combination, i.e. the combination with the lowest χ^2-value, fails to fulfill the top quark mass and W boson mass constraints and returns a very high χ^2_{min} value. In Figure 5.14 it was observed that processes other than semi-muonic $t\bar{t}$ events, with no significant radiation, have a distribution of the χ^2_{min} of the event with a long tail. To reject these events a cut on the χ^2_{min}-value of the event is applied. In Figure 6.20 the bias on the estimation of scale factor $\widehat{F}^{rew}$ is shown as a function of a cut on the χ^2_{min}-value of the event, rejecting events with $\chi^2_{min} > \chi^2_{cut}$. The bias on the estimation of the scale factor decreases when cutting harder on the χ^2_{min}-value

until a χ^2_{cut}-value of about 10. Applying a harder cut on the χ^2_{min}-value introduces again a bias on the estimation of the b-tag efficiency. Based on this observation a cut on the χ^2_{min}-value of 25 is applied. For this cut the estimated scale factor $\widehat{F}^{rew}$ is equal 1.121 $\pm$ 0.007 whereas the expected scale factor is F^{exp}= 1.12 $\pm$ 0.01. The relative bias is now 0.1 $\pm$ 1.4% and is thus compatible with zero. This cut is now applied to estimate the efficiency in a completely data driven way.

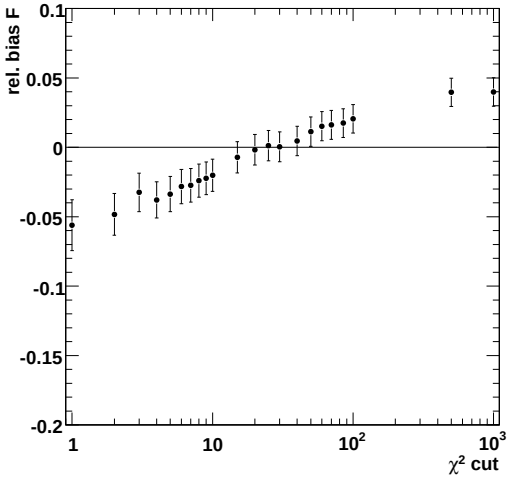

Figure 6.20: Bias on the estimation of the scale factor F as a function of the cut on the χ^2_{min}-value of the event.

To illustrate the origin of the bias on the estimation of the scale factor the effect of the cut on the reconstructed top quark mass and W boson mass distribution is shown in Figure 6.21. When applying the χ^2-cut, only a fixed mass window centered around the central mass value is allowed. This is due to the fact that the same mass resolutions, $\sigma(m_W)$ and $\sigma(m_{top})$, are used in the χ^2 expression (cf. Equation 5.2) for all selected events. The long tails of high mass values are mainly due to semi-muonic $t\bar{t}$ events where one of the leading four jets is not originating from the hadronic decaying top quark and from other processes.

When now reconstructing the jet-muon mass in the control sample, it is unfavorable that the kinematical properties of semi-muonic $t\bar{t}$ events are altered by applying a χ^2-cut. To quantify the optimal χ^2-cut where the kinematics of the event are not yet altered, the symmetry of the reconstructed top quark mass distributions is studied. This symmetry is quantified as the difference between the peak-value and the maximal observed mass value and the peak-value and the minimal observed mass value, calculated after several χ^2-cuts and shown in Figure 6.22. It is found that the symmetry of the top quark mass distribution is improved when imposing a cut harder than a χ^2-cut of about 20. Another data-driven estimator that can define the preferred cut on the χ^2-value is the fraction of events that is removed when applying a cut on the χ^2-value. A Gaussian fit is applied to the top quark mass peak and an interval around the mean value of $\pm 2\sigma$ is defined. For each cut the number of events within this interval is compared to the initial number of events before the χ^2-cut. This fraction is shown in Figure 6.23. It is found that when cutting around a χ^2-cut of 20, events are being removed in the top quark mass peak.

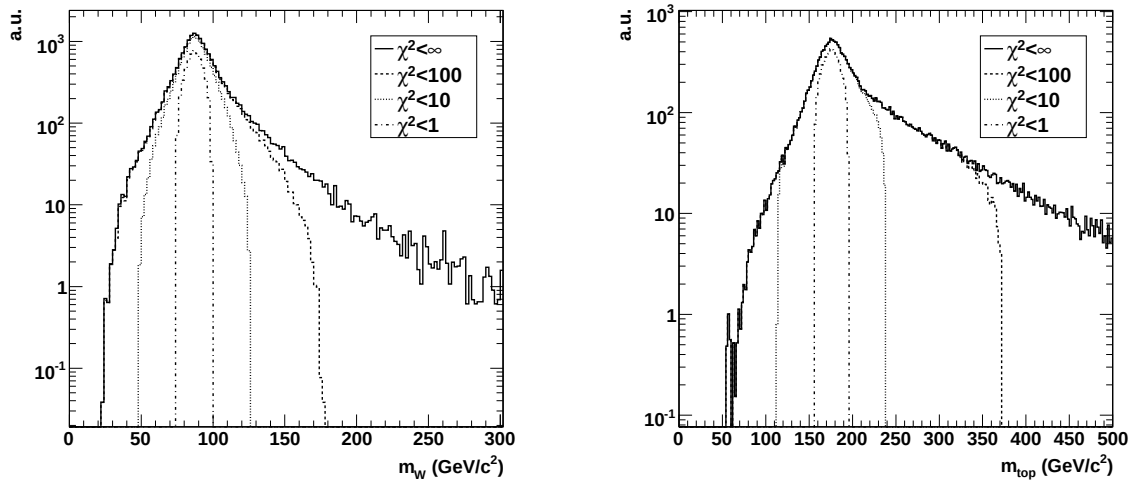

Figure 6.21: Reconstructed W boson mass and reconstructed top quark mass distributions for different cuts on the χ^2_{min} of the event.

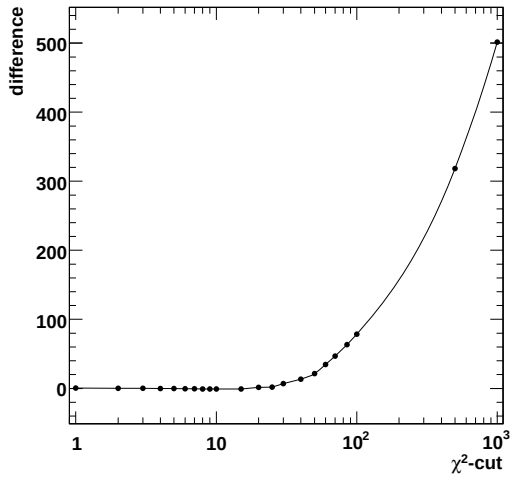

Figure 6.22: Difference between the low side and the high side of the reconstructed mass window of the top quark as a function of the cut on the χ^2.

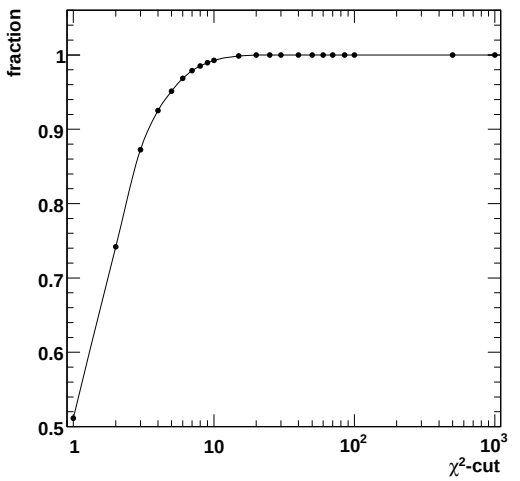

Figure 6.23: The fraction of events in the 2-σ interval around the top quark mass peak that remains when applying a cut on the χ^2.

6.3 Inclusive estimation of the b-tag efficiency

Based on the observations made in the previous section the results for the estimation of the b-tag efficiency using the scale factor from the reweighted jet-muon mass obtained in the control sample, $\widehat{F}^{rew}$, is overviewed in this section. In Figure 6.24 the b-tag efficiency is shown as a function of the cut on the b-tag discriminator. The relative bias between the expected and estimated b-tag efficiency shows the same trend as observed in Section 6.1.3. It is found that the χ^2-cut does not introduce a bias on the estimated b-tag efficiency, neither does the data driven estimate of the scale factor.

In Table 6.5 an overview of the expected and estimated b-tag efficiency is given for the

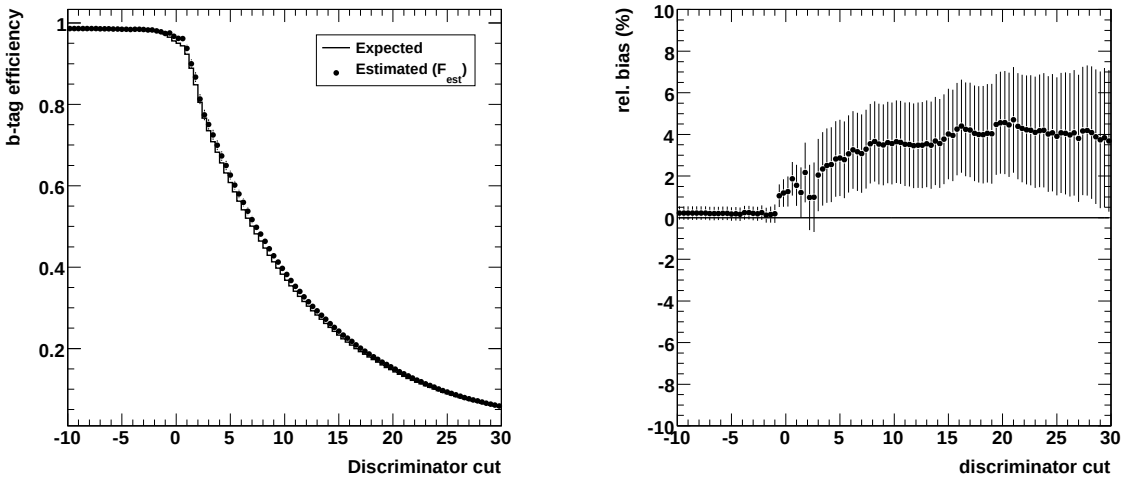

Figure 6.24: The estimated and expected b-tag efficiency (left) and the relative bias between them (right).

previously defined loose, medium and tight working points. A relative bias on the estimated b-tag efficiency of $1.7 \pm 1.7\%$ for the loose working point, up to $3.9 \pm 2.1\%$ for the tight working point, is observed. The relative and absolute statistical uncertainties expected for an integrated luminosity of 1 fb^{-1} are indicated in the last column.

working point	ϵ_b^{tot}	ϵ_b^{enr}	$\widehat{\epsilon}_b^{\,enr,rew}$	rel. bias	stat. uncertainty rel.	abs.
loose	73.8 ± 0.1	74.1 ± 0.2	75.4 ± 1.2	$(1.7 \pm 1.7)\%$	3.8%	2.8
medium	48.1 ± 0.2	48.2 ± 0.2	49.9 ± 0.9	$(3.3 \pm 1.9)\%$	4.4%	2.2
tight	23.6 ± 0.1	23.5 ± 0.2	24.5 ± 0.5	$(3.9 \pm 2.1)\%$	5.7%	1.4

Table 6.5: The estimated and expected b-tag efficiency for three working points corresponding to a loose, a medium and a tight cut on the b-tag discriminator value. The expected relative statistical uncertainty for an integrated luminosity of 1 fb^{-1} is given.

6.3.1 Statistical properties of the estimator

The distribution of the pull of the estimated b-tag efficiency at the medium working point with reweighting of the b-tag discriminator distribution and with a scale factor obtained from the control sample for 450 pseudo-experiments is shown in Figure 6.25. Each pseudo-experiment is a random selection of events from the total samples to reflect an integrated luminosity of 1 fb^{-1}. The pull distribution of the estimated scale factor $\widehat{F}^{rew}$ is shown as

well. The distribution was fitted with a Gaussian function resulting in a width equal to 1.11 ± 0.04 and 1.44 ± 0.05 for respectively the estimation of the b-tag efficiency and the scale factor. The pull of the scale factor estimator is significantly larger than unity indicating an underestimation of the statistical uncertainty on the estimation. This can be understood by the reweighting of the control sample. The reweighting factor was applied without accounting for the statistical uncertainty on the weight factor. The effect on the pull of the estimated b-tag efficiency is of the order of 6%.

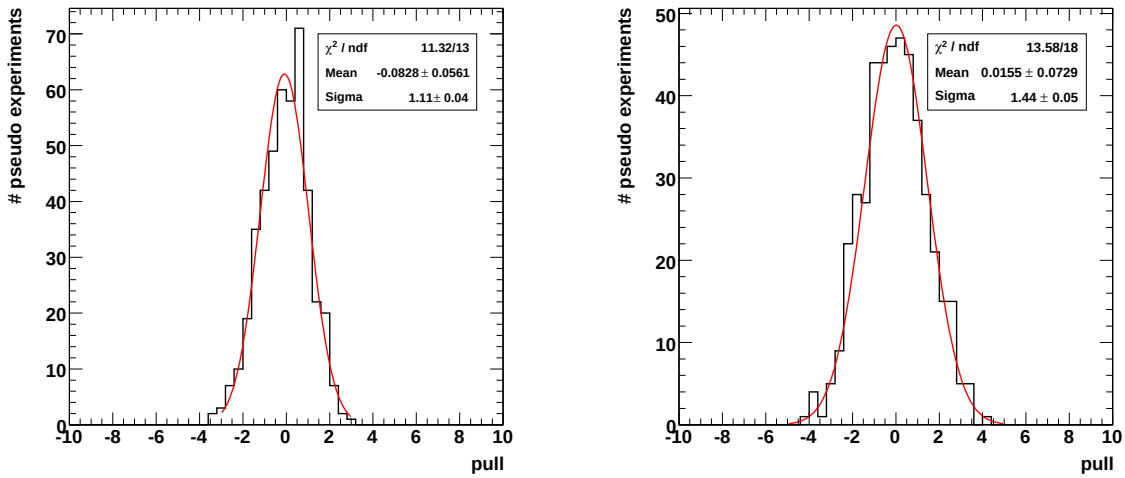

Figure 6.25: The distribution of the pull the b-tag efficiency estimator (left) and the scale factor estimator (right) for 450 pseudo-experiments each reflecting a sample with an integrated luminosity of 1 fb^{-1}.

6.4 Systematic uncertainties

The method developed to estimate the b-tag efficiency is subject to various systematic effects. These systematic uncertainties originate from several sources such as the robustness of the reconstructed objects, in particular the calibration of the jets. Another important source of uncertainty is the simulation of the proton-proton collision, i.e. the generation of the initial and final state radiation. The uncertainties on the cross section of the background processes will induce additional systematic uncertainties on the estimated b-tag efficiency.

The aim of the estimation of the scale factor F based on a control sample is to increase the robustness of the method against systematic uncertainties. The systematic effects on the estimation of the scale factor and on the estimation of the b-tag efficiency at the medium working point are shown. An overview of the uncertainties on the estimation of the b-tag efficiency is shown in Table 6.10 at the end of this section.

6.4.1 Jet energy scale

Jets are reconstructed with the Seedles Infrared Safe cone algorithm and are calibrated using absolute and relative correction factors to obtain a flat jet energy response with respect to the transverse momentum and the pseudo-rapidity of the jets. To estimate the

b-tag efficiency, cuts are applied on the calibrated jets to obtain the b quark jet candidate sample as well as the control sample. These calibrated jets in the b candidate sample are then combined with the muon to obtain the mass of the jet-muon system. Based on this distribution a b-enriched and a b-depleted subsample are defined. A bias on the reconstructed jet energy can lead to a degradation of the separation between the enriched and the depleted subsample. Besides this the shape of the jet-muon mass in the control sample is used to estimate the non-b quark jet contribution in the b candidate sample. The shape of this distribution is obtained from a control sample based on the jets identified as the quarks from the hadronic decaying W boson. A bias on the reconstructed jet energy will lead to changes in the shape of the jet-muon mass introducing a possible bias on the estimation of the scale factor.

The effect of a systematic shift on the inclusive jet energy scale is studied by scaling the four-momenta of the jets by a factor α,

$$P_{jet}^{\pm\alpha} = ((1 \pm \alpha)E, (1 \pm \alpha)p_x, (1 \pm \alpha)p_y, (1 \pm \alpha)p_z). \tag{6.9}$$

A conservative shift of α=$\pm$10% is applied, reflecting the initial jet energy scale uncertainty based on simulations and cosmic ray and test beam data [98]. This shift is applied on all jets before the event selection and the complete analysis is repeated with the shifted jet four momenta.

The bias on the estimation of the scale factor F varies about 1% when comparing the negative and the positive shift on the jet energy scale. Given the uncertainty due to the limited size of the simulated samples this is compatible with zero but since the samples are highly correlated the bias is considered significant. The bias on the scale factor from the simulation was found to shift about 7%, while now a shift of only 1% is observed. This proves the robustness of the estimation of the scale factor from the control sample. Comparing the shift on the estimation of the b-tag efficiency at the medium working point a bias of about 1.5-2.5% is observed comparing the samples with a $\pm$10% jet energy scale variation with the nominal sample. The systematic uncertainty, induced by the uncertainty of the jet energy scale is taken to be the mean of this bias, i.e. 2%. No uncertainty is considered on this value since the samples with altered jet energy scale are highly correlated and thus no propagation of uncertainties can be applied to obtain an uncertainty. It is expected that the uncertainty on the jet energy scale will be reduced significantly when more data is collected by the CMS experiment.

6.4.2 Initial and final state radiation

The effect of the amount of initial and final state radiation is based on data simulated with the PYTHIA event generator. To model the amount of initial and final state radiation in the proton collision process, dedicated parameters need to be set. To study the systematic uncertainty due to the amount of initial and final state radiation the nominal parameters are varied. A description of the applied variations is given in Section 3.3.1. Two $t\bar{t}$ samples with varied parameters have been generated for a comparison with the nominal sample. Although the jet multiplicity changes significantly in these samples, i.e. the average number of jets increases or decreases by a few units, the effect in the analysis is found to be small. In Table 6.7 an overview is given of the estimated and expected values of the b-tag efficiency and the

	F^{exp}	$\widehat{F}$	$(\widehat{F} - F^{exp})/F^{exp}$ - $(\widehat{F}_{nom} - F^{exp}_{nom})/F^{exp}_{nom}$
nominal	1.12 ± 0.01	1.121 ± 0.007	-
α=-10%	1.17 ± 0.01	1.186 ± 0.007	1 %
α=+10%	1.08 ± 0.01	1.071 ± 0.007	-0.8 %
	ϵ_b^{enr}	$\widehat{\epsilon}_b$	$(\widehat{\epsilon}_b - \epsilon_b^{enr})/\epsilon_b^{enr}$ -$(\widehat{\epsilon}_{b,nom} - \epsilon^{enr}_{b,nom})/\epsilon^{enr}_{b,nom}$
nominal	48.2 ± 0.2	49.9 ± 0.9	-
α=-10 %	49.0 ± 0.2	51.8 ± 0.9	2.5 %
α=+10 %	48.3 ± 0.2	49.2 ± 0.9	-1.5 %

Table 6.6: Overview of the estimated scale factor and b-tag efficiency at the medium working point after shifting the jet energy scale.

scale factor. The mean of the bias on the estimated b-tag efficiency induced by the samples with respectively less and more radiation is found to be compatible with zero within the systematic uncertainty. The samples with more and less radiation and the nominal sample have been generated independently, therefore as a conservative choice, the uncertainty on the bias is taken to be the systematic uncertainty. An uncertainty of 2.6% is used for the effect modeling of the initial and final state radiation.

6.4.3 Background cross section

The cross sections of the W+jets and Z+jets samples are obtained from LO calculations and are extrapolated to NLO using an inclusive K-factor, therefore the uncertainties can be rather large. To study the influence of this uncertainty, the cross sections of the W+jets and Z+jets samples have been increased with a factor 1.5. To study the impact of the heavy flavour content of these samples, the V+qq+jets (V=W/Z) and the W+c+jets have been added independently to the samples used for the analysis. In Table 6.8 an overview is given of the effect of the background cross sections on the estimation of the scale factor and the b-tag efficiency. For most of the background processes the effect on the estimation of the scale factor is negligible, except for the W+jets sample where a larger effect might be present. For the estimation of the b-tag efficiency an effect is present for the W+jets process which is mainly due to the worse estimation of the scale factor. The resulting biases are highly correlated, therefore no uncertainties can be calculated for each difference. The largest one, being 1.9% when scaling the W+jets cross section with a factor of 1.5, is assumed to be the systematic uncertainty. It is expected that with more data collected by

	F^{exp}	$\widehat{F}$	$(\widehat{F} - F^{exp})/F^{exp}$ - $(\widehat{F}_{nom} - F^{exp}_{nom})/F^{exp}_{nom}$
nominal	1.05 ± 0.01	1.117 ± 0.007	-
less ISR/FSR	1.06 ± 0.02	1.12 ± 0.01	-0.1 ± 0.02 %
more ISR/FSR	1.16 ± 0.02	1.22 ± 0.01	0.0 ± 0.02 %
	ϵ_b^{enr}	$\widehat{\epsilon}_b$	$(\widehat{\epsilon}_b - \epsilon_b^{enr})/\epsilon_b^{enr}$ -$(\widehat{\epsilon}_{b,nom} - \epsilon^{enr}_{b,nom})/\epsilon^{enr}_{b,nom}$
nominal	49.3 ± 0.3	53.4 ± 0.7	-
less ISR/FSR	48.8 ± 0.4	53.3 ± 0.9	0.7 ± 2.6 %
more ISR/FSR	49.0 ± 0.4	52.9 ± 0.9	-0.5 ± 2.5 %

Table 6.7: Overview of the estimated scale factor and b-tag efficiency at the medium working point after comparing $t\bar{t}$ samples with less or more radiation.

the CMS experiment the uncertainties on the background cross sections will be reduced.

6.4.4 Event generator

To study the influence of the event generator for $t\bar{t}$ events, the $t\bar{t}$ events generated using the `MadGraph` event generator were replaced by the $t\bar{t}$ events generated using the `ALPGEN` generator. In Table 6.9 an overview is given of the bias on the estimation of the scale factor and the b-tag efficiency at the medium working point. Within the statistical precision of this test there is no observed difference between the bias obtained with respectively the `MadGraph` and the `ALPGEN` event generator. The statistical precision of the test is 6.5% but will not be added to the total systematic uncertainty because of the limited size of the simulated sample.

6.4.5 Combination

In Table 6.10 the systematic uncertainties on the b-tag efficiency at the medium working point of the track counting high efficiency b-tagging algorithm, obtained in the previous sections, are listed. The bias on the estimation is considered a part of the systematic uncertainty since it is considered as significant. The relative statistical uncertainty is combined with the systematic uncertainty to obtain a combined relative uncertainty of 6.7% for an integrated luminosity of 1 fb^{-1} at a center of mass energy of 10 TeV. An important contribution to the systematic uncertainty is coming from the modeling of the initial and final state radiation. It must be emphasized that a conservative choice was made by using the

	F^{exp}	$\widehat{F}$	$(\widehat{F} - F^{exp})/F^{exp}$ - $(\widehat{F}_{nom} - F^{exp}_{nom})/F^{exp}_{nom}$
nominal	1.12 ± 0.01	1.121 ± 0.007	-
W+jets $\times$ 1.5	1.11 ± 0.01	1.098 ± 0.007	-1.2 %
Z+jets $\times$ 1.5	1.12 ± 0.01	1.121 ± 0.007	-0.1 %
add W+c+jets	1.12 ± 0.01	1.123 ± 0.007	0.3 %
add V+qq+jets	1.12 ± 0.01	1.121 ± 0.007	0.1 %
	ϵ_b^{enr}	$\widehat{\epsilon}_b$	$(\widehat{\epsilon}_b - \epsilon_b^{enr})/\epsilon_b^{enr}$ -$(\widehat{\epsilon}_{b,nom} - \epsilon^{enr}_{b,nom})/\epsilon^{enr}_{b,nom}$
nominal	48.2 ± 0.2	49.9 ± 0.9	-
W+jets $\times$ 1.5	48.2 ± 0.2	48.9 ± 0.9	-1.9 %
Z+jets $\times$ 1.5	48.2 ± 0.2	49.9 ± 0.9	0.2 %
add W+c+jets	48.2 ± 0.2	50.2 ± 0.9	0.7 %
add V+qq+jets	48.2 ± 0.2	50.0 ± 0.9	0.3 %

Table 6.8: Overview of the estimated scale factor and b-tag efficiency at the medium working point after increasing the cross section of the various background processes.

	F^{exp}	$\widehat{F}$	$(\widehat{F} - F^{exp})/F^{exp}$ - $(\widehat{F}_{nom} - F^{exp}_{nom})/F^{exp}_{nom}$
MadGraph	1.12 ± 0.01	1.121 ± 0.007	-
ALPGEN	1.08 ± 0.03	1.08 ± 0.02	-0.2 $\pm$ 3.6 %
	ϵ_b^{enr}	$\widehat{\epsilon}_b$	$(\widehat{\epsilon}_b - \epsilon_b^{enr})/\epsilon_b^{enr}$ -$(\widehat{\epsilon}_{b,nom} - \epsilon^{enr}_{b,nom})/\epsilon^{enr}_{b,nom}$
MadGraph	48.2 ± 0.2	49.9 ± 0.9	-
ALPGEN	50.6 ± 1.1	52.0 ± 2.9	-0.6 $\pm$ 6.5 %

Table 6.9: Overview of the estimated scale factor and b-tag efficiency at the medium working point for ALPGEN and MadGraph event generators.

statistical uncertainty on the observed bias as systematic uncertainty. Also for the systematic uncertainty due to the jet energy scale and the cross section of the background rather conservative assumptions have been made.

The complete procedure to determine the systematic uncertainties for the medium working point was repeated for the loose and the tight working points. In the table a breakdown of the systematic uncertainty for the three working points is given together with the corresponding relative statistical uncertainty. It is found that a combined relative uncertainty, respectively for the loose and the tight working point, of 5.7% and 8.4% can be obtained.

relative uncertainty	loose	medium	tight
statistical (1 fb^{-1})	3.8 %	4.4 %	5.7 %
systematic			
relative bias	1.7 ± 1.7 %	3.3 ± 1.9 %	3.9 ± 2.1 %
JES	2.4 %	2.0 %	2.6 %
ISR/FSR	≤ 2.0 %	≤ 2.6 %	≤ 3.8 %
background cross section	2.3 %	1.9 %	1.5 %
total	4.2 %	5.0 %	6.2 %
combined	5.7 %	6.7 %	8.4 %

Table 6.10: Overview of the systematic uncertainties, the statistical uncertainty for 1 fb^{-1} and the combined uncertainty of the estimation on the b-tag efficiency at the loose, medium and tight working points of the track counting high efficiency b-tagging algorithm.

6.5 Results for other b-tagging algorithms

The method to estimate the b-tag efficiency developed in the previous section was applied for the track counting high efficiency b-tag algorithm. The method is however not limited to this algorithm and is more generally applicable for other b-tag algorithms. In Figure 6.26 and 6.27 the estimation of the b-tag efficiency and the corresponding relative bias is shown for the simple secondary vertex b-tag algorithm and the combined secondary vertex algorithm. In general it can be concluded that the method performs equally well for these jet algorithms with relative biases on the estimated b-tag efficiency below 3-4%.

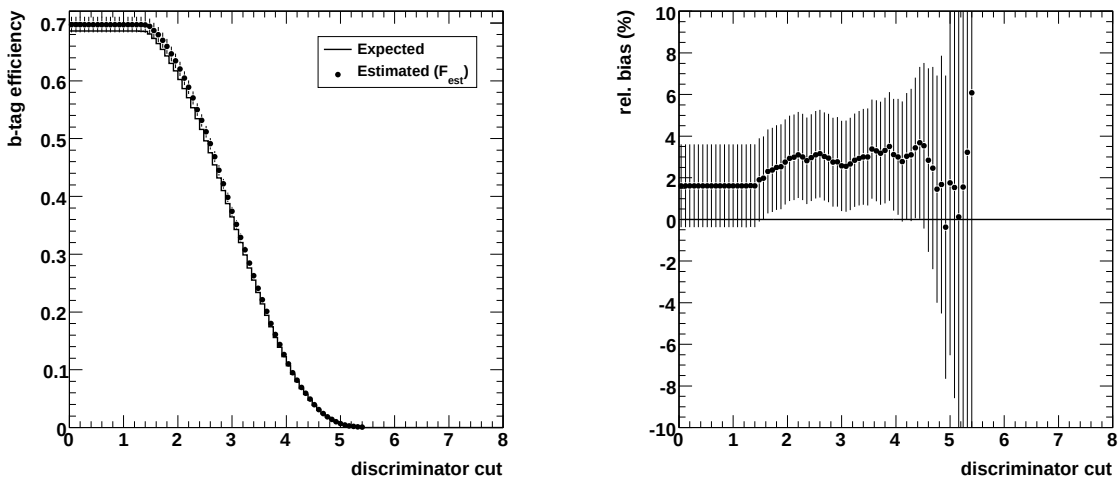

Figure 6.26: The estimated and expected b-tag efficiency, for the simple secondary vertex algorithm, (left) and the bias between the estimated and expected b-tag efficiency (right).

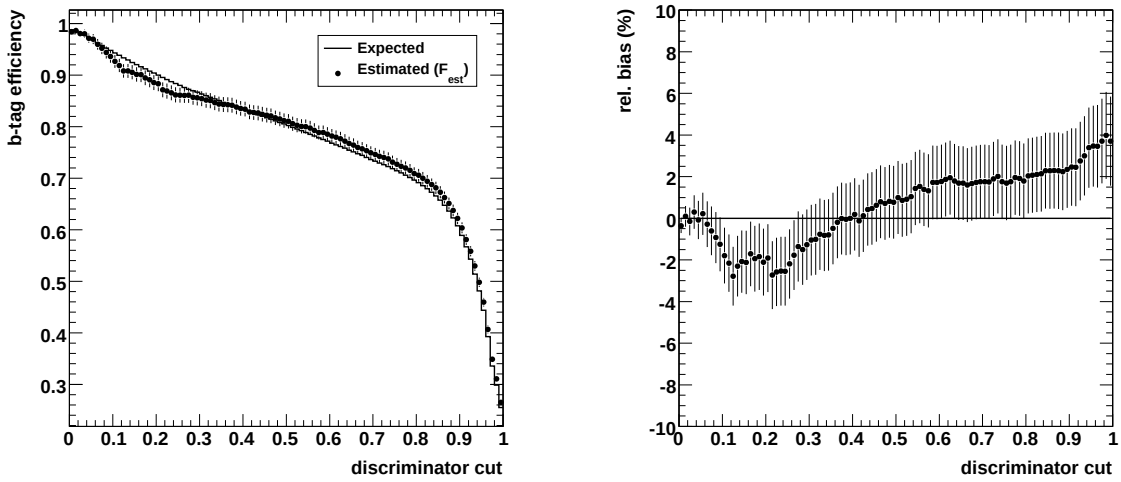

Figure 6.27: The estimated and expected b-tag efficiency, for the combined secondary vertex algorithm, (left) and the bias between the estimated and expected b-tag efficiency (right).

Chapter 7

Differential estimation of the b-tag efficiency

In the previous chapter a method to estimate the b-tag efficiency based on semi-muonic $t\bar{t}$ events is introduced. The $t\bar{t}$ events are obtained by applying an event selection to reduce the enormous multi-jet background and the large W+jets and Z+jets background. After this event selection a b quark jet candidate sample is extracted from the selected events and from this jet sample a b-enriched and a b-depleted subsample are obtained. Based on the b-tag discriminator distribution of the jets in the b-enriched subsample and the b-depleted subsample the b-tag efficiency is estimated for any given working point.

It is shown in Section 4.3.5 that the performance of b-tagging algorithms is correlated with the kinematic properties of the jets. The b-tag efficiency degrades for jets with a low or a very high transverse momentum and for forward jets. The event selection, prior to the estimation of the b-tag efficiency, extracts a jet sample with a specific p_T- and η-spectrum. Therefore the b-tag efficiency has to be estimated as a function of the kinematic properties of the jets to make the results generally applicable. The b-tag efficiency is affected by other variables besides p_T and η, like e.g. the number of constituents of the jet, but these are less important and therefore no differentiation is made for them.

In this chapter an extension of the method to estimate the b-tag efficiency as a function of the transverse momentum and the pseudo-rapidity of the jets is described. It will be argumented in Section 7.1 that, in contrast to the inclusive measurement, the estimation of the scale factor F from a control sample is not feasible in the differential measurement. Nevertheless the estimation of the differential b-tag efficiency can be obtained using the dependency of the scale factor on the transverse momentum or the pseudo-rapidity from simulation as shown in Section 7.2. In Section 7.3 the systematic uncertainties on the estimation of the differential b-tag efficiency are discussed and the chapter is concluded with Section 7.4 where the differential estimation of the b-tag efficiency is shown for different b-tagging algorithms.

7.1 Extension of the method to estimate the inclusive b-tag efficiency

In this section the method to estimate the inclusive b-tag efficiency, developed in the previous chapter, is extended towards an estimation of the differential b-tag efficiency. To estimate the b-tag efficiency differentiated as a function of the transverse momentum or the pseudo-rapidity of the jets, the b quark jet candidate sample is divided in bins. A jet in the b candidate sample with a transverse momentum p_T or a pseudo-rapidity η is considered to be part of bin i if it, respectively, fulfills the requirement $(p_T^{min})_i < p_T < (p_T^{max})_i$ or $(|\eta|^{min})_i < |\eta| < (|\eta|^{max})_i$. In principle to estimate the b-tag efficiency for the jets in each p_T- and η-bin i, an identical procedure can be applied as for the inclusive method. It is however found that the data-driven procedure to estimate the scale factor F from the control sample is not feasible.

In the previous chapter the scale factor F was defined as the ratio of the number of non-b quark jets in the b-enriched subsample divided by the number of non-b quark jets in the b-depleted subsample. The b-enriched subsample was defined as the jets in the b candidate sample which yield a jet-muon mass of $90\ \text{GeV}/\text{c}^2 < m_{\mu j} < 160\ \text{GeV}/\text{c}^2$ whereas the b-depleted subsample was defined as the jets satisfying the condition $160\ \text{GeV}/\text{c}^2 < m_{\mu j} < 300\ \text{GeV}/\text{c}^2$. For the jets in the control sample two subsamples were defined based on the same conditions and the ratio of jets in these subsamples provided an estimation of the scale factor.

In Figure 7.1 the distributions of the jet-muon mass for the non-b quark jets in the b candidate sample and for all jets in the control sample after reweighting are displayed for five p_T-bins. A strong correlation between the jet-muon mass and the p_T of the non-b quark jets in the b candidate sample is observed, a similar correlation is found for the jets in the control sample.

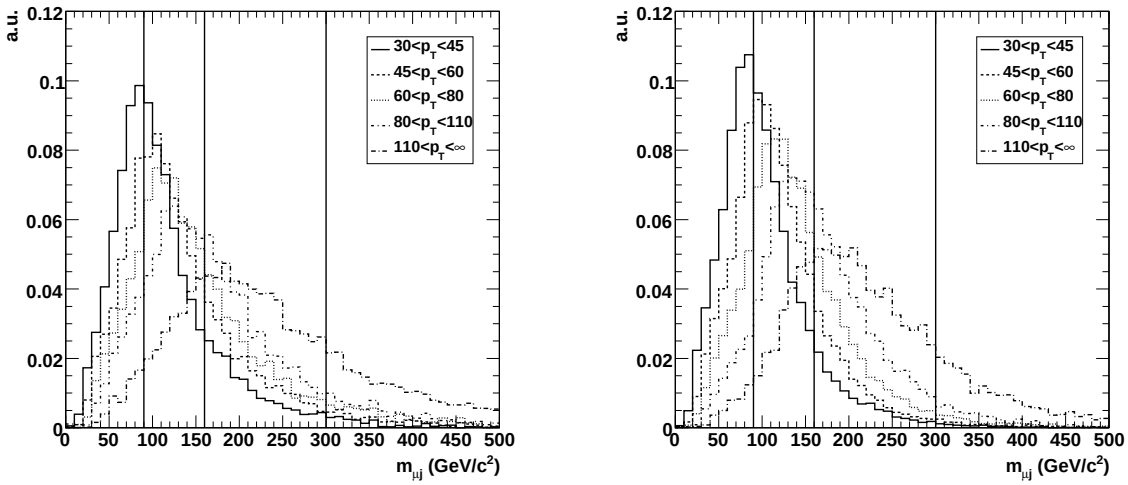

Figure 7.1: The jet-muon mass distribution for the non-b quark jets in the b candidate sample (left) and for all reweighted jets in the control sample (right), differentiated in bins of the transverse momentum of the jets.

To obtain an unbiased estimation of the b-tag efficiency, by applying Equation 6.3 with

F estimated from a control sample, the jet-muon mass distribution of the non-b quark jets in the b candidate sample should match the jet-muon mass distribution of all jets in the control sample, for each bin.

The jet-muon mass distribution for both the non-b quark jets in the b candidate sample and all the jets in the control sample shift towards a higher average value for higher p_T jets. Although the distributions in both samples shift, to first order, in a similar way, the estimation of the scale factor from the control sample, $\widehat{F}^{rew}$, has a bias with respect to the expected scale factor, F^{exp}, from the non-b quark jets in the candidate sample. In Figure 7.2 the relative bias on the estimation of the scale factor, defined as

$$(\widehat{F}^{rew} - F^{exp})/F^{exp} \tag{7.1}$$

is shown as a function of the average transverse momentum of the jets in a given p_T-bin. It can be seen that the proposed data-driven estimation of the scale factor from the control sample is problematic for most of the bins. It should be remarked that the distribution of the jet-muon mass from jets from the control sample can be validated with data. In Figure 7.2 the relative bias is also shown as a function of the pseudo-rapidity, a similar, although less profound, problematic estimation of the scale factor from the control sample can be seen.

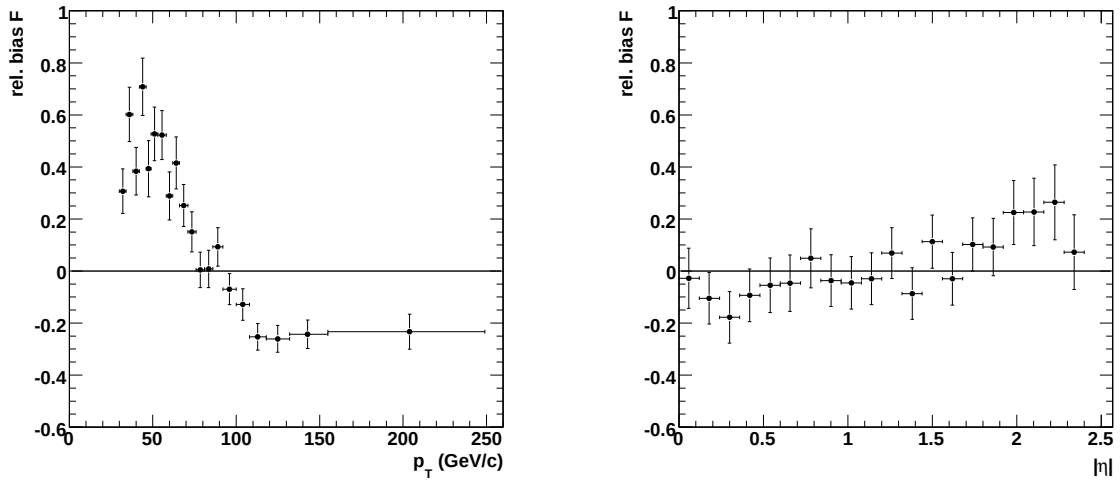

Figure 7.2: Bias on the estimation of the scale factor as a function of the average transverse momentum (left) and the pseudo-rapidity (right) of the jets in a given bin.

The boundaries to define the b-enriched and b-depleted subsamples have been fixed on the inclusive sample and are kept constant for the various bins. This is done to maintain the increased b quark jet purity in the b-enriched subsample and the decreased b quark jet purity in the b-depleted subsample. In Figure 7.3 the jet muon mass for b quark jets in the b candidate sample is shown.

The evolution of the bias on the estimation of the scale factor from the control sample indicates that the jet-muon mass evolves differently, as a function of the p_T, for jets in the control sample compared to the non-b quark jets in the b candidate sample. To quantify this different evolution the mean and the standard deviation of the jet-muon mass is compared between both samples as a function of the average p_T of the jets in a given bin. The

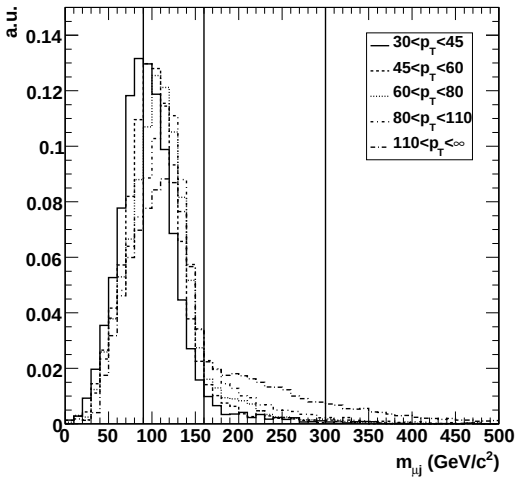

Figure 7.3: The jet-muon mass distribution for b quark jets in the b candidate sample, differentiated in bins of the transverse momentum of the jets.

mean and the standard deviation of the jet-muon mass are calculated with the jets, in a given bin, that yield a jet-muon mass between 90 GeV/ c^2 and 300 GeV/ c^2, the relevant region for the calculation of the scale factor. In Figure 7.4 the dependency of the mean and the standard deviation of the jet-muon mass is shown as a function of the average transverse momentum of the jets in a specific bin. A different evolution is visible for the jets in the control sample with respect to the non-b quark jets in the b candidate sample. The disagreement between the mean of the jet-muon mass shows the same dependency as a function of the p_T as the bias on the scale factor. When the mean of the jet-muon mass is the same in both samples, for example in the bins around 80-100 GeV/ c^2, the estimation of the scale factor is found to be unbiased. The evolution of the mean and the standard deviation of the jet-muon mass of the jets in the control sample can be made with data for validation of the simulation. This can give information whether the scale factor F in the control sample as a function of the transverse momentum is well modeled in the simulated data.

The control sample is composed of both b quark jets and non-b quark jets. The fraction of b quark jets as a function of the average transverse momentum of the jets is shown in Figure 7.5. The fraction is found to change slightly and does not explain the difference in jet-muon mass as is observed in Figure 7.4. In Figure 7.5 also the distribution of the average χ^2_{min} of the events contributing a jet to the control sample and the b candidate sample as a function of the average transverse momentum of the jets is shown. The same tendency as for the jet-muon mass is observed. The χ^2_{min} of the event is found to depend differently on the transverse momentum of the jets for the jets in the control sample and the non-b quark jets in the b candidate sample.

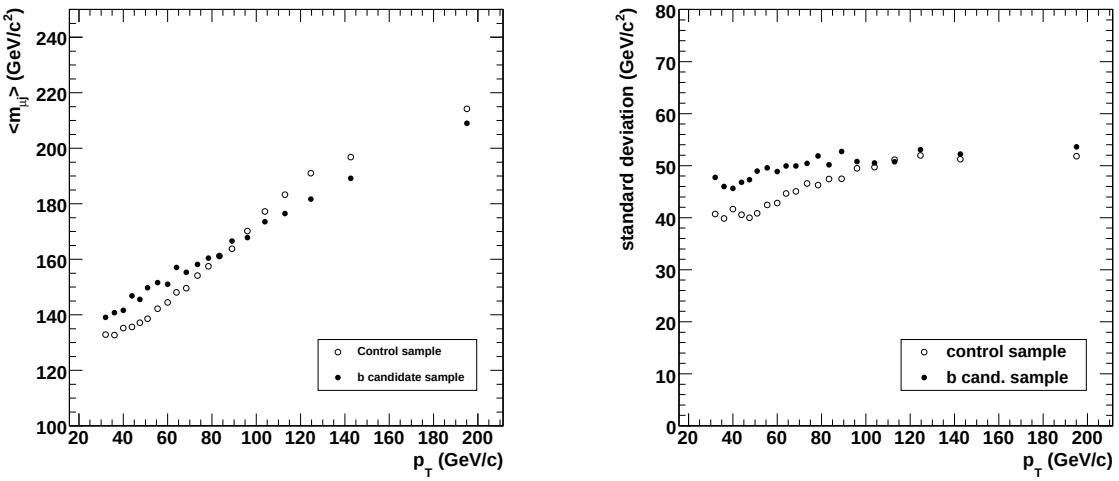

Figure 7.4: The dependency of the mean (left) and the standard deviation (right) of the jet-muon mass as a function of the average transverse momentum of the jets in a given bin for the control sample and the non-b quark jets in the b candidate sample.

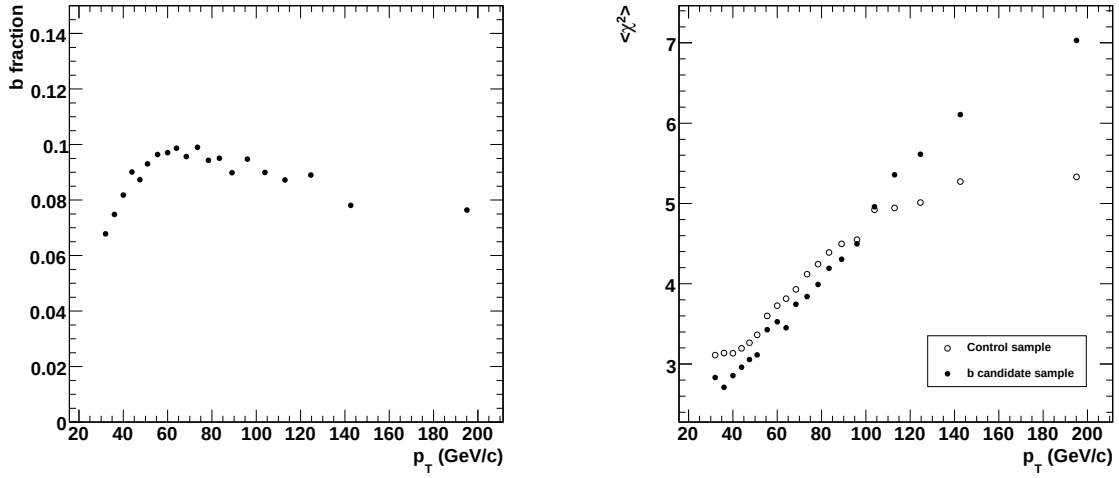

Figure 7.5: The fraction of b quark jets in the control sample (left) and the mean χ^2_{min} of the events (right) as a function of the average transverse momentum of the jets in a given bin.

7.2 Differential estimation of the b-tag efficiency

In the previous section is was shown that the estimation of the scale factor F from a control sample is not feasible when applying the b-tag efficiency measurement differentiated in bins of the transverse momentum. Also as a function of the pseudo-rapidity it is found that the estimation of the scale factor fails. To obtain now an estimation of the b-tag efficiency, the scale factor F is defined as the ratio of non-b quark jets in the b-enriched to the b-depleted sample, is determined from simulation. To acquire this scale factor in a given range of transverse momentum or in a given range of the pseudo-rapidity, the evolution of F is determined as a function of respectively p_T and η. This is done by dividing the b candidate sample in respectively 20 p_T- or 20 η-bins. For each subsample, F is calculated and the

resulting dependency is shown in Figure 7.6. It is fitted with an exponential function and a line for respectively the transverse momentum and the pseudo-rapidity. In contrast to the inclusive method in the previous chapter where the b-depleted subsample ranges up to 300 GeV/ c^2 it is now limited up to 200 GeV/ c^2, in order to not extend too far in the tail of the jet-muon mass distribution for the low p_T bins.

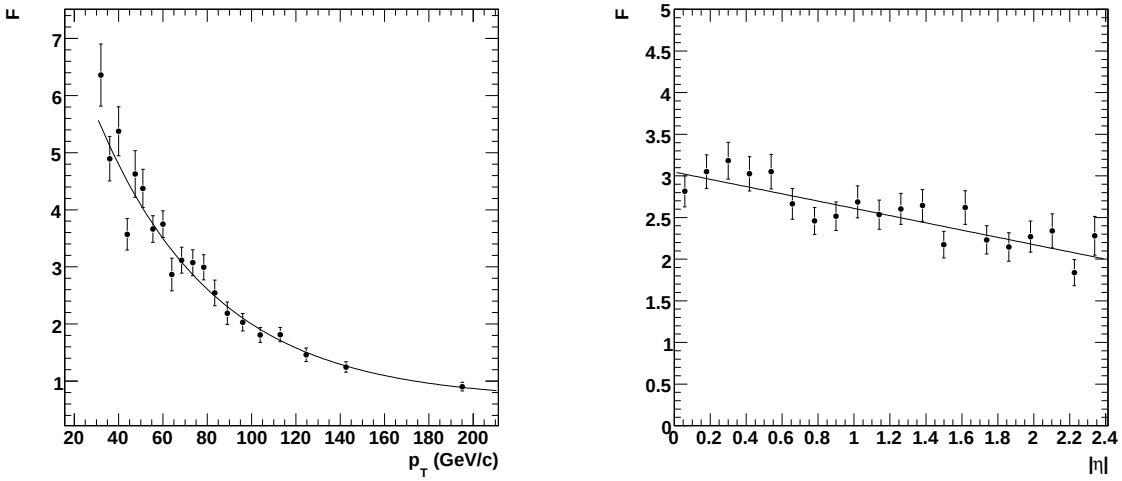

Figure 7.6: The dependency of the scale factor on the transverse momentum (left) and the pseudo-rapidity (right) of the jets.

This fitted functions from simulations now serve as an input to determine the scale factor needed for the estimation of the b-tag efficiency in a given p_T- or η-bin. The differentiated estimation of the b-tag efficiency is now performed by dividing the candidate sample in five bins. For each bin the method as described in Section 6.1.2 is applied where the value of the scale factor F is based on the evaluation of the fitted function at the average transverse momentum or at the average pseudo-rapidity of the jets in that bin.

In Figure 7.7 the estimated and expected b-tag efficiency is shown for the medium working point of the track counting high efficiency b-tagging algorithm as a function of the transverse momentum and the pseudo-rapidity. The relative bias between the estimated and the expected b-tag efficiency defined as,

$$\frac{\widehat{\epsilon}_b^{\,i} - \epsilon_b^{exp,i}}{\epsilon_b^{exp,i}} \tag{7.2}$$

for each bin i and is shown as well in Figure 7.7. The uncertainties reflect the limited size of the simulated sample. No significant bias is observed for any of the p_T- and η-bins given the limited size of the simulated sample except maybe for the highest p_T-bin. The discrepancy in this bin is mainly due to an overestimation of the scale factor. The fit function to derive the scale factor is evaluated at the average transverse momentum of the jets in this bin, this yields a slightly biased scale factor due to the exponential shape of the function. The expected statistical uncertainty for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV is indicated as well. The highest p_T-bin is for convenience only indicated until 200 GeV/ c in the figure, in practice all jets with a transverse momentum higher than 110 GeV/ c have been included in this bin.

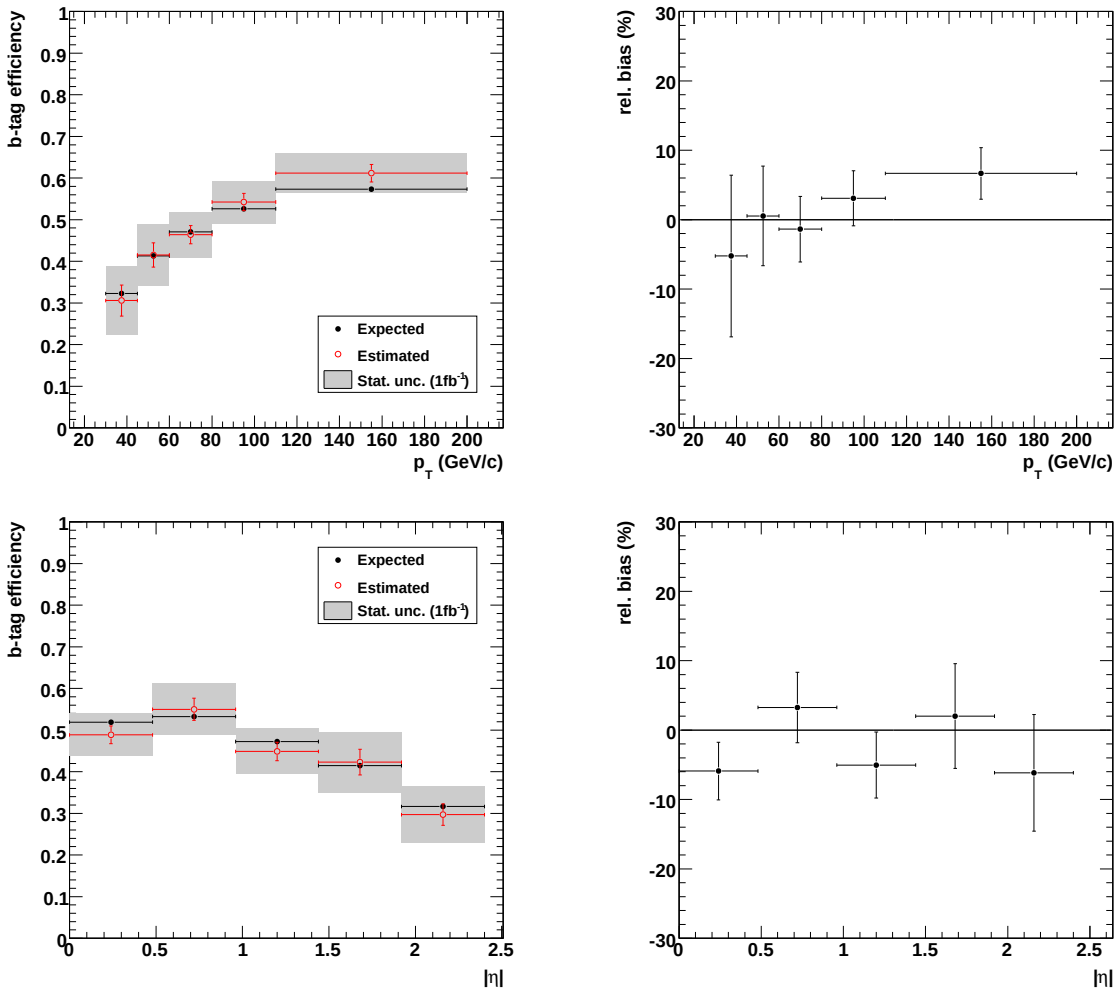

Figure 7.7: The estimated and expected b-tag efficiency (left) for the medium working point of the track counting high efficiency b-tagging algorithm and the relative bias (right), differentiated as a function of the transverse momentum (upper) and the pseudo-rapidity (lower). Uncertainties indicate the limited size of the simulated samples. The expected statistical uncertainty for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV is indicated.

7.3 Systematic uncertainties

The systematic uncertainties on the differential estimation of the b-tag efficiency is evaluated in a similar way as in the previous chapter except that now the scale factor F is obtained from the fitted function obtained from the simulated nominal sample. To obtain the uncertainty due to a given systematic effect the following equation is used,

$$\Delta\epsilon^{i}_{b,sys} = \frac{\widehat{\epsilon}^{\,i}_{b,sys} - \epsilon^{exp,i}_{b,sys}}{\epsilon^{exp,i}_{b,sys}} - \frac{\widehat{\epsilon}^{\,i}_{b,nom} - \epsilon^{exp,i}_{b,nom}}{\epsilon^{exp,i}_{b,nom}} \tag{7.3}$$

where for each bin i the relative difference between the estimated b-tag efficiency $\widehat{\epsilon}^{\,i}_{b,sys}$ and the expected b-tag efficiency $\epsilon^{exp,i}_{b,sys}$ when including a given systematic effect is compared to

the relative difference between the the estimated b-tag efficiency $\widehat{\epsilon}^{\,i}_{b,nom}$ and the expected b-tag efficiency $\epsilon^{exp,i}_{b,nom}$ from the nominal sample. For each systematic effect the obtained systematic uncertainty is compared to the expected statistical uncertainty in each bin for an integrated luminosity of 1 fb^{-1} at 10 TeV.

Jet energy scale

To study the effect of the jet energy scale, the jets have been modified according to the description in Section 6.4.1. The energy scale was increased and decreased with 10% with respect to the nominal samples. In Figure 7.8 the systematic uncertainty is shown for the five p_T- and η-bins. The uncertainties on the observed shifts are due to the limited size of the simulated samples. Although the observed shifts are in general compatible with zero the shift is considered significant since the nominal sample is strongly correlated with the samples with increased and decreased jet energy scale. The average absolute shift is quoted as the systematic uncertainty due to the jet energy scale variation.

Initial and Final state radiation

The systematic uncertainty due to the description of the initial and final state radiation in the simulation is studied using three dedicated samples. These samples have been generated with modified values for the parameters defining the amount of radiation as discussed in Section 3.3.1. Since these samples are compared to a dedicated nominal sample the distribution of F as a function of the pseudo-rapidity and the transverse momentum was obtained from this nominal sample. In Figure 7.8 the systematic uncertainty is shown for the five p_T- and η-bins. The uncertainties reflect the limited size of the simulated samples. Since the samples are independent the uncertainties indicate that the systematic effect due to initial and final state radiation is compatible with zero. Therefore the statistical precision is taken as a measure for the systematic uncertainty.

Event generator

The uncertainty due to the event generator for $t\bar{t}$ events is studied by replacing the nominal $t\bar{t}$ events, generated with the `MadGraph` event generator, by $t\bar{t}$ events generated with the `ALPGEN` event generator. In Figure 7.8 the systematic uncertainty is shown of the five p_T- and η-bins. Due to the limited size of the $t\bar{t}$ event sample generated with the `ALPGEN` event generator large uncertainties on the bias can be observed. Therefore the uncertainty due to the event generator is not taken into account since the shift is found to be compatible with zero.

Background cross sections

The cross sections of the W+jets and Z+jets processes have important uncertainties. This can induce a systematic effect on the estimation of the b-tag efficiency due to the uncertainty of the number of background events. Analogue to the strategy followed for the inclusive estimation, the W+jets and Z+jets samples are scaled with a factor 1.5 to study the effect of this uncertainty. Additionally, the samples with additional heavy flavour jets,

V+qq+jets (V=W/Z) and the W+c+jets, have been added to the samples used in this analysis to study the effect of the amount of background events with a significant heavy flavour content. In Figure 7.9 these effects as a function of the five p_T- and η-bins are shown. Due to the large correlation between the nominal sample and the systematic sample the shifts are considered significant and are added to the systematic uncertainty.

Combination of the systematic uncertainties

The systematic uncertainties on the estimated b-tag efficiency at the medium working point of the track counting high efficiency b-tagging algorithm, obtained in the previous sections, are combined and are shown in Figure 7.9. The dominating systematic uncertainty is the jet energy scale which was taken rather conservatively. Another important contribution to the systematic uncertainty is due to the modeling of the initial state and final state radiation, which is mainly due to the limited size of the systematic sample. The relative bias on the estimation of the b-tag efficiency is considered a systematic uncertainty as well.

The total systematic uncertainty is combined with the expected statistical uncertainty for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV. The total relative uncertainty on the estimation of the b-tag efficiency is found to range from 11% up to about 33% for the b-tag efficiency as a function of the transverse momentum. For the estimation of the b-tag efficiency as a function of the pseudo-rapidity the total relative uncertainty is found to range from 15% up to about 27%.

7.4 Results for other b-tagging algorithms

In the previous chapter it was concluded that the estimation of the b-tag efficiency is applicable for b-tagging algorithms other than the track counting high efficiency b-tagging algorithm. This conclusion still holds when estimating the b-tag efficiency differentiated in bins of the transverse momentum or the pseudo-rapidity of the jet. In Figure 7.10 the b-tag efficiency as a function of the transverse momentum and the pseudo-rapidity is shown for the simple secondary vertex b-tagging algorithm and the combined secondary vertex b-tagging algorithm. The cut on the b-tag discriminator or working point was chosen to yield an average b-tag efficiency of 50%. For both b-tagging algorithms it is found that the estimation of the b-tag efficiency can be made with a similar precision as for the track counting high efficiency b-tagging algorithm. Comparing the estimated b-tag efficiency, differentiated as a function of the transverse momentum, a similar tendency is observed as for the track counting high efficiency b-tagging algorithm.

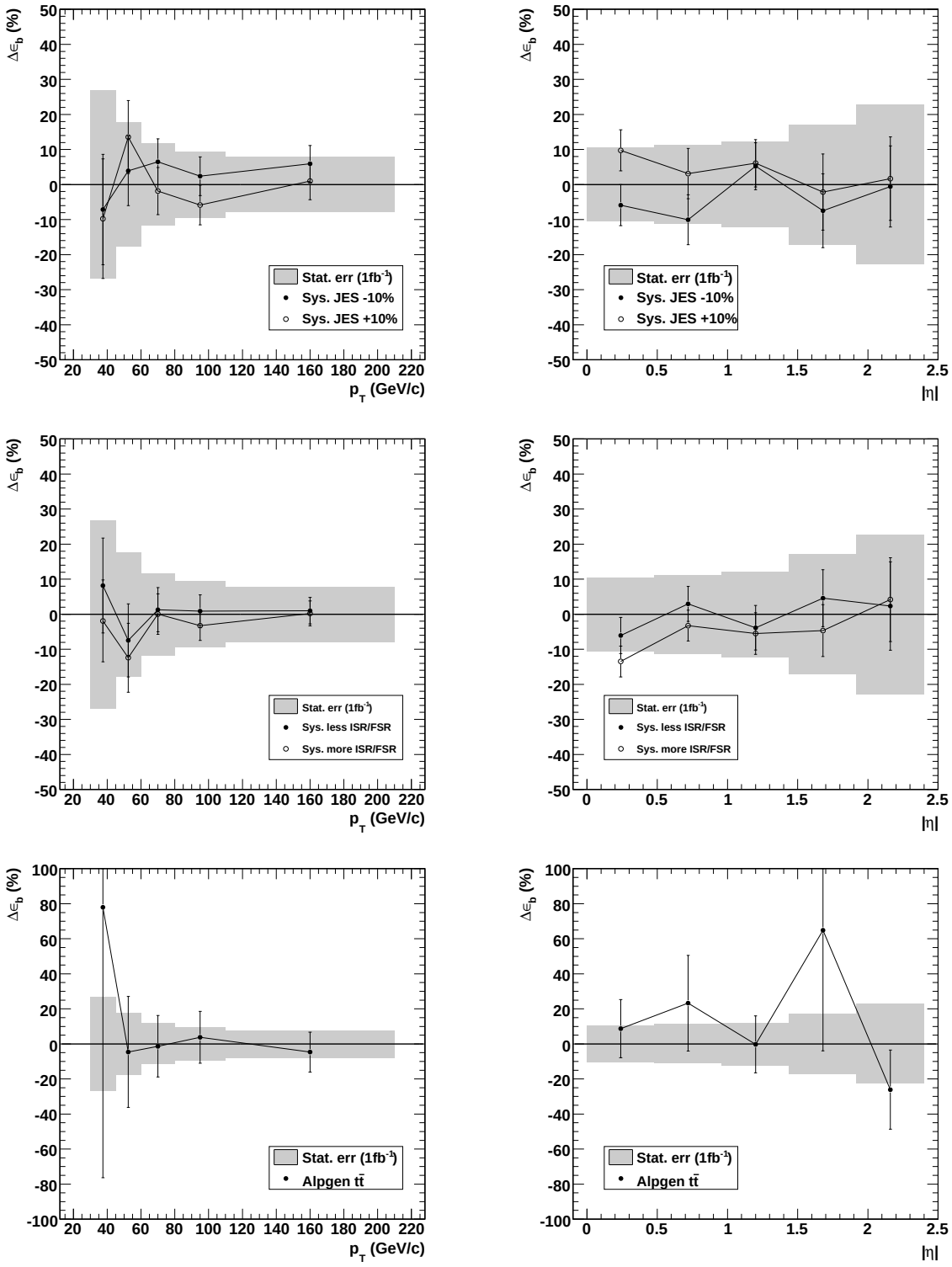

Figure 7.8: Overview of various relative systematic uncertainties as a function of the transverse momentum (left) and the pseudo-rapidity (right). The expected relative statistical uncertainty for an integrated luminosity of 1 fb^{-1} is indicated as well.

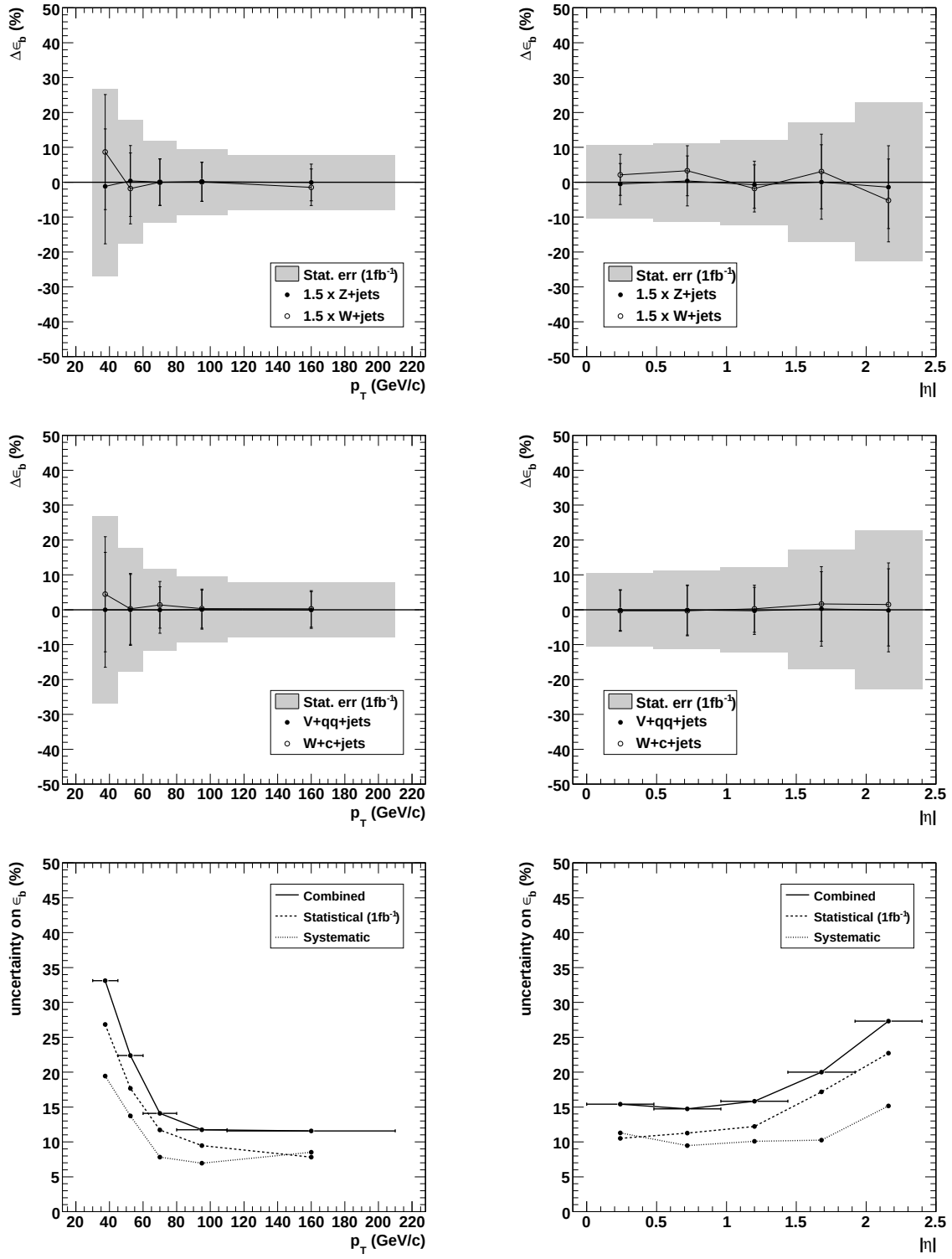

Figure 7.9: Overview of various relative systematic uncertainties as a function of the transverse momentum (left) and the pseudo-rapidity (right). The expected relative statistical uncertainty for an integrated luminosity of 1 fb^{-1} is indicated as well. An overview of the expected relative statistical uncertainty, the combined relative systematic uncertainty and the corresponding total relative uncertainty is shown (lower plots).

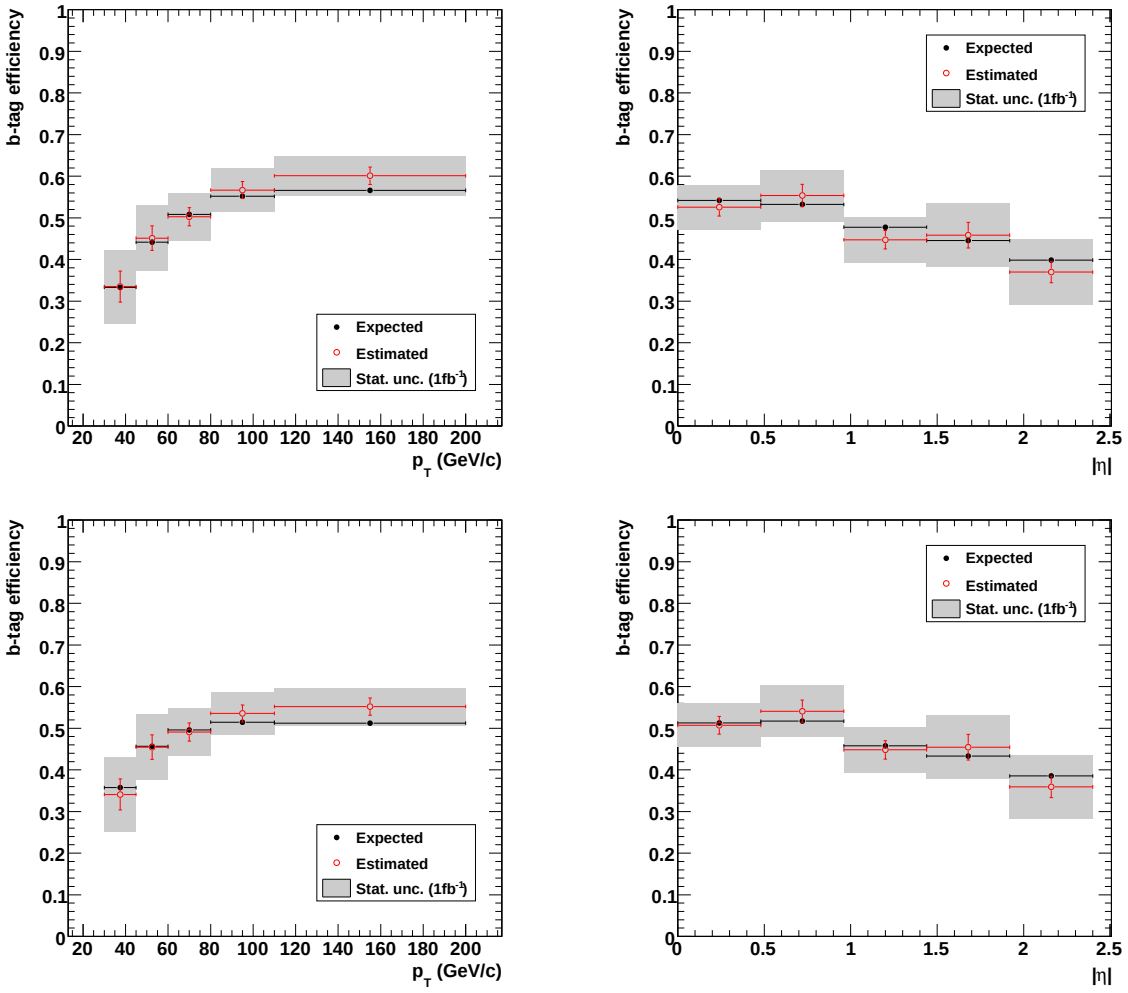

Figure 7.10: Estimation of the b-tag efficiency at a medium working point for the simple secondary vertex b-tagging algorithm (upper) and the combined secondary vertex b-tagging algorithm (lower). Both are differentiated as a function of the transverse momentum (left) and the pseudo-rapidity (right). The expected statistical uncertainty for an integrated luminosity of 1 fb^{-1} for 10 TeV collisions is indicated.

Chapter 8

Conclusions and perspectives

The top quark, the heaviest quark in the Standard Model, was discovered in 1995 by the DØ and CDF experiment at the Tevatron collider. During the following years some of its most important properties, like its mass and decay modes, were measured with good precision. At the LHC, operating at a center-of-mass energy up to 7 times higher than the Tevatron, a very large sample of top quarks will be produced. This allows the CMS experiment and the ATLAS experiment to measure its properties with greater precision and will also allow them to use the top quark as a handle for calibration of reconstruction algorithms. In the Standard Model the top quark decays almost exclusively to a b quark and a W boson providing a large sample of b quarks from top quark decays. Based on this sample of b quarks a method to calibrate the b quark jet identification algorithms, or b-tagging algorithms, on data is developed in this thesis. A good discrimination between jets originating from b quarks and jets not originating from b quarks is required for many measurements within the Standard Model and searches beyond the Standard Model. To meet these requirements, calibration techniques for the b-tagging algorithms based on collision data are necessary.

In Section 8.1 an overview is given of the potential of the method developed in this thesis for an inclusive estimation of the b-tag efficiency. The potential of the method is presented for a data sample of proton-proton collisions corresponding to 1 fb^{-1} at a center-of-mass energy of 10 TeV. Additional to the inclusive estimation of the b-tag efficiency, potential results are summarized for a differential estimation of the b-tag efficiency as a function of the transverse momentum and the pseudo-rapidity of the jets. In Section 8.2 the perspectives are given for the estimation of the b-tag efficiency as a function of the integrated luminosity. The LHC is operating at a center-of-mass energy of 7 TeV for the data taking period of 2010-2011, therefore the prospects are given for the potential of the method at center-of-mass energies other than 10 TeV. In the final section of this chapter a short proposal is made for a method to combine the estimation of the b-tag efficiency with the measurement of the production cross section of $\text{t}\bar{\text{t}}$ pairs.

8.1 Estimation of the b-tag efficiency

8.1.1 Inclusive estimation

In Chapter 4 the reconstruction of high level objects in the CMS experiment is introduced. The jets in the simulated samples used in this thesis have been reconstructed with the Seedless Infrared Safe cone algorithm and are calibrated with level 2 and level 3 jet energy correction factors. The muon is reconstructed with the global muon reconstruction algorithm. To select semi-muonic $t\bar{t}$ events, event selection cuts are applied on these objects as discussed in Chapter 5. By requiring at least four jets with a high transverse momentum and exactly one isolated muon, the large multi-jet, W+jets and Z+jets backgrounds can be significantly reduced. It is expected that for an integrated luminosity of 1 fb^{-1} at 10 TeV center-of-mass energy a sample with about 10249 semi-muonic $t\bar{t}$ events and about 8444 background events is retained after the event selection.

The four selected jets in each event are assigned to the quarks in the final state of the semi-muonic $t\bar{t}$ decay process using a kinematic jet-quark matching algorithm as discussed in Section 5.2. This matching algorithm exploits the top quark mass and the W boson mass constraint in $t\bar{t}$ events. The jet identified as the jet originating from the b quark coming from the top quark associated with the leptonically decaying W boson is then used to construct a b quark jet candidate sample. This sample is divided into two subsamples, one with an enriched b purity and one with a depleted b purity. By subtracting the b-tag discriminator distribution of the jets in the b-depleted subsample from the b-tag discriminator distribution of the jets in the b-enriched sample, the b-tag discriminant distribution for b quark jets is obtained. This needs to be carefully balanced to cancel the contribution of non-b quark jets, therefore a scale factor is introduced. This scale factor is obtained in a data-driven way from a control sample with a very low b quark jet content.

This results in an estimation of the b-tag discriminant distribution of b quark jets which is used to estimate the b-tag efficiency at any given working point. The estimated b-tag efficiency for the track counting high efficiency b-tagging algorithm at a loose, medium and tight working point[1] is found to be, with absolute statistical and systematic uncertainty,

$$\widehat{\epsilon}_b(\text{loose}) = 75.4 \pm 2.9\ (\text{stat}) \pm 3.2\ (\text{sys})\ \%, \quad (8.1)$$

$$\widehat{\epsilon}_b(\text{medium}) = 49.9 \pm 2.2\ (\text{stat}) \pm 2.5\ (\text{sys})\ \%, \quad (8.2)$$

$$\widehat{\epsilon}_b(\text{tight}) = 24.5 \pm 1.4\ (\text{stat}) \pm 1.5\ (\text{sys})\ \%. \quad (8.3)$$

Given the statistical uncertainty due to the limited size of the simulated sample, a small bias is observed on the estimated b-tag efficiency. This bias is absorbed in the systematic uncertainty. The systematic uncertainties on the jet energy scale and the cross section of background processes is taken rather conservative. It is expected that with an increasing amount of collision data these systematic uncertainties will be reduced. The systematic uncertainty due to the modeling of the initial and final state radiation is taken very conservative due to the limited size of the simulated samples, a larger set of simulated data will provide a more accurate estimation of this systematic uncertainty.

[1] These working point have been chosen to yield a b-tag efficiency of approximately 25%, 50% and 75%. A more precise definition is given in Section 6.1.3.

8.1.2 Differential estimation

In Chapter 7 a method to estimate the b-tag efficiency differentiated as a function of the transverse momentum, p_T, and the pseudo-rapidity, η, of the jets has been developed. It was found that the estimation of the scale factor from a control sample was not feasible when estimating the b-tag efficiency in bins of the transverse momentum or in bins of the pseudo-rapidity. Therefore the scale factor was parametrized as a function of the transverse momentum and the pseudo-rapidity based on the simulation. From this function the scale factor for any given p_T- or η-bin can be obtained. Since the scale factor is obtained from simulations the differential estimation of the b-tag efficiency is found to be more sensitive to systematic uncertainties such as systematic effects related to the jet energy scale uncertainties and effects due to the modeling of the initial and final state radiation. In Figure 8.1 the differential estimation of the b-tag efficiency is shown. The uncertainties indicate the statistical precision expected for a sample corresponding to an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV. Additionally the total uncertainty, equal to the quadratic sum of the statistical uncertainty and the systematic uncertainty, is shown. The distributions of the estimated b-tag efficiency are found to show the expected tendency as observed in Section 4.3.5.

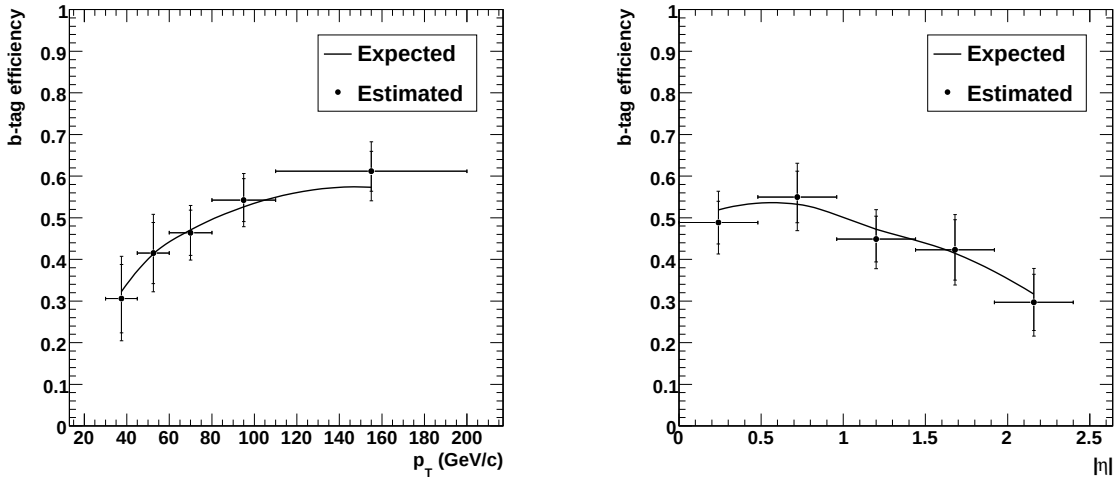

Figure 8.1: Estimation of the b-tag efficiency differentiated as a function of the transverse momentum (left) and the pseudo-rapidity (right) of the jets. The uncertainties indicate the expected statistical uncertainty for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV and the total uncertainty by adding the systematic uncertainty to the statistical uncertainty.

8.2 Perspectives

The method to estimate the b-tag efficiency from top quark events in this thesis has been developed on simulated events reflecting an integrated luminosity of 1 fb^{-1} for a center-of-mass energy of 10 TeV. The LHC is designed to operate at a center-of-mass energy of 14 TeV, however currently the LHC is operating at a center-of-mass energy of 7 TeV.

It is therefore relevant to extrapolate the potential performance of the method at other center-of-mass energies and other integrated luminosities.

8.2.1 Potential performance at other integrated luminosities

Figure 8.2 shows an extrapolation of the systematic and statistical uncertainties on the inclusive estimation of the b-tag efficiency for the medium working point of the track counting high efficiency b-tagging algorithm as a function of the integrated luminosity collected at a center-of-mass energy of 10 TeV. The statistical uncertainty obtained for an integrated luminosity of 1 fb^{-1} is rescaled to obtain the expected uncertainty at other integrated luminosities. The systematic uncertainty is taken to remain constant. A conservative approach was adopted for determining the systematic uncertainties. It is expected that with an increasing amount of collected data the uncertainty on the jet energy scale and the background cross sections will be reduced leading to a smaller systematic uncertainty. The statistical uncertainty and the systematic uncertainty are found to be of equal size at an integrated luminosity of 1 fb^{-1}.

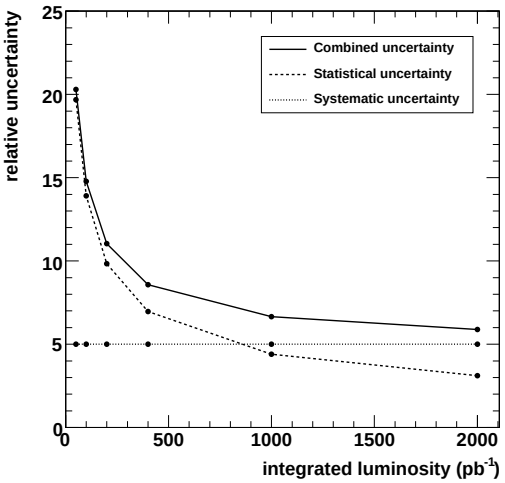

Figure 8.2: Expected relative uncertainty on the inclusive estimation of the b-tag efficiency at the medium working point of the track counting high efficiency b-tagging algorithm as a function of the integrated luminosity for a center-of-mass energy of 10 TeV.

8.2.2 Potential performance at other center-of-mass energies

The method in this thesis was developed on simulated proton collisions at a center-of-mass energy of 10 TeV. For the data taking period of 2010-2011 however, the center-of-mass energy at which the protons collide is 7 TeV. It is foreseen that over this period around 1 fb^{-1} of data will be collected. The principle of the method presented in this thesis is expected to work at this lower center-of-mass energy. The main limiting factor will be the decreased cross section of $t\bar{t}$ production. It is predicted in [106] that the next-to-next-to-leading order cross section of top quark production at a center-of-mass energy of 7 TeV is 170 ± 10 pb, compatible with a reduction of about 2.5 compared to the top quark cross

section predicted at 10 TeV. Similarly to the decrease of the production cross section of top quark pair processes, the production cross section of W+jets events and Z+jets events will decrease as well, albeit with a smaller reduction factor. It is shown in section 6.4 that a relatively larger amount of background events does not significantly influence the estimation of the b-tag efficiency. When the cross section of $t\bar{t}$ processes is decreased by a factor of 2.5 at a center-of-mass energy of 7 TeV a relative statistical uncertainty of 7% is expected on the inclusive estimation of the b-tag efficiency as can be seen in Figure 8.2 by looking at the uncertainty for an integrated luminosity of 400 pb^{-1}. Assuming that the systematic uncertainty remains constant, the total relative uncertainty for an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 7 TeV is 8.5%.

The LHC is designed to operate at a center-of-mass energy of 14 TeV, it is expected that at this collision energy the cross section for top quark pair processes increases with a factor 2.2 with respect to the cross section at 10 TeV. The relative statistical uncertainty expected at a center of mass energy of 14 TeV for 1 fb^{-1} is approximately 3%. In the conservative assumption that no improvement is made for the systematic uncertainty a total relative uncertainty of about 6% can be expected.

8.2.3 Combination with the estimation of the $t\bar{t}$ production cross section

A potential extension of the method to estimate the b-tag efficiency is a simultaneous measurement of the top quark pair production cross section and the b-tag efficiency. In [107] the top quark pair production cross section is estimated by performing a template fit to the *M3* distribution. The *M3* variable is the mass of the three jets with the highest vectorial sum p_T. Three template functions, obtained from simulations, are used; one for the $t\bar{t}$ contribution, one for the W+jets and one for the single top events.

Rather than fitting these components to the observed *M3* distribution in data, a fit to the jet-muon mass distribution for the jets in the b candidate sample could be performed. This would require templates obtained from simulations for the jet-muon mass distribution. The distribution of the jet-muon mass for $t\bar{t}$ events and for single top quark, W+jets and Z+jets events are shown in Figure 8.3. In the method to estimate the b-tag efficiency developed in this thesis a control sample with a very low b quark jet content is constructed. This control sample is used to estimate in a data-driven way the contribution to the jet-muon mass of non-b quark jets in the b candidate sample. The jet-muon mass of W+jets and the less important Z+jets events is mainly containing non-b quark jets. Therefore the jet-muon mass constructed in the control sample could potentially serve as a data driven template for the jet-muon mass distribution of the single top quark, the W+jets and the Z+jets background events. In Figure 8.3 the distribution is shown of the jet-muon mass of the single top quark, W+jets and Z+jets events contributing to the b candidate sample. It is compared to the jet-muon mass of all jets from all processes in the control sample. A good similarity is observed indicating that it is feasible to obtain the jet-muon mass template for single top quark, W+jets and Z+jets events in a data-driven way.

In Figure 8.4 a simultaneous estimation of the b-tag efficiency and the top quark pair production cross section is shown for 600 pseudo-experiments corresponding to an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV. The figure displays the b-

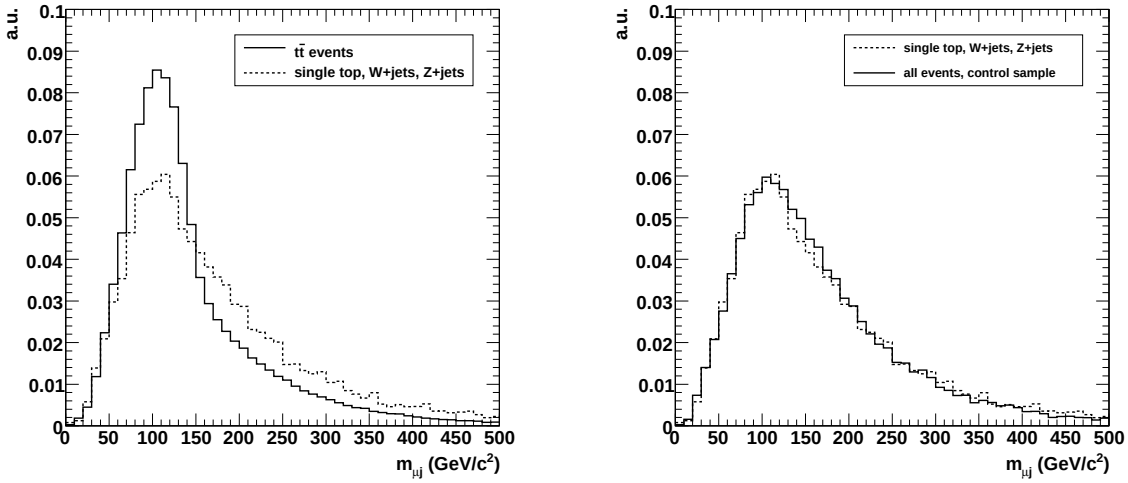

Figure 8.3: The normalized jet-muon mass distribution for $t\bar{t}$, single top quark, W+jets and Z+jets in the b candidate sample (left). The comparison between the normalized jet-muon mass distribution for single top quark, W+jets and Z+jets in the b candidate sample and the normalized jet-muon mass distribution for all jets in the control sample (right).

tag efficiency estimation for the track counting high efficiency b-tagging algorithm at the medium working point. The estimation of the number of $t\bar{t}$ events is performed based on a simultaneous fit of a template for $t\bar{t}$ events from simulation and a template for non-$t\bar{t}$ events obtained from the control sample as explained in the previous paragraph. The estimated number of $t\bar{t}$ events is divided by the expected number of $t\bar{t}$ events. This result shows that it is feasible to estimate the production cross section for $t\bar{t}$ pairs simultaneously with the estimation of the b-tag efficiency.

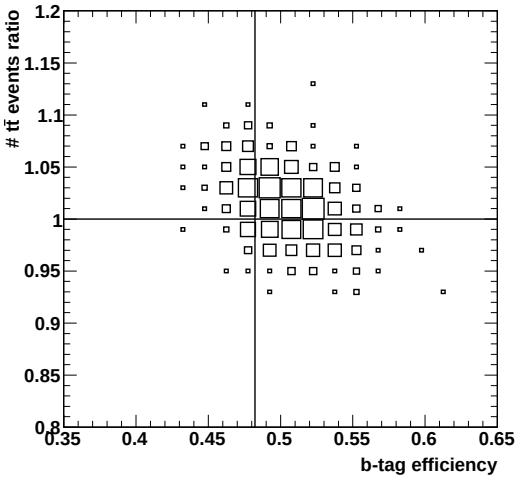

Figure 8.4: A simultaneous estimation of the b-tag efficiency and the estimated number of top quark events divided by the expected number of top quark events, for 600 pseudo-experiments corresponding to an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV.

Bibliography

[1] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*, USA: Addison-Wesley (1995) 842 p.

[2] F. Halzen and A. D. Martin, *Quarks and Leptons: an Introductory Course in Modern Particle Physics*, New York, Usa: Wiley (1984) 396p.

[3] S. Weinberg, *The Quantum theory of fields. Vol. 1: Foundations*, Cambridge, UK: Univ. Pr. (1995) 609 p.

[4] Particle Data Group Collaboration, C. Amsler *et al.*, Phys. Lett. **B667** (2008) 1.

[5] S. Weinberg, Phys. Rev. Lett. **19** (1967) 1264–1266.

[6] A. Salam and J. C. Ward, Phys. Lett. **13** (1964) 168–171.

[7] S. L. Glashow, Nucl. Phys. **22** (1961) 579–588.

[8] F. Englert and R. Brout, Phys. Rev. Lett. **13** (1964) 321–322.

[9] P. W. Higgs, Phys. Rev. Lett. **13** (1964) 508–509.

[10] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, Phys. Rev. Lett. **13** (1964) 585–587.

[11] G. 't Hooft and M. J. G. Veltman, Nucl. Phys. **B44** (1972) 189–213.

[12] *The LEP Electroweak Working Group*, `http://lepewwg.web.cern.ch/LEPEWWG/`.

[13] S. F. King, Contemporary Physics **48 (4)** (2007) pp. 195–211.

[14] A. M. Szelc, Acta Phys. Polon. **B41** (2010) 1417–1440.

[15] S. P. Martin, *A Supersymmetry Primer*, `hep-ph/9709356`.

[16] L. Randall and R. Sundrum, Phys. Rev. Lett. **83** (Oct, 1999) 3370–3373.

[17] K. D. Lane, *Technicolor 2000*, `hep-ph/0007304`.

[18] L. Evans, (ed.) and P. Bryant, (ed.), JINST **3** (2008) S08001.

[19] *The D0 experiment*, `http://www-d0.fnal.gov/`.

[20] *The CDF experiment*, `http://www-cdf.fnal.gov/`.

[21] *The Tevatron Collider*, http://www-bdnew.fnal.gov/tevatron/.

[22] A. Quadt, The European Physical Journal C - Particles and Fields **Volume 48, Number 3** (2006).

[23] CDF Collaboration, T. E. W. Group, *Combination of CDF and D0 Measurements of the Single Top Production Cross Section*, 0908.2171.

[24] CDF Collaboration, and others, *Combination of CDF and D0 Results on the Mass of the Top Quark*, 1007.3178.

[25] J. D. Hobbs, M. S. Neubauer, and S. Willenbrock, *Tests of the Standard Electroweak Model at the Energy Frontier*, 1003.5733.

[26] CMS Collaboration, *Measurement of jet energy scale corrections using top quark events*, CMS-PAS-TOP-07-004.

[27] CMS Collaboration, S. Lowette, J. D'Hondt, J. Heyninck, and P. Vanlaer, *Offline Calibration of b-Jet Identification Efficiencies*, CERN-CMS-NOTE-2006-013.

[28] J. M. Campbell, J. W. Huston, and W. J. Stirling, Reports on Progress in Physics **Volume 70, Number 1** (2007).

[29] CMS Collaboration, R. Adolphi *et al.*, JINST **3** (2008) S08004.

[30] ATLAS Collaboration, G. Aad *et al.*, JINST **3** (2008) S08003.

[31] LHCb Collaboration, A. A. Alves *et al.*, JINST **3** (2008) S08005.

[32] ALICE Collaboration, K. Aamodt *et al.*, JINST **3** (2008) S08002.

[33] TOTEM Collaboration, G. Anelli *et al.*, JINST **3** (2008) S08007.

[34] LHCf Collaboration, O. Adriani *et al.*, JINST **3** (2008) S08006.

[35] CMS Collaboration, G. L. Bayatian *et al.*, *CMS physics: Technical design report*, CERN-LHCC-2006-001.

[36] CMS Collaboration, *Track reconstruction in the CMS Tracker*, note in preparation.

[37] R. Fruhwirth, Nucl. Instrum. Meth. **A262** (1987) 444–450.

[38] *Tracking and Vertexing Results from First Collisions*,.

[39] W. Waltenberger, *Adaptive vertex reconstruction*, CERN-CMS-NOTE-2008-033.

[40] CMS Collaboration, *Performance of Jet Algorithms in CMS*, CMS-PAS-JME-07-003.

[41] D. Bonacorsi, Nucl. Phys. Proc. Suppl. **172** (2007) 53–56.

[42] A. Fanfani *et al.*, J. Grid Comput. **8** (2010) 159–179.

[43] I. Bird, (ed.) *et al.*, *LHC computing Grid. Technical design report*, CERN-LHCC-2005-024.

[44] G. Bagliesi *et al.*, NSSC (2008).

[45] J. M. Hernandez *et al.*, J. Phys. Conf. Ser. **119** (2008) 052019.

[46] A. Mohapatra *et al.*, Nucl. Phys. Proc. Suppl. **177-178** (2008) 324–325.

[47] T. Sjostrand, *Monte Carlo Tools*, 0911.5286.

[48] M. A. Dobbs, S. Frixione, E. Laenen, K. Tollefson, H. Baer, E. Boos, B. Cox, R. Engel, W. Giele, J. Huston, S. Ilyin, B. Kersevan, F. Krauss, Y. Kurihara, L. Lonnblad, F. Maltoni, M. Mangano, S. Odaka, P. Richardson, A. Ryd, T. Sjostrand, P. Skands, Z. Was, B. R. Webber, and D. Zeppenfeld, *Les Houches Guidebook to Monte Carlo Generators for Hadron Collider Physics*,.

[49] S. Agostinelli *et al.*, Nuclear Instruments and Methods **A 506** (2003) 250–303.

[50] T. Sjostrand, S. Mrenna, and P. Z. Skands, JHEP **05** (2006) 026.

[51] G. Corcella *et al.*, *HERWIG 6.5 release note*, hep-ph/0210213.

[52] S. Frixione and B. R. Webber, *The MC@NLO 3.3 event generator*, hep-ph/0612272.

[53] F. Maltoni and T. Stelzer, JHEP **02** (2003) 027.

[54] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, JHEP **07** (2003) 001.

[55] J. Alwall *et al.*, Comput. Phys. Commun. **176** (2007) 300–304.

[56] J. Alwall *et al.*, JHEP **09** (2007) 028.

[57] *The*, http://www.phys.psu.edu/ cteq/.

[58] *The MRST Collaboration*, http://durpdg.dur.ac.uk/hepdata/mrs.html.

[59] J. Pumplin *et al.*, JHEP **07** (2002) 012.

[60] W. K. Tung *et al.*, JHEP **02** (2007) 053.

[61] *On-line PDF Plotting and Calculation*, http://durpdg.dur.ac.uk/hepdata/pdf3.

[62] J. Pumplin *et al.*, Phys. Rev. **D65** (2001) 014013.

[63] A. D. Martin, R. G. Roberts, W. J. Stirling, and R. S. Thorne, Eur. Phys. J. **C28** (2003) 455–473.

[64] V. V. Sudakov, Sov. Phys. JETP **3** (1956) 65–71.

[65] L. N. Lipatov, Sov. J. Nucl. Phys. **20** (1975) 94–102.

[66] V. N. Gribov and L. N. Lipatov, Sov. J. Nucl. Phys. **15** (1972) 438–450.

[67] G. Altarelli and G. Parisi, Nucl. Phys. **B126** (1977) 298.

[68] Y. L. Dokshitzer, Sov. Phys. JETP **46** (1977) 641–653.

[69] P. Bartalini, R. Chierici, and A. De Roeck, *Guidelines for the estimation of theoretical uncertainties at the LHC*, CERN-CMS-NOTE-2005-013.

[70] R. K. Ellis, W. J. Stirling, and B. R. Webber, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **8** (1996) 1–435.

[71] OPAL Collaboration, G. Abbiendi *et al.*, Eur. Phys. J. **C13** (2000) 1–13.

[72] DELPHI Collaboration, P. Abreu *et al.*, Phys. Lett. **B405** (1997) 202–214.

[73] L3 Collaboration, M. Acciarri *et al.*, Phys. Lett. **B476** (2000) 243–255.

[74] ALEPH Collaboration, A. Heister *et al.*, Phys. Lett. **B561** (2003) 213–224.

[75] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, JHEP **01** (2007) 013.

[76] S. Hoche *et al.*, *Matching parton showers and matrix elements*, `hep-ph/0602031`.

[77] B. Andersson, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **7** (1997) 1–471.

[78] J. S. Schwinger, Phys. Rev. **82** (1951) 664–679.

[79] C. Peterson, D. Schlatter, I. Schmitt, and P. M. Zerwas, Phys. Rev. **D27** (1983) 105.

[80] CMS Collaboration, V. Khachatryan *et al.*, *Measurement of the Underlying Event Activity in Proton- Proton Collisions at 0.9 TeV*, `1006.2083`.

[81] M. Cacciari, S. Frixione, M. L. Mangano, P. Nason, and G. Ridolfi, JHEP **09** (2008) 127.

[82] N. Kidonakis and R. Vogt, Phys. Rev. **D78** (2008) 074005.

[83] A. D. Martin *et al.*, Phys. Lett. B (2007) 292.

[84] M. Cacciari, S. Frixione, M. L. Mangano, P. Nason, and G. Ridolfi, JHEP **04** (2004) 068.

[85] A. Giammanco and A. Perrotta, *Fast simulations of the ATLAS and CMS experiments at LHC*, CERN-CMS-CR-2007-010.

[86] D. O. (on behalf of the CMS collaboration), J. Phys.: Conf. Ser. (2010) 219 032053.

[87] M. Hansen *et al.*, *The Top Quark Analysis Framework*, CMS Internal Note IN-2007/068.

[88] B. W. Harris, E. Laenen, L. Phaf, Z. Sullivan, and S. Weinzierl, Phys. Rev. **D66** (2002) 054024.

[89] Z. Sullivan, Phys. Rev. **D70** (2004) 114012.

[90] J. M. Campbell, R. K. Ellis, and F. Tramontano, Phys. Rev. **D70** (2004) 094012.

[91] J. M. Campbell and F. Tramontano, Nucl. Phys. **B726** (2005) 109–130.

[92] J. Ohnemus, Phys. Rev. **D47** (1993) 940–955.

[93] U. Baur, T. Han, and J. Ohnemus, Phys. Rev. **D48** (1993) 5140–5161.

[94] UA1 Collaboration, G. Arnison *et al.*, Phys. Lett. **B132** (1983) 214.

[95] G. C. Blazey *et al.*, *Run II jet physics*, `hep-ex/0005012`.

[96] G. P. Salam and G. Soyez, JHEP **05** (2007) 086.

[97] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Nucl. Phys. **B406** (1993) 187–224.

[98] *Plans for Jet Energy Corrections at CMS*, CMS PAS JME-07-002.

[99] V. Chetluru, F. Pandolfi, P. Schieferdecker, and M. Zielinski, *Jet Reconstruction Performance at CMS*, CMS AN-2009/067.

[100] A. Rizzi, F. Palla, and G. Segneri, *Track impact parameter based b-tagging with CMS*, CERN-CMS-NOTE-2006-019.

[101] C. Saout, A. Scheurer, F.-P. Schilling, and A. Schmidt, *Performance of b-Tagging Algorithms with Realistic Detector Scenarios at CMS*, CERN-CMS-AN-2007-047.

[102] C. Weiser, *A combined secondary vertex based B-tagging algorithm in CMS*, CERN-CMS-NOTE-2006-014.

[103] P. Demin, S. De Visscher, A. Bocci, and R. Ranieri, *Tagging b jets with electrons and muons at CMS*, CERN-CMS-NOTE-2006-043.

[104] CMS Collaboration, *Evaluation of udsg Mistags for b-tagging using Negative Tags*, CMS-PAS-BTV-07-002.

[105] CMS Collaboration, *Performance Measurement of b tagging Algorithms Using Data containing Muons within Jets*, CMS-PAS-BTV-07-001.

[106] N. Kidonakis, *Higher-order corrections to top-antitop pair and single top quark production*, `0909.0037`.

[107] CMS Collaboration, D. Green, F. Yumiceva, G. Hanson, and J. Geng-Yuan, *Early observation of top quark pair production in muon plus jets channel*, CMS AN-2009/048.

Summary

The recently started Large Hadron Collider located at the CERN laboratory near Geneva, collides protons at unprecedented energies. This allows the experiments, like the CMS detector, located at the Large Hadron Collider to study the Standard Model of elementary particles at the TeV scale. It is expected that at this scale discoveries of new physics phenomena will be made and that the existence of the Higgs boson will be confirmed or ruled out. In many data analyses at these experiments the b quark plays an important role in discriminating between interesting signal events and the abundantly produced multi-jet background events which are omnipresent at hadron colliders. To identify jets originating from b quarks several powerful b-tagging algorithms have been developed at CMS making use of the specific properties of b quark jets. A good calibration of the performance of these algorithms is crucial for the success of the LHC physics program. An important challenge is to develop calibration methods that are independent of information obtained from simulation.

In this thesis a data-driven method is developed to calibrate b-tagging algorithms based on events where a top quark pair ($t\bar{t}$) has been created, detected by the Compact Muon Solenoid detector. The top quark decays nearly always to a b quark resulting thus in a large sample of b quark jets in $t\bar{t}$ events. The method is designed to select semi-muonic decaying $t\bar{t}$ events, $pp \rightarrow t\bar{t} \rightarrow bqqb\mu\nu_{\mu}$, and reduce the large amount of background events coming from multi-jet, W+jets and Z+jets processes. This is done by applying dedicated selection criteria based on the topology of the semi-muonic $t\bar{t}$ events. For an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV, about 10000 semi-muonic $t\bar{t}$ events are expected to be selected while the expected number of background events is about 8400. In the selected events the jets are assigned within the expected decay topology of the event using a jet-quark matching algorithm making use of the top quark mass and W boson mass constraint in $t\bar{t}$ events. Based on the assignment of the jets, a jet sample is selected which is found to contain around 33% b quark jets. On this selected jet sample the estimation of the b-tag efficiency is performed.

The method developed in this thesis is designed to estimate the b-tag discriminant distribution for b quark jets by selecting a b-enriched (45%) and a b-depleted (15%) subsample in the sample of selected jets. The contribution of jets not originating from a b quark in the subsamples is estimated in a data-driven way from a control sample. In this way the b-tag efficiency can be estimated from collision data for various b-tagging algorithms. For an integrated luminosity of 1 fb^{-1} at a center-of-mass energy of 10 TeV it is expected that an absolute (relative) statistical uncertainty of respectively 2.9% (3.8%), 2.2% (4.4%) and 1.4% (5.7%) can be reached for a b-tag efficiency of approximately 25%, 50% and 75%. The systematic uncertainty is expected to be respectively 3.2% (4.2%), 2.5% (5.0%) and

1.5% (6.2%), making rather conservative assumptions about the systematic uncertainties.

The method for the inclusive estimation of the b-tag efficiency is extended to perform a differential estimation of the b-tag efficiency as a function of the transverse momentum and the pseudo-rapidity of the jets. The estimation is performed in five bins of the transverse momentum or the pseudo-rapidity and for an average b-tag efficiency of 50%. For a dataset corresponding to 1 fb^{-1} the total relative uncertainty on the estimation of the b-tag efficiency is found to range from 11% up to about 33% for the b-tag efficiency as a function of the transverse momentum. For the estimation of the b-tag efficiency as a function of the pseudo-rapidity the total relative uncertainty is found to range from 15% up to about 27%.

The method developed on simulated proton collisions in this thesis is found to be able to provide an estimation of the b-tag efficiency. Additionally a differential estimation of the b-tag efficiency as a function of the transverse momentum and the pseudo-rapidity of the jets is explored. These results can be applied to calibrate the performance of the b-tagging algorithms in a data-driven way and can be cross-checked with the estimation of the b-tag efficiency obtained with other methods in the CMS experiment.

Samenvatting

Schatting van de b-tag efficiëntie met behulp van top quarks bij CMS

De Large Hadron Collider (LHC), operationeel sinds eind 2009, bevindt zich in het CERN laboratorium nabij Genève en produceert protonbotsingen bij nooit eerder bereikte energieën. Deze hoge botsingsenergieën maken het mogelijk om de fundamentele interacties tussen de elementaire deeltjes te bestuderen. Dit gebeurt door experimenten die zich vlakbij de LHC bevinden, zoals het CMS experiment. Men verwacht dat bij deze energieschaal nieuwe fysica-fenomenen ontdekt zullen worden. In de verschillende studies die uitgevoerd worden in de experimenten bij de LHC speelt het identificeren van b quarks een belangrijke rol om de interessante botsingen te onderscheiden van de veelvuldig geproduceerde multi-jet achtergrondprocessen. Deze laatste zijn alomtegenwoordig bij hadron versnellers zoals de LHC. Om jets afkomstig van b quarks te identificeren, werden verscheidene krachtige b-tagging algoritmes ontwikkeld in het CMS experiment, die gebruik maken van de specifieke kenmerken van b quark jets. Een goede calibratie van de performantie van deze algoritmes is cruciaal voor het succes van het onderzoeksprogramma van de LHC. Een belangrijke uitdaging is het ontwikkelen van calibratietechnieken die onafhankelijk zijn van informatie bekomen uit gesimuleerde protonbotsingen.

In deze thesis werd een methode ontwikkeld die geen gebruik maakt van informatie bekomen uit simulaties om de b quark identificatie algoritmes, ook wel b-tagging algoritmes genoemd, te calibreren. Deze methode is gebaseerd op botsingsgebeurtenissen waar een top quark paar ($t\bar{t}$) gecreëerd werd en die gedetecteerd worden door de CMS detector. De top quark vervalt hoofdzakelijk in een b quark wat resulteert in een groot aantal b quarks in $t\bar{t}$ gebeurtenissen. De methode werd ontwikkeld om semi-muonisch vervallende $t\bar{t}$ gebeurtenissen, $pp \rightarrow t\bar{t} \rightarrow bqqb\mu\nu_{\mu}$, te selecteren en om de grote hoeveelheid achtergrond afkomstig van multi-jet, W+jets en Z+jets gebeurtenissen te reduceren. Dit gebeurt door het toepassen van specifieke selectiecriteria die gebaseerd zijn op de topologie van de semi-muonische $t\bar{t}$ gebeurtenissen. Voor een geïntegreerde luminositeit van 1 fb^{-1} bij een massamiddelpuntsenergie van 10 TeV worden ongeveer 10000 semi-muonische $t\bar{t}$ gebeurtenissen verwacht terwijl het aantal verwachte achtergrond gebeurtenissen gelijk is aan ongeveer 8400. In de geselecteerde gebeurtenissen worden de geobserveerde jets toegewezen aan de quarks aanwezig in het semi-muonisch vervallende $t\bar{t}$ gebeurtenissen met behulp van een specifiek algoritme dat gebruik maakt van de gereconstrueerde top quark massa en de W boson massa in de $t\bar{t}$ gebeurtenissen. Op basis van de toewijzing van

de jets wordt een verzameling jets geselecteerd dat ongeveer 33% b quark jets bevat. Deze verzameling geselecteerde jets wordt gebruikt om de b-tag efficiëntie te schatten.

De techniek werd ontworpen om de verdeling van de b-tag discriminator voor b quark jets te schatten door middel van het selecteren van twee deelverzamelingen van jets, één met een verhoogde fractie b quark jets (45%) en één met een verminderde fractie b quark jets (15%). De bijdrage van jets die niet afkomstig zijn van een b quark wordt gecorrigeerd door een schaalfactor. Deze factor wordt bepaald aan de hand van een controleverzameling jets van data om zodoende geen gebruik te maken van gesimuleerde gegevens. Op deze manier kan de b-tag efficiëntie geschat worden met reële databotsingen, voor verscheidene b-tagging algoritmes. Voor een geïntegreerde luminositeit van 1 fb^{-1} bij een massamiddelpuntsenergie van 10 TeV kan een absolute (relatieve) onzekerheid van respectievelijk 2.9% (3.8%), 2.2% (4.4%) en 1.4% (5.7%) bekomen worden voor een b-tag efficiëntie gelijk aan 25%, 50% en 75%. De systematische onzekerheid is gelijk aan 3.2% (4.2%), 2.5% (5.0%) en 1.5% (6.2%), gebaseerd op conservatieve veronderstellingen in verband met de effecten van de systematische onzekerheden.

De methode voor de inclusieve schatting van de b-tag efficiëntie werd uitgebreid om een schatting te bekomen als functie van de transverse impuls en de pseudo-rapiditeit van de jets. De schatting werd uitgevoerd voor vijf deelverzamelingen van de transverse impuls en de pseudo-rapiditeit en dit voor een gemiddelde b-tag efficiëntie van 50%. Voor een hoeveelheid botsingsgegevens dat correspondeert met 1 fb^{-1} wordt verwacht dat de totale relatieve onzekerheid varieert van 11% tot 33% voor de schatting van de b-tag efficiëntie als functie van de transverse impuls. Voor de schatting als functie van de pseudo-rapiditeit wordt verwacht dat de totale relatieve onzekerheid varieert van 15% tot 27%.

De in deze thesis ontwikkelde methode, gebaseerd op gesimuleerde protonbotsingen, toont aan dat het mogelijk is om een schatting van de b-tag efficiëntie te bekomen zonder gebruik te maken van gesimuleerde botsingen. Verder kan een schatting bekomen worden van de b-tag efficiëntie als functie van de transverse impuls en de pseudo-rapiditeit van de jets. Deze resultaten kunnen toegepast worden om de efficiëntie van b-tagging algoritmes te calibreren zonder gebruik te maken van gesimuleerde gegevens. Ook kunnen de resultaten met betrekking tot de b-tag efficientie vergeleken worden met schattingen bekomen uitgaande van alternatieve methodes.

Acknowledgments

Doing research and writing a PhD thesis can be difficult but is luckily not a solitary job. There are many people that have supported me in one way or another during the past years. Therefore I would like to address a few words of thanks to them.

In the first place I would like to thank the IIHE and the IWT for making it possible to do research and to write a PhD thesis. I'm also very grateful to the members of the jury for their interest and valuable comments, it improved the quality of this manuscript.

A key player in this work is certainly my promotor Jorgen D'Hondt. Jorgen, I have appreciated your enthusiasm, patience and the many fruitful discussions we had, thank you for the excellent support and guidance during the past years.

I have spent four years in an office which I had the luck to share with three fantastic people. Gregory, Ilaria and Petra we have shared many hours silently working or loudly discussing about physics or whatever came to our minds, this has made working very pleasant. Petra, it was really nice to study and work together with you, we were often on the same wavelength, many thanks for all the good conversations. Ilaria, I very much appreciated your opinion on various non-physics topics, many thanks for broadening my view on the world. Gregory, you were always in for a joke and ready to have some fun, thanks pal, I will never forget the nice time we have spent together.

Within the top quark group in Brussels I have met some great people. Eric, you have been there during a crucial period of my research, you were a very valuable resource of advice, thank you for that! Jan, during the first year you often walked into our office to find yourself trapped for hours, thanks for your patience with a starting PhD student. Steven, thanks for helping me and introducing me to the b-tag community. Catherine, Stéphanie and Volker, thanks for the invaluable insights from senior researchers, also thanks to Nadjieh and Maryam for the usefull discussions. Alexis, Michael and Stijn, many thanks for your enthusiasm and distraction during the writing phase of my thesis, it is really appreciated.

I'm also grateful to many people at the IIHE. In particular Shkelzen, Olivier, Stijn, Stéphane, Danny and Abdel, many thanks for solving all computing issues at nearly any moment of the day. Marleen, Rosine, Daisy and Annie, a warm thanks for the friendly help with the administration. Also thanks to some very nice colleagues I have met during my years at the lab, Otman, Julie, Sherif, Mateusz, Vincent, Sven,... it was nice to have you all around.

In the international CMS collaboration I have had the pleasure to work together with Wolf Behrenhoff, Daniele Bonacorsi, Juka Klem, Peter Kreuzer, James Letts, Lukas Vanelderen and Christoph Wissing for various computing tasks. I'm also thankful to Thomas Speer for many discussions on b-tagging. Many others were always available for help and discussions about physics, Wolfgang Adam, Tommaso Boccali, Andrea Giammanco, Gena Kukarzev,

Mathias Mozer, Andrea Rizzi, Pascal Vanlaer, Roger Wolf and Fransisco Yumiceva thanks, for the nice cooperation.

It is always a relief to disconnect from the work and meet my dearest friends, Raf, Wim, Lander, Koen, Pieter and Lennart, many thanks for all the nice times, I think I owe you all a drink, at least one. A very special thanks goes to Sara for being there during these intense times of writing a PhD. Last but certainly not least I am grateful to my parents, my sisters and my brother for supporting me during my studies and during my PhD, thank you!

List of publications

A list of papers to which I have made significant contributions is given below. Additionally to this list, I'm co-author of 31 papers published by the CMS experiment.

1. Maes, J., *Leptons, b-tagging and MET reconstruction at CMS after the first data*, Nuovo Cim., (2010), *to be published*.

2. Fanfani, A. *et al.*, *Distributed analysis in CMS*, J.Grid Comput. **8**, (2010) 159-179.

3. Bagliesi, G. *et al.*, *Debugging Data Transfers in CMS*, CMS Conference Report CR-2009/093, (2009).

4. Andreeva, J. *et al.*, *CMS Analysis Operations*, CMS Conference Report CR-2009/088, (2009).

5. Hernandez, J. M. *et al.*, *CMS Monte Carlo production in the WLCG computing grid*, J. Phys. Conf. Ser., (2008) 119:052019.

6. Mohapatra, A. *et al.*, *CMS Monte Carlo production operations in a distributed computing environment*, Nucl. Phys. Proc. Suppl., (2008) 177-178.

7. Hansen, M. *et al.*, *The Top Quark Analysis Framework*, CMS Internal Note IN-2007/068, (2007).

8. Bagliesi, G. *et al.*, *The CMS data transfer test environment in preparation for LHC data taking*, NSS-IEEE, (2008) 3475-3482.

9. Maes, J., *Measurements of the b-tag performance from data in CMS*, Nuovo Cim. **123B**, (2008) 1305-1307

10. D'Hondt, J. *et al.*, *Measurement of flavour tagging efficiencies using top quark events in CMS*, CMS Analysis Note AN-2007/030.

11. Sevrin, A. and Maes, J., *A Note on N=(2,2) superfields in two dimensions*, Phys. Lett. **B642** (2006) 535-539.

Het denken mag zich nooit onderwerpen,
noch aan een dogma,
noch aan een partij,
noch aan een hartstocht,
noch aan een belang,
noch aan een vooroordeel,
noch aan om het even wat,
maar uitsluitend aan de feiten zelf,
want zich onderwerpen betekent het einde van alle denken.

Henri Poincaré
21 november 1909

Uit een redevoering ter gelegenheid van de 75ste verjaardag van de Université Libre de Bruxelles.