Early observation of t-quarks in single-leptonic $t\bar{t}$ decays at CMS

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde (Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Dimitrios Tsirigkas

aus

Griechenland

Promotionskomitee

Prof. Dr. Claude Amsler (Leitung der Dissertation)

Prof. Dr. Vincenzo Chiochia

Dr. Martijn Mulders

Prof. Dr. Luigi Rolandi

Zürich, 2009

Abstract

This thesis investigates the possibility of identifying a significant top quark signal and measuring the production cross-section using the first $O(10\ \mathrm{pb}^{-1})$ of data to be collected by the CMS experiment. This will be among the first goals of the experiment and will be possible due to the large cross-section and the characteristic experimental signature of the single-muon decay channel (one isolated muon and four jets). The motivation comes from both physics and detector commissioning reasons. From the physics point of view, it is important to confirm the Standard Model prediction for the cross-section. Knowledge of the cross-section is also important to Higgs boson searches, to which $t\bar{t}$ production is a background. From the commissioning point of view, once a signal-rich sample has been identified, it will be possible to reconstruct the invariant mass of the hadronically decaying W-boson and use it to correct the Monte Carlo-based jet calibration.

In the early period of data taking knowledge of the CMS detector will be limited, the analysis should thus only rely in simple reconstructed quantities. The main sources of systematic uncertainties will be the jet energy scale, the startup detector conditions (tracker misalignment and calorimeter miscalibration) and pileup events (the importance of which will depend on the initial LHC machine parameters). Due to the large $t\bar{t}$ production cross-section, the systematic uncertainty is expected to surpass the statistical uncertainty soon after the LHC startup. Studying these effects was therefore essential for an early measurement. Such studies can be performed by comparing a Monte Carlo sample produced in the presence of the effect to one produced in its absense, from the same initial set of generator-level events. Studying multiple systematic effects thus calls for producing multiple Monte Carlo samples. This was only feasible using the Fast Simulation software of the experiment, which was incompatible with the analysis software of the top quark physics group when the current work began. The first step was therefore to interface the analysis software to the Fast Simulation, a task which did not only serve this work, but also other top quark-related measurements in different channels. Furthermore, it was in the context of this thesis that the Fast Simulation was first made to run under startup detector conditions. After this was achieved, all the samples used in this analysis were locally produced using the CERN computing facilities.

Having addressed the technical issues, the focus moved into defining a selection strategy to ensure the collection of a sample containing a significant $t\bar{t} \rightarrow W^-(\rightarrow \mu\nu)bW^+(\rightarrow q\bar{q}')\bar{b}$ signal. There were two steps to this process. The first was to find efficient ways to suppress the background from other physical processes. Each of the major sources of background (W-boson and QCD multijet events as well as single top events) was treated separately and a different method to control it was developed. With the background under control, an investigation into different ways to assign the jets in a selected event to the final state partons was made. This is necessary in order to determine the three jets from the hadronic top quark decay and reconstruct the invariant mass. A number of different options were considered and systematically compared. It was found that the definition of certain jet-parton matching criteria can not only reduce the combinatorial background, but also the W-boson, QCD and single top background, leading to a final signal purity of up to 72%.

With a high-purity sample ensured, the possibility of an early "rediscovery" of the top quark was investigated. It was found that acquiring a statistically significant (5σ) signal will indeed be possible using the first $O(10\mathrm{pb}^{-1})$ of integrated luminosity. Three methods of measuring the $t\bar{t}$ production cross-section with this amount of data were explored. The first is a simple event-counting method, in which the observed excess of events is attributed to the signal. The other two involve a maximum likelihood fit of the top invariant mass and the jet multiplicity distributions obtained from the Monte Carlo to the experimental data. The best result was obtained from the jet multiplicity fit. The corresponding statistical uncertainty is 11% and the systematic uncertainty 33%. The most important source of systematic uncertainty was found to be the jet energy scale. Reducing the sensitivity to the jet energy scale is possible at the cost of a larger statistical uncertainty. Either way, beyond the first $O(10\ \mathrm{pb}^{-1})$, reducing the jet energy scale uncertainty becomes more important than reducing the statistical uncertainty.

Contents

Chapter 1

Introduction

1.1 A brief history of the top quark

In 1977 the E288 experiment at Fermilab discovered a resonance with a mass of 9.5 GeV [1]. This resonance was named Υ and was soon identified as a bound state of a new elementary particle and its antiparticle. This new particle was the b-quark. Studies to determine the properties of this new quark followed. Soon its charge was determined to be $Q^b = -\frac{1}{3}$ and its weak isospin $T_3^b = -\frac{1}{2}$. It therefore became obvious that the b-quark was one of the two members of a doublet, the other member of which had $Q^b = +\frac{2}{3}$ and weak isospin $T_3^b = +\frac{1}{2}$ and completed the three-generation structure of the Standard Model.

The search for the t-quark started almost immediately after the discovery of the b. Experiments at e^+e^- colliders looked for pair production in $e^+e^- \rightarrow t\bar{t}$ events and already in 1985 PETRA placed a $m_t > 23.3$ GeV limit on its mass [2]. This limit was further improved by subsequent similar searches reaching 30.2 GeV at TRISTAN [3] and 45.8 GeV at LEP [4]. Meanwhile, searches for top quarks in W decays were already taking place at hadron colliders. After a false claim to evidence for a top quark of 40 GeV in 1984, the UA1 and UA2 experiments at CERN had excluded a $m_t < 69$ GeV top quark by 1990 [5][6]. In the meantime, the CDF experiment had joined the search and it was soon discovered that the top quark had a mass higher than 91 GeV [7], practically eliminating the possibility of observation in W decays.

Subsequent research was therefore limited to the Tevatron. In the beginning of 1994 DØ raised the top mass limit to 131 GeV [8] and in the middle of the year CDF published the results of the first fruitful search [9]. An excess of events in the final states with two opposite sign leptons and with one lepton plus jets compared to top-less predictions was observed. This excess translated to 2.8 σ evidence. Fitting the mass in the seven events that were possible to fully reconstruct as $t\bar{t}$ yielded a first direct indication of the top mass, $m_t = 174 \pm 10^{+13}_{-12}$ GeV. This was fully consistent with previous precision electroweak measurements at LEP placing the top mass at 177^{+11+18}_{-11-19} (see section 1.2 and [10]). The discovery finally came in 1995 [11],[12] and was accompanied by a $m_t = 176^{+8+10}_{-8-10}$ GeV measurement of the mass and a $6.8^{+3.6}_{-2.4}$ pb measurement of the cross-section, which was fully consistent with the Standard Model.

Important progress has been achieved since then. Figure 1.1 shows the results of a list of mass measurements undertaken in Tevatron. The current world average, including Tevatron Run I and Run II data is $m_t = 173.1 \pm 0.6 \pm 1.1$ while intensive efforts to determine the $t\bar{t}$ cross-section and the other properties of the top remain well in agreement with Standard Model predictions.

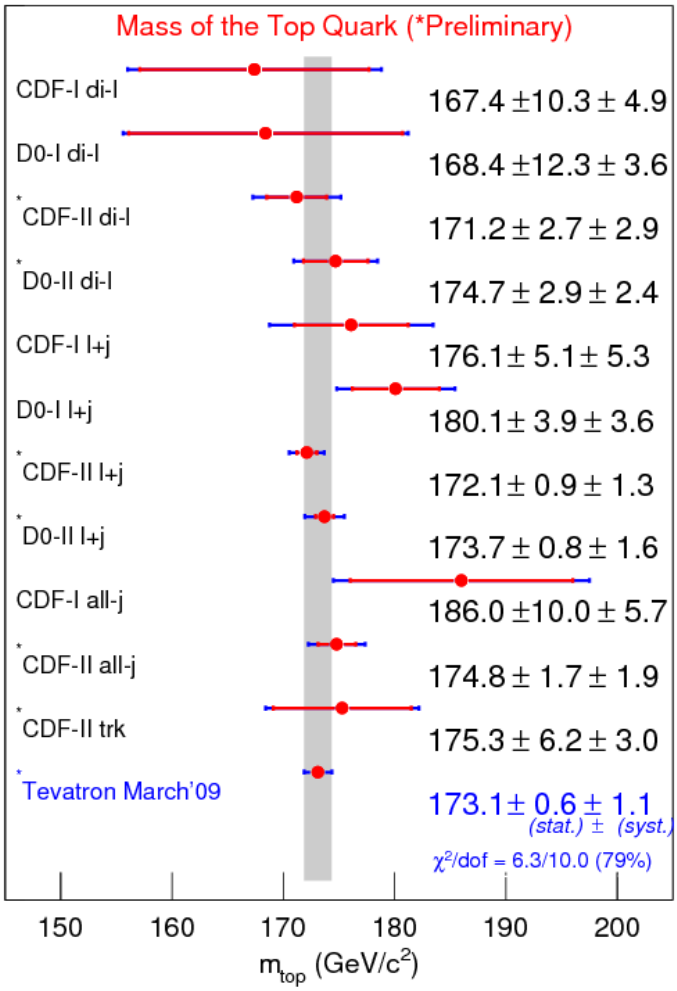


Figure 1.1: Tevatron measurements of the top mass quark. Taken from the CDF results website.

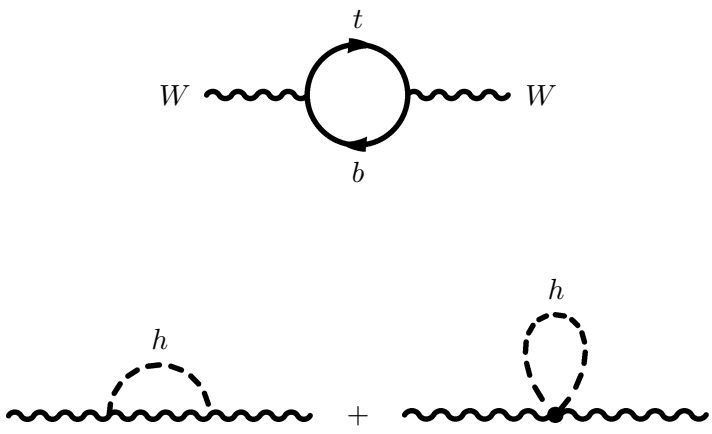


Figure 1.2: Loop diagrams contributing to the W-boson mass.

1.2 Electroweak measurements and the top quark

It was mentioned in section 1.1 that the top mass had been predicted correctly from precision electroweak measurements before it was actually discovered at the Tevatron. This deserves some further explanation.

At tree level, the mass of the W boson can be expressed as a function of the three most accurately measured electroweak quantities. The fine structure constant α, the Fermi constant, G_F, and the mass of the Z boson, M_Z.

$$M_W^2 = \frac{1}{2}M_Z^2(1 + \sqrt{1 - \frac{4\pi\alpha}{\sqrt{2}G_F M_Z^2}}) = \frac{\pi\alpha}{\sqrt{2}G_F s_W^2} \tag{1.1}$$

where we have defined

$$s_W^2 \equiv \sin^2\theta_{\mathrm{W}} \equiv 1 - \frac{\mathrm{M_W^2}}{\mathrm{M_Z^2}} \tag{1.2}$$

and $c_W^2 \equiv \cos^2\theta_{\mathrm{W}}, \mathrm{t_W^2} \equiv \tan^2\theta_{\mathrm{W}}$. However, one loop diagrams including the top quark and the Higgs boson (Fig. 1.2) also contribute to the W mass by introducing a factor $1 - \Delta R$ to the denominator

$$M_W^2 = \frac{1}{2}M_Z^2(1 + \sqrt{1 - \frac{4\pi\alpha}{\sqrt{2}G_F M_Z^2}}) = \frac{\pi\alpha}{\sqrt{2}G_F s_W^2(1 - \Delta R)}. \tag{1.3}$$

The contribution of the top one-loop diagrams to ΔR is

$$(\Delta R)_t \approx -\frac{3G_F m_t^2}{8\sqrt{2}\pi^2 t_W^2} \tag{1.4}$$

so the dependence of the W mass on the top mass is strong. The contribution of the Higgs diagrams

$$(\Delta R)_h \approx \frac{11G_F M_Z^2 c_W^2}{24\sqrt{2}\pi^2}\ln\frac{\mathrm{m_h^2}}{\mathrm{M_Z^2}} \tag{1.5}$$

is proportional to the logarithm of the Higgs mass, therefore the W mass is much less dependent on the the Higgs boson mass than on the top mass. This is the reason why the use of W mass measurements for estimations of the mass of the top has been possible - because the W mass had already been measured reasonably well before direct observation of top quarks was achieved. In turn, our current good knowledge of the top mass, can now be used to predict a mass for the Higgs boson from the mass of the W, since the top contribution to the W mass can now be accurately calculated (figure 1.3).

1.3 $t\bar{t}$ production at the LHC

The primary source of top quarks at the LHC and the signal for this analysis is the production of $t\bar{t}$ pairs by the strong interaction. The leading order Feynman digrams can be seen in figure 1.4. The first diagram corresponds to quark-antiquark annihilation ($q\bar{q} \rightarrow t\bar{t}$) and the other three to gluon fusion ($gg \rightarrow t\bar{t}$). The qq or gg pair that participate in the hard scattering process originate from the colliding protons.

In its simplest quantum representation the proton is a bound state of two u and one d-quark. In the context of a high energy collision however, the proton should be treated as a complex object, constantly subject to quantum fluctuations, wherein a plethora of partons (quarks and gluons) are emitted and recombine. Thus, appart from the three valence quarks (uud), there is a chance that the particle participating in the collision is one of the partons belonging to this quantum "sea". It is this "sea" that the antiquarks and the gluons necessary for $t\bar{t}$ production

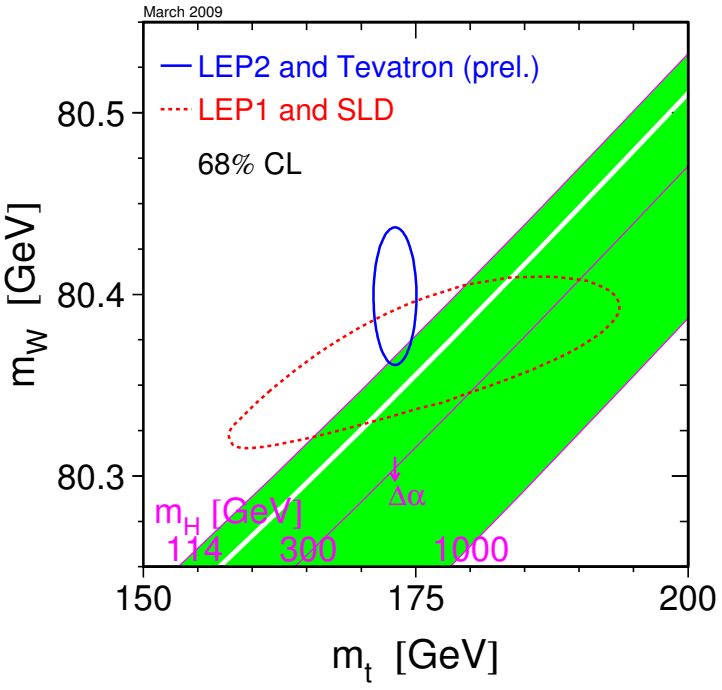


Figure 1.3: W-boson mass dependence on the mass of the top quark. The dashed (solid) ellipse corresponds to the 68% confidence limit that LEP1 (LEP2) placed on the two masses. The diagonal lines show the dependence under different constant Higgs masses. From http://lepewwg.web.cern.ch/LEPEWWG/.

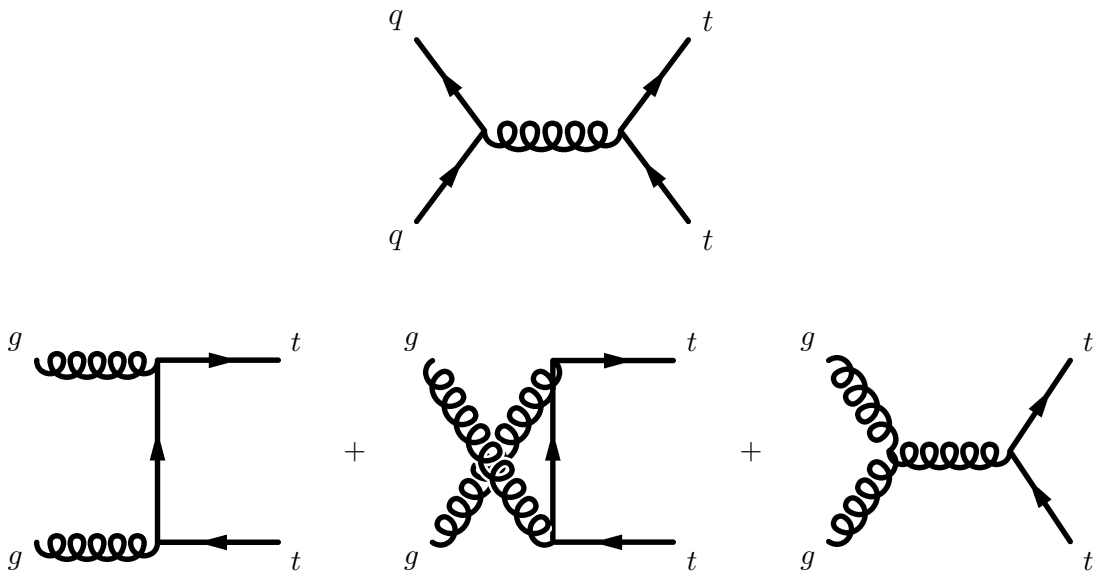


Figure 1.4: Leading order $t\bar{t}$ production Feynman diagrams. Quark-antiquark annihilation (above) is responsible for only 10% of the production at the LHC. The remaining 90% of $t\bar{t}$ are produced via gluon fusion (below).

come from. The probability density of a parton of type i, carrying a fraction x of the proton's momentum appearing is given by its Parton Distribution Function (PDF) $f_i(x, \mu_f)$. PDFs are currently impossible to calculate theoretically. Instead, parametrizations exist based on fits to experimental data. These methods to compute the PDFs rely on separating the hard scattering process from the rest of the low-momentum transfer processes by a factorization scale μ_f. The factorization scale reflects the momentum transfer of the process and is usually taken to be equal to the top mass for $t\bar{t}$ cross-section calculations.

Figure 1.5 shows example PDFs for different partons within a proton. This plot is very useful in understanding $t\bar{t}$ production. The first thing to notice is that as the momentum fraction goes down it becomes more likely for a parton to appear inside the proton. This means that a process with a threshold x_{min} will mostly occur for x values near its threshold. For $t\bar{t}$ production this means that the total energy in the center of momentum frame of the two partons will be very close to $2m_t$ or

$$\hat{s} \approx 4m_t^2 \tag{1.6}$$

But if P_1, P_2 are the momenta of the colliding protons, $\hat{s}^2$ can be written (neglecting the mass of the partons and of the protons)

$$\hat{s} = (x_1 P_1 + x_2 P)^2 = x_1^2 P_1^2 + x_2^2 P_2^2 + 2x_1 x_2 P_1 \cdot P_2 \approx 2x_1 x_2 P_1 \cdot P_2 = x_1 x_2 S \tag{1.7}$$

therefore

$$x_1 x_2 \approx \frac{\hat{s}}{S} \approx \frac{4m_t^2}{S} \tag{1.8}$$

We can calculate the threshold at the LHC by setting $x_1 = x_2 = x$ and $S = 14$ TeV. The result is $x = 0.025$. Looking back at figure 1.5, it is obvious that gluons dominate that x region. This means that the cross-section for gluon fusion will be much higher than for $q\bar{q}$ annihilation. Incidentally, for the Tevatron, $S = 1.96$ TeV results in $x = 0.179$ and the light quark distribution functions obtain larger values than that of the gluon. That coupled with the fact that Tevatron is a $p\bar{p}$ collider, which means that there will be no shortage of antiquark partons, explains why the situation is reversed for the Tevatron.

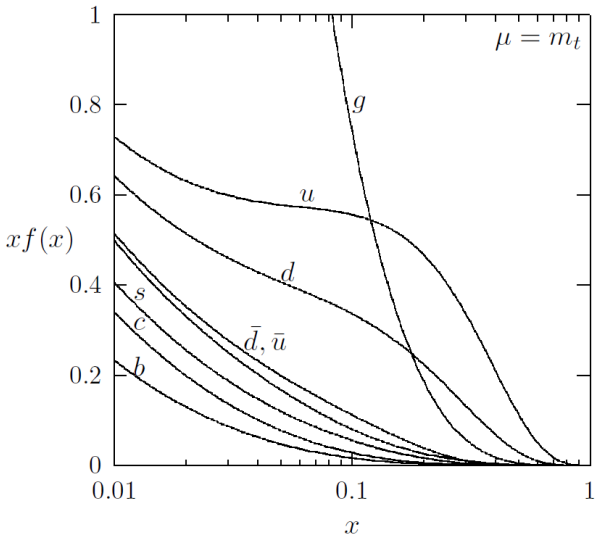


Figure 1.5: Example parton distribution functions for the proton. The product of the longitudinal momentum fraction x and the distribution functions f is plotted versus x. The factorization scale used is equal to the top mass. Taken from [13].

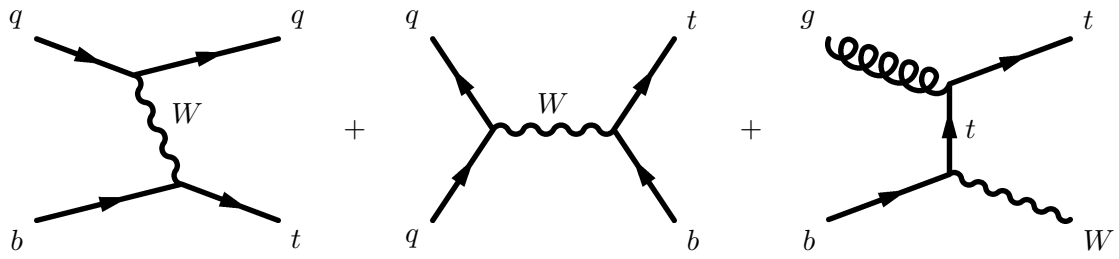


Figure 1.6: Feynman diagrams for single top production. The diagrams illustrate (from left to right) the t-channel (75%), s-channel (3%) and Wt associated production (22%).

Once the PDFs are given, the cross-section of $t\bar{t}$ production can be written

$$\sigma_{p\bar{p}\to t\bar{t}} = \sum_{i,j=q,\bar{q},g} \int dx_i dx_j f_i(x_i,\mu_f^2) f_j(x_j,\mu_f^2) \hat{\sigma}_{ij\to t\bar{t}}(\hat{s},\alpha_s(\mu_r)) \tag{1.9}$$

At the time of writing, the latest calculations at Next to Leading Order (NLO) plus Next to Leading Logarithm (NLL) [14] result in an LHC cross-section of 961^{+89+11}_{-91-12} pb, 90% of which comes from gluon fusion and 10% from quark-antiquark annihilation. The first uncertainty is the result of the absence of higher perturbative orders from the calculation and the second is the PDF-induced uncertainty.

1.4 Single top production at the LHC

The top quark can also be produced by the weak interaction. This is referred to as single top production, as the top quark is not produced in pairs. The leading order Feynman diagrams for the three subprocesses can be found in figure 1.6. The first involves a time-like W-boson as the mediator of the interaction (t-channel), in the second a space-like boson is exchanged between a light quark and a b-quark from the proton sea (s-channel) and in the third one the W-boson is real and comes as a by-product of the top production (Wt associated production).

There are a number of reasons why single top production should be studied at the LHC. For example, its cross-section is proportional to $|V_{tb}|^2$, allowing for a direct measurement of this element of the CKM matrix. V_{tb} is very well known only under the assumption of unitarity and from the values of V_{ts}, V_{td}. No direct measurements have otherwise been made. If a direct measurement showed that its value differs significantly from unity, it would be a strong indication of physics beyond the Standard Model.

Of equal if not greater importance is the role of single top as an important background for other signals. A notable example would be a Higgs boson search in $WH \to l\nu bb$ decays, the final state of which has the same objects as the t-channel single top decay.

1.5 $t\bar{t}$ decay

The top quark decays by the charged-current weak interaction. The dominant decay mode is $t \to Wb$. Transition to the two lighter generations ($t \to Ws, t \to Wd$) is highly suppressed, due to the very small V_{ts}, V_{td} elements of the CKM matrix. By the unitarity of the CKM matrix, $\sum_{i=d,s,b} |V_i| = 1$ and since $|V_{ts}| << 1, |V_{td}| << 1$, we know $|V_{tb}| \approx 1$ without directly measuring it [1], assuming there is no fourth generation of quarks. Flavour changing neutral current decays

[1] The exact current value according to [15] is $|V_{tb}| = 0.999133 \pm 0.000044$.

name	decay mode	branching ratio
di-leptonic	$t\bar{t} \to e\nu_e be\nu_e b$	1/81
	$t\bar{t} \to e\nu_e b\mu\nu_\mu b$	2/81
	$t\bar{t} \to \mu\nu_\mu b\mu\nu_\mu b$	1/81
di-leptonic τ channels	$t\bar{t} \to e\nu_e b\tau\nu_\tau b$	2/81
	$t\bar{t} \to \mu\nu_\mu b\tau\nu_\tau b$	2/81
	$t\bar{t} \to \tau\nu_\tau b\tau\nu_\tau b$	1/81
semi-leptonic	$t\bar{t} \to e\nu_e bq\bar{q}'b$	12/81
	$t\bar{t} \to \mu\nu_\mu bq\bar{q}'b$	12/81
semi-leptonic τ channels	$t\bar{t} \to \tau\nu_\tau bq\bar{q}'b$	12/81
hadronic	$t\bar{t} \to q\bar{q}'bq\bar{q}'b$	36/81

Table 1.1: $t\bar{t}$ decay modes and branching ratios

are forbidden in the Standard Model at tree level and loop calculations indicate a negligible branching ratio [16].

NLO calculations for the top quark decay rate yield

$$\Gamma(t \to Wb) \approx 1.42 \text{ GeV} \Rightarrow \tau = \frac{1}{\Gamma} \approx 4 \times 10^{-25} \text{ s} \tag{1.10}$$

The scale of the hadronization process is an order of magnitude larger than the top lifetime. This practically means that the top quark does not get the chance to form hadrons, but decays as a free quark.

A decay of the top quark is characterized as leptonic or hadronic based on how the W-boson decays subsequently. The W-boson can decay into a $q\bar{q}'$ pair of the first two generations and of any of the three colors or into a lepton and its neutrino of any of the three flavors. These nine outcomes are equally probable. When discussing $t\bar{t}$ pairs, it is customary to speak of di-leptonic decays when the W-bosons from both quarks decay leptonically ($BR(t\bar{t} \to l\nu bl\nu b) = 9/81$), semi-leptonic decays when one decays leptonically and the other to a $q\bar{q}'$ pair ($BR(t\bar{t} \to l\nu bq\bar{q}'b) = 36/81$) and hadronic decays when both decay to $q\bar{q}'$ pairs ($BR(t\bar{t} \to q\bar{q}'bq\bar{q}'b) = 36/81$). Due to the particular nature and detection requirements of the tau lepton, which can decay both leptonically and hadronically, we usually refer to the tau-including channels ($BR(t\bar{t} \to \tau\nu_\tau bl\nu b) = 5/81, BR(t\bar{t} \to \tau\nu_\tau bq\bar{q}'b) = 12/81$) separately. Table 1.1 summarizes the branching ratios of all different channels.

1.6 Summary

In this chapter, we have described how the prediction of the top quark was experimentally confirmed. The theory of the production and decay of the heaviest known elementary particle is in good agreement with the results of the Tevatron experiments. Nevertheless, the study of the top quark remains one of the high priorities for the two general-purpose LHC experiments (CMS and ATLAS) and will hold great importance during the first period of their operation, because

- The high production cross-section will provide large statistics enabling better determination of the top quark properties.
- Studying top quark events is an essential step towards the commissioning of the detectors.
- Top quark processes are an important source of background for Higgs boson searches and searches for new physics.

Chapter 2

The Large Hadron Collider

In this chapter we first summarize the basics of particle colliders. Having given a very short introduction on the relevant parameters we then focus on the Large Hadron Collider (LHC) [17] currently under construction at CERN, providing a general overview of its design and of the planned experiments.

2.1 Collider parameters

2.1.1 Center of mass energy

An important parameter for a collider is the center of mass energy of the colliding particles, E_{CM}. In order to maximize the energy that can be put in a physics process, thus reaching the threshold of more physics processes, colliding particles have to have opposite and equal momenta. It is for that reason that modern particle colliders aiming for the discovery of new resonances utilize accelerated beams of equal and opposite momenta crossing each other at one of more collision points. It should however be noted that when it comes to measuring the lifetime of such resonances it is useful to produce them with a boost, expanding their lifetime in the laboratory frame. Another important thing is the precise determination of the collision spot. This is the reason why the beams in B-factories are some times asymmetrical and do not collide head-on but at a certain angle.

In the case of a linear machine the limiting factor to the E_{CM} is the length. If the collider is circular then the bunches can be accelerated multiple times by the same electric fields, however the energy is limited by the synchrotron radiation. The energy emitted by a single particle per revolution is $\Delta E \propto \frac{1}{R}(\frac{E}{m})^4$. It is therefore understood that in order to reach higher energies we need to build larger colliders and that it is much easier to reach higher energies by accelerating heavier particles.

2.1.2 Instantaneous luminosity

The instantaneous luminosity L is defined as the number of particles crossing each other per unit area per unit time. In the case of two colliding beams each having a cross-section α and consisting of n bunches, the instantaneous luminosity is given by $L = nf\frac{N_1 N_2}{\alpha}$, where f is the revolution frequency and N_1, N_2 are the number of particles per bunch in the colliding beams. The usual unit for luminosity is $\mathrm{cm}^{-2}\mathrm{s}^{-1}$.

The luminosity connects the cross section of a certain physics process to the rate at which this process occurs. To obtain the rate of a given process of cross-section σ at a collider operating at a given luminosity we multiply $R = \sigma\ L$.

2.1.3 Total reaction rate

When we refer to the reaction rate R of a collider, we mean the total number of scattering events occurring at this collider per unit time. The reaction rate is therefore the product of the luminosity and the sum of the cross-sections of all scattering processes at the center of mass energy.

2.1.4 Integrated luminosity

When referring to the collider's production of collisions over large periods of time it is customary to quote the luminosity integrated over a year of operation and expressed in fb^{-1}/year. It is also customary to refer to any amount of data collected over a period of time by the corresponding integrated luminosity in fb^{-1}. This amount of data will then include a number of events of a certain process equal to the cross-section of this process times the integrated luminosity.

2.2 Choice of particles

Electrons and protons make excellent candidate particles for colliders because they are easy to produce, electrically charged therefore possible to accelerate and do not decay. Each of the two possibilities has its advantages and disadvantages.

2.2.1 e^+e^- colliders

Given that the e^+e^- interaction is well understood and described by the Standard Model, it is easy to manage the event rate in an electron collider. Besides that, electrons do not have an inner structure. This means that if we have particles of equal energy colliding, the laboratory frame is the CM frame and all the energy can be used to overcome the threshold of physics processes. Perhaps even more importantly, initial state gluon radiation (see 4.1.2) is absent from e^+e^- colliders[1] and the underlying event is much smaller in size. This results in a "clean" environment, where the signal can be isolated and triggered on with relative ease. Finally electron beams are easy to manipulate (e.g. polarize) enabling measurements related to chiral and other asymmetries.

One important disadvantage of e^+e^- colliders is that large energies are very difficult to achieve in circular accelerators due to the small mass of the electron leading to energy loss by synchrotron radiation. Linear colliders are still possible but would have to operate at lower luminosity. Another problem is that e^+e^- predominantly couple to a spin-1 state in the s-channel and resonant production of spin-0 states is all but excluded, whereas higher spin final states can only be reached through the interaction of higher order partial waves.

2.2.2 Hadron colliders

Compared to e^+e^- colliders, hadron colliders possess the important advantage of less synchrotron radiation resulting in higher center of mass energies and higher luminosities (due to the possibility of storing the beam in storage rings). In addition protons can interact through the strong force leading to larger rates for the same luminosities. Finally the interacting partons can be $q\bar{q}$, qg or gg pairs leading to different charge and spin final states and opening multiple possibilities of resonant production.

The main disadvantage is that, due to the scattering taking place among partons, the center of mass energy of the interaction is not the same as the energy of the partons in the laboratory frame (the total momentum might well have a longitudinal component) and each of the partons only carries an unknown fraction of the hadron's total energy. Also important is the production

[1] It should be noted that initial state photon radiation is present in e^+e^- colliders.

of a large number of particles in every event which result in a multitude of reconstructed objects (tracks, calorimeter hits etc). This makes the process of assigning reconstructed objects to products of the hard signal process much more difficult. Additionally, it creates the requirement for sophisticated trigger systems, as even events without physics interest will produce enough objects to survive the selection of a simple and/or loose trigger.

This concludes our brief discussion of particle colliders. For a more complete coverage of the topic the reader is referred to [18].

2.3 The Large Hadron Collider

The LHC is a particle accelerator and collider currently under construction at CERN, on the Franco-Swiss border near the city of Geneva. It is an international collaboration of over two thousand physicists and engineers coming from thirty-four countries. Scheduled to begin operation in May 2008, it will be the highest center of mass energy particle accelerator ever built and is expected to expand our physics knowledge to a new frontier, leading to new, exciting discoveries.

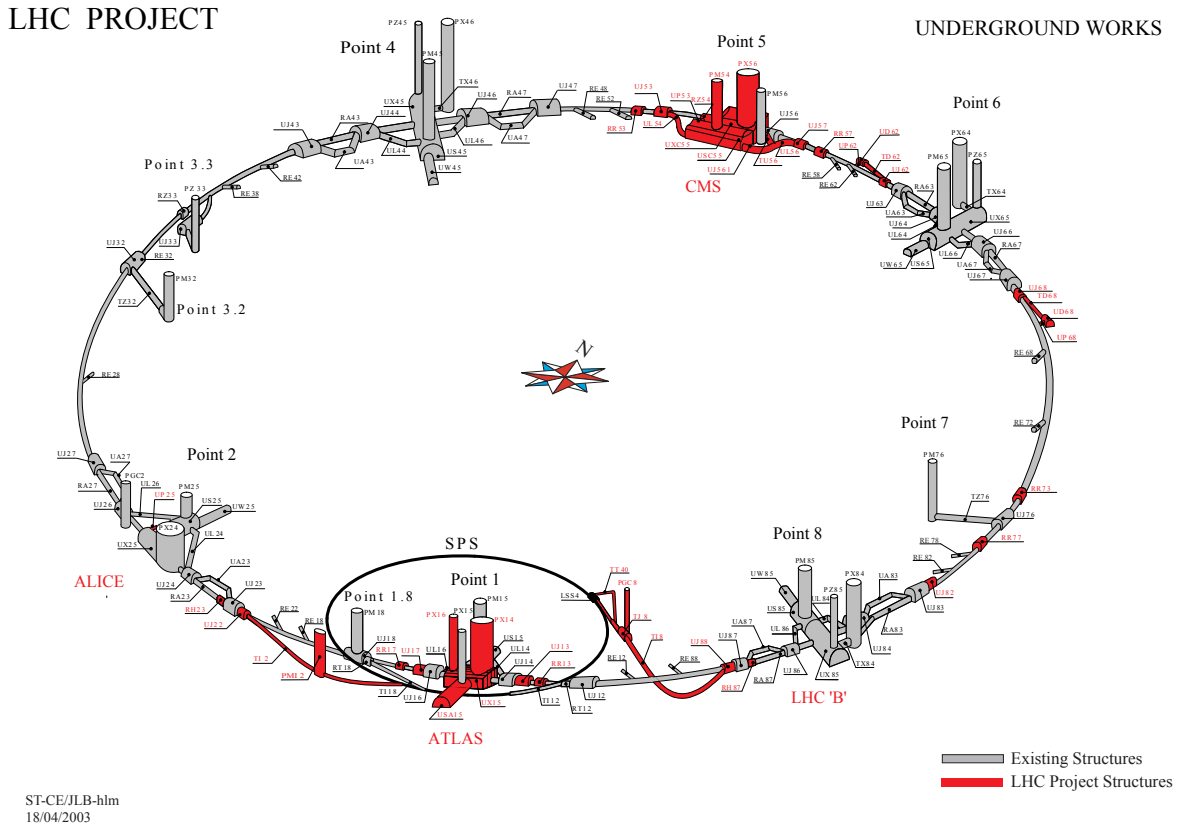


Figure 2.1: Schematic layout of the accelerator tunnel showing the previously existing LEP structures as well as the new ones built for LHC and the four experiments at the intersection points. Picture taken from the LHC webpage [17].

2.3.1 Technical design

A schematic view of the LHC can be seen in figure 2.1. The accelerator is contained in an underground tunnel of a circumference of 27 km at a depth ranging from 50 to 175 m. This tunnel was originally excavated in the 80's to contain the Large Electron-Positron Collider

Parameter	pp	HI	
Energy per nucleon	7	2.76	TeV
Design Luminosity	10^{34}	10^{27}	$\mathrm{cm}^{-2}\mathrm{s}^{-1}$
Bunch Separation	25	100	ns
No. of bunches	2808	592	
No. of particles per bunch	1.15×10^{11}	7×10^{7}	
No. of collisions per crossing	≈ 20	-	

Table 2.1: LHC machine parameters for pp and heavy ion (HI) collisions.

(LEP), which was shut down and decommissioned in 2000 after eleven years of operation. Along the tunnel run two pipes, each of which will contain a 7 TeV proton beam. Each of the beams will be driven and manipulated by hundreds of superconducting magnets positioned around the beam pipe and kept at low temperatures by helium cooling systems. There are four intersection points along the circumference of the LHC, where the opposite-headed beams cross each other. A cavern has been excavated at each of those points for the installation of the detectors.

To allow for accelerating and driving the beams, they are not continuous but come in separated bunches of particles. After an initial period of operation at a reduced luminosity, the number of bunches in each of the beams will rise to 2808 bringing the time interval between crossings at any given point at 25 ns and reaching the design luminosity of 10^{34} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$.

There is a series of accelerating systems that prepare the particle bunches before injection into the main accelerator. The first one is Linac2, a linear accelerator providing bunches of 50 MeV protons. These are fed into the Proton Synchrotron Booster (PSB), which increases their energy to 1.4 GeV, before passing them to the Proton Synchrotron (PS), which further accelerates them to 26 GeV. The last pre-injector the particles pass through is the Super Proton Synchrotron (SPS). When finally entering the LHC ring at the correct 25 ns spacing, the protons have reached an energy of 450 GeV.

Besides protons, the LHC is designed to create collisions of Pb ions as well. As a heavy ion collider it will operate at a luminosity of $10^{27}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The center of mass energy of the beams will be 5.5 TeV per nucleon. The machine parameters for the LHC are detailed in table 2.3.1.

2.3.2 LHC experiments

Six detectors have been approved for the LHC.

The two largest are CMS [19],[20] and ATLAS [21],[22]. These are general-purpose detectors, the construction of which has required the collaboration of thousands of people around the world. They are designed to take data at even the highest luminosity possible at the LHC and share very extensive physics programs, most notably including the search for the Higgs boson and Supersymmetric particles, but have different design characteristics. Since this thesis concerns an analysis to be carried out at CMS, the next chapter will cover this detector in some detail.

A Large Ion Collider Experiment (ALICE) [23], [24] is also being prepared with the purpose of studying heavy ion collisions. At center of mass energies of 5.5 TeV, collisions of Pb ions are expected to result in the generation of quark-gluon plasma, a very high-energy state of matter in which quarks and gluons are not confined within hadrons but exist in free form. ALICE hopes to observe and study quark-gluon plasma.

LHCb [25],[26] is an experiment designed for Beauty physics measurements. It will take data at a luminosity two orders of magnitude lower than the peak LHC luminosity and its main focus is on the measurement of the CP violation in the interactions of b-hadrons.

Finally, TOTEM [27], [28] and LHCf [29], [30] are specialized experiments for very specific purposes. They share interaction points 5 and 1 with CMS and ATLAS respectively. TOTEM stands for Total Cross Section, Elastic Scattering and Diffraction Dissociation. It is

an experiment dedicated to the measurement of total cross section at the LHC by a luminosity-independent method based on the simultaneous detection of inelastic or quasi elastic interactions. LHCf comprises two detectors positioned at 140 m on either side of the interaction point. Its purpose is to study the particles generated in the very "forward" region of collisions and draw conclusions related to the validity of existing models for cosmic ray showers.

Chapter 3

The CMS detector

This chapter is dedicated to the detection of particles produced at high energy collisions with the Compact Muon Solenoid (CMS) detector. We first examine how different kinds of particles behave inside a detector, what kind of experimental signature we could expect them to leave and how we can hope to identify them. From that follows the idea for the typical layered structure of modern high energy physics experimental devices. Most importantly, a short description of how CMS realizes this idea is provided.

3.1 Detection of particles

It is useful to classify particles produced in collisions based on how they behave in the detector and what kind of an experimental signature they leave behind, allowing their identification.

Very massive particles (t, W, Z etc.) as well as particles decaying through the strong or electromagnetic interaction will decay without traveling considerable distances in the detector. Such short-lived particles can only be detected through their decay products.

Weakly decaying particles will move a certain distance before decaying. Their decay vertex will therefore be displaced from the initial interaction point. Typically this secondary vertex is reconstructed from the tracks of its charged products.

Particles with lifetimes long enough to travel through the detector are detected by its different components. A typical detector layout can be seen in figure 3.1.

Electrons will leave hits in a tracking system as they follow curved trajectories through the magnetic field. Interaction with the detector material results in electromagnetic cascades of Bremsstrahlung photons and electrons from pair creations, which mainly take place in the electromagnetic (e/m) calorimeter. The typical signature of an electron is therefore a localized e/m calorimeter energy deposition combined with a matching track. However, particularly in cases in which the tracking system is made up by a lot of material, the Bremsstrahlung photons might be emitted early. This results in topologically separated energy depositions in the calorimeter and even multiple tracks per initial electron in cases where the Bremsstrahlung photons converted into electron-positron pairs before reaching the e/m calorimeter.

Photons also produce the same kind of electromagnetic showers in the e/m calorimeter as the electrons, however they do not follow curved trajectories, being neutral. As in the case of electrons, if a pair creation takes place within the tracking system then it may lead to tracks and separated clusters of energy deposited in the e/m calorimeter.

Stable hadrons like charged π's and protons do not lose much of their energy through Bremsstrahlung, however they will interact strongly with the nuclei of the detector material, ionizing them and producing new hadrons until their energy is absorbed. They thus result in tracks (if charged) as well as energy depositions in both the e/m and the hadronic calorimeters.

Muons, being heavier than electrons, radiate much less. Therefore they typically pass through both the e/m and the hadron calorimeter without depositing significant amounts of

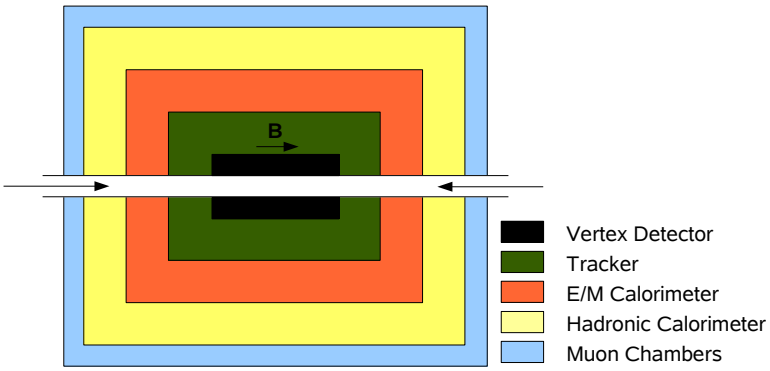


Figure 3.1: Typical layout of a general-purpose detector with a vertex detector and tracker at the center, electromagnetic and hadronic calorimeters and ionization chambers to track muons on the outside layer.

energy and subsequently escape the detector. For this reason, it is not possible to measure their energy by absorbing it, however, an external tracking system can be used to identify them as most other charged particles will not reach the area outside the calorimeters.

Neutrinos do not interact with detector matter and exit the detector without leaving any signal. They can therefore only be indirectly identified as "missing transverse energy". Since the colliding particles momenta should in principle not have a large transverse component, the vectorial sum of the transverse momenta of all detected products will in an ideal detector be zero. A non-zero sum can only be explained in the context of the Standard Model by assuming that one or more neutrinos were emitted. The total transverse momentum of those neutrinos will then in principle be equal to that sum and in the opposite direction.

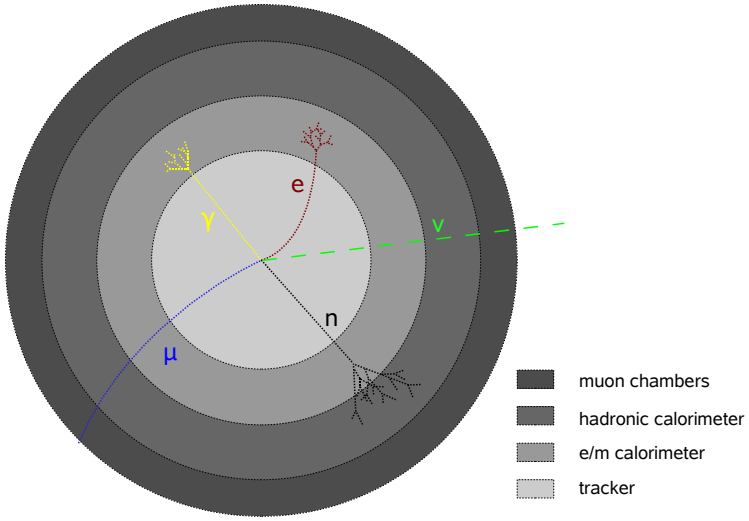


Figure 3.2: Different kinds of particles travelling through a typical general-purpose detector.

3.2 CMS

The variety of particles to identify and reconstruct give rise to the need to employ several different detection techniques in the form of different detector subsystems. Furthermore, those need to be combined in such a way that the existence of each one only minimally obstructs the operation of the others. The rest of this chapter is a description of CMS [20]. As a modern general-purpose detector, CMS is designed to fulfill all the requirements mentioned above.

3.2.1 Detector overview

A 13 m long, 5.9 m inner diameter superconducting solenoid is the central element of the CMS design. The magnet provides a uniform 4 T magnetic field with the purpose of bending the particle trajectories, enabling the precise measurement of their momenta from the curvature of their tracks. On the outside of the solenoid, the magnetic flux is returned via a 1.8 m-thick saturated iron yoke. The field inside the iron of the yoke is 1-2 T. At the center of the solenoid lies the cylindrical tracking system. The inner part consists of silicon pixel detectors positioned very close to the interaction region. The surrounding outer part is made up of ten layers of silicon microstrip detectors bringing the total tracking volume to 5.8 m in length and 2.6 m in diameter. A lead tungstate electromagnetic calorimeter (ECAL) covers a $|\eta| < 3.0$ pseudorapidity region around the tracker. Hadron showers passing through the ECAL crystals are detected in the hadronic sampling calorimeter, which comprises layers of brass and scintillator and fills the remaining space between the ECAL and the magnet solenoid. Additional scintillators are installed outside the solenoid enhancing the effective thickness of the HCAL barrel to more than 10 interaction lengths for $|\eta| < 1.2$. The outermost layers of the CMS detector are the muon systems, made up of drift tube chambers (barrel, $|\eta| < 1.2$), cathode strip chambers (endcaps, $|\eta| < 2.4$) and resistive plate chambers (barrel and endcaps, $|\eta| < 1.6$) and installed within the return yoke.

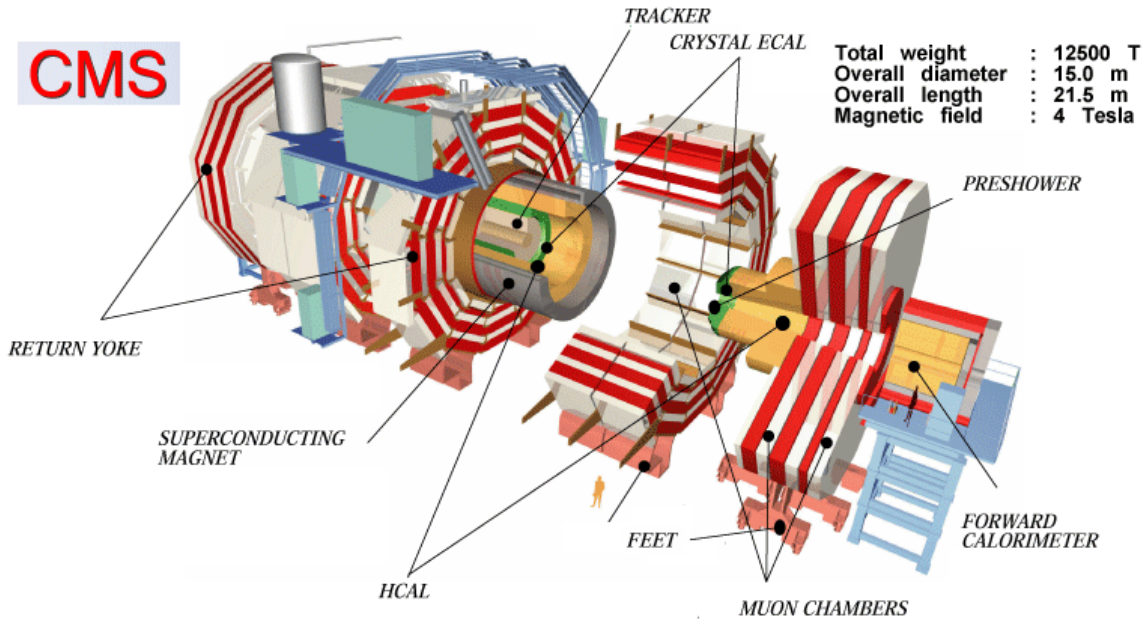


Figure 3.3: The CMS detector. Taken from the CMS website [31].

3.2.2 The inner tracker

The cylindrical tracking system of CMS extends from a radial distance of a few centimeters to almost 110 cm of the interaction point and is 540 cm long. It covers a pseudorapidity range of $|\eta| < 2.4$. The requirement to keep the occupancy low has dictated the size of the detectors used in each area. Thus, very close to the interaction point ($r < 20$ cm) the design calls for pixel detectors with small size (100×150 μ m^2) to compensate for the high particle flux, whereas further away, silicon microstip detectors with a minimum cell size of 10 cm$\times$80 μ m and a maximum cell size of 25 cm$\times$180 μ m are used.

The pixel detector

The pixel detector (Fig. 3.4) comprises three barrel layers and four endcap disks (two on each side). The barrel length is 53 cm and the layers are located at radii of 4.4, 7.3, 10.2 cm. The endcaps have an inner radius of 6 cm and an outer radius of 15 cm and are located at $z = \pm 34.5$ and $z = \pm 46.5$ cm. There are in total 768 pixel modules in the barrel, arranged in half ladders of 4 and another 672 in the endcaps, arranged in 7-module blades placed in a turbine like fashion.

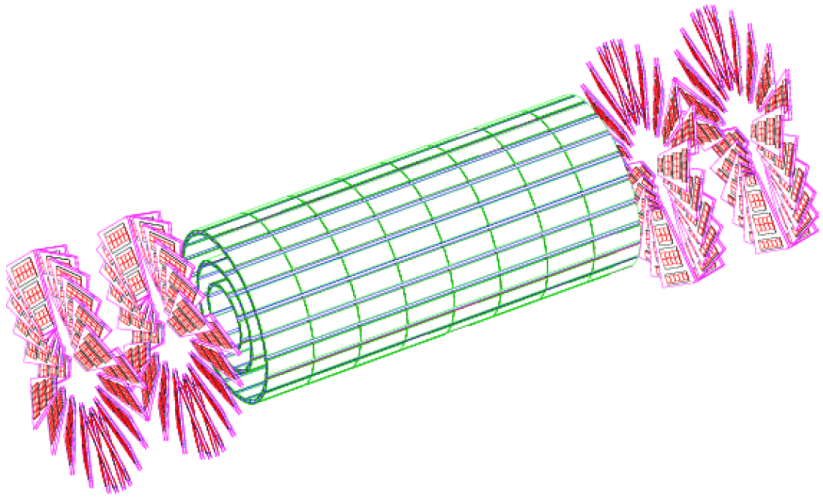


Figure 3.4: Drawing of the CMS pixel detector. Taken from [32].

The silicon strip tracker

The silicon strip tracker is divided into three regions. The Tracker Inner Barrel and Disks (TIB/TID) occupy a cylinder of 55 cm in radius and 130 cm in length. They form 4 barrel layers and 3 disks at each z-side. The detector modules are placed with their strips parallel to the beamline in TIB and radially in in TID offering a measurement of the ϕ coordinate. In order to get a two dimensional measurement for a hit, a second module can be mounted back to back with the first one with a stereo angle of 100 mrad. This gives a measurement of the r coordinate for the TID and of the z coordinate for the TIB, albeit at a much worse resolution. Such double modules make up the two innermost layers/rings of the TIB/TID respectively. Surrounding the TIB/TID and expanding to a radius of 116 cm there is the Tracker Outer Barrel (TOB). The TOB microstrips have thickness of 500 μ m and are positioned in 6 layers covering a $|z| < 118$ cm range. Similarly to the TIB, the two innermost layers of the TOB have double modules. Finally, the Tracker EndCaps (TEC) cover the 124 cm$< |z| <$ 282 cm regions. There are 9 ring-like disks in each endcap. The outer radius of each of the disks extends to the end of the tracker at $r = 113.5$ cm. The inner radii are bigger for the disks more distant from the interaction point,

as there is no tracker coverage for $|\eta| > 2.5$. Rings 1, 2 and 5 are made up of double modules. A drawing of the silicon strip tracker can be found in Fig. 3.5.

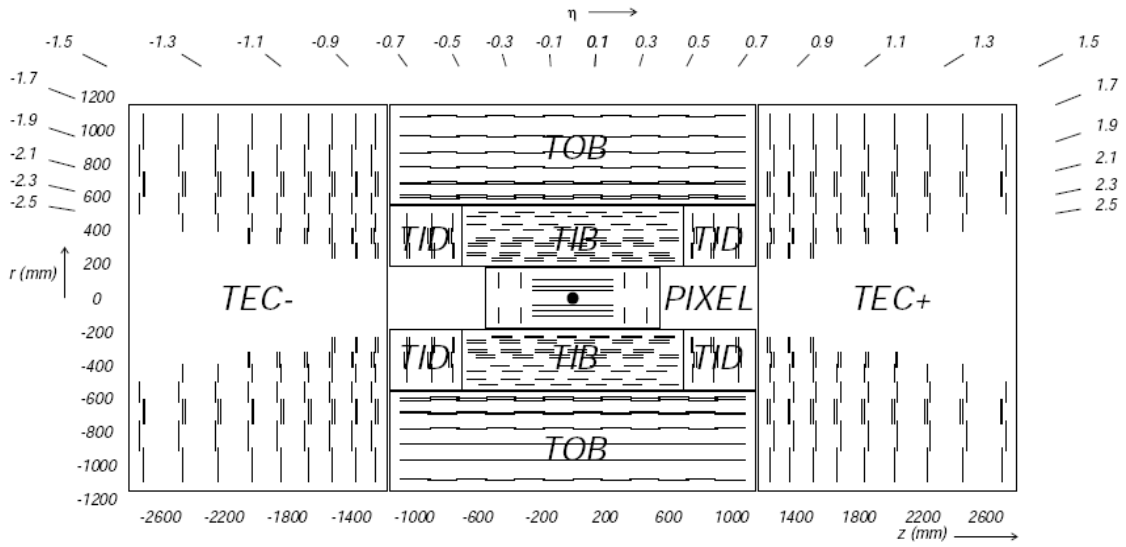


Figure 3.5: The CMS tracker. Taken from [20].

3.2.3 The electromagnetic calorimeter

The CMS ECAL is made of lead tungstate ($PbWO_4$) crystals. Lead tungstate was selected because of its short radiation length and Moliere radius and quick response time. These advantages come at the cost of a relatively low light yield (thirty photons per MeV). To compensate for that, photodetectors with large intrinsic gain are used.

The barrel crystals have a front face cross-section of 22×22 mm^2 and are 230 mm long. This is equivalent to $25.8X_0$. The crystals are positioned 129 cm from the beamline, slightly tilted, with their axes pointing 3^o off the nominal vertex position. Each of them covers a 0.0174×0.0174 area in $\Delta\phi$ and $\Delta\eta$. There are 61200 such crystals and they are organized in 36 supermodules, each covering half of the barrel length in a $20^o\phi-$region. Silicon Avalanche Photodiodes are used as photodetectors in the barrel.

The endcap crystals are slightly shorter and wider than those of the barrel, having a front face of 28.6×28.6 mm^2 and a length of 220 mm ($24.7X_o$). Each of the endcaps is placed 314 cm from the nominal vertex position with its crystals arranged in a x-y grid (instead of a $\eta - \phi$ grid in the barrel) with their axes similarly tilted. The crystals in each endcap are mounted on aluminum "Dees" and organized in 5×5 supercrystals. The pseudorapidity coverage offered by the ECAL endcaps is $1.479 < |\eta| < 3.0$. Vacuum phototriodes are used as photodetectors here.

In front of the crystal endcaps there is a preshower device employing two planes of silicon strip detectors behind lead absorber disks with a thickness of $2X_o$ and $3X_o$ respectively.

3.2.4 The hadronic calorimeter

The CMS HCAL is required to have good hermiticity and containment. Consequently it was important to maximize the interaction length of the HCAL on the inside of the magnet and also to complement this with an additional layer of scintillators lining the outside of the coil. Brass, being non-magnetic and having a short interaction length, makes a very suitable choice of absorber material. The tile/fibre technology was chosen for the active medium. Plastic

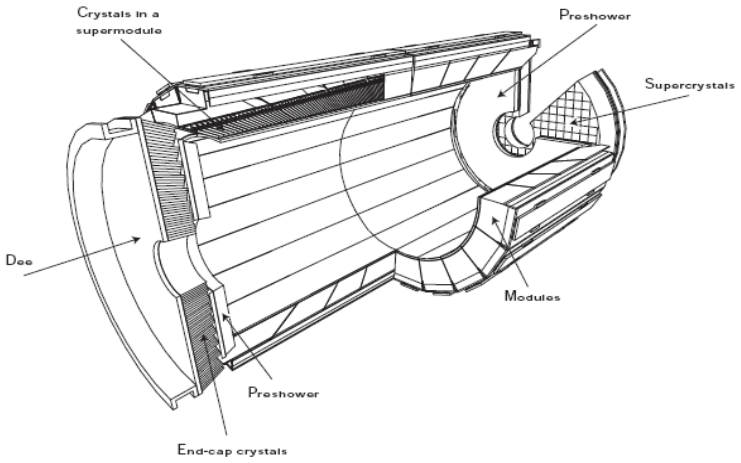


Figure 3.6: Drawing of the CMS ECAL showing the arrangement of crystals in the barrel and endcap. Taken from [20].

scintillator tiles are placed behind layers of absorber and read out with embedded wavelength-shifting fibres. The small thickness of the tiles (3.7 mm) allows for more space for the brass.

The HCAL barrel (HB) comprises 32 towers in the η direction and 72 towers in the ϕ direction to a total of 2304. It covers the pseudorapidity range of $|\eta| < 1.4$ at a segmentation of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$, matching that of the ECAL (1 HCAL tower for each 5×5 crystal cluster).

An additional layer of scintillators outside the magnet coil is referred to as the outer barrel (HO). The magnet acts as the absorber for the HO. The tiles are much thicker than those of the HB (10 mm). They are located inside the muon system and are organized in 5 ring-like sections 2.5 m wide in z. The purpose of the HO is to increase the effective thickness of the HCAL to 10 interaction lengths thus reducing the tails in jet energy resolution.

The HCAL endcaps (HE) cover the pseudorapidity range $1.3 < |\eta| < 3.0$. There are 14 towers in the eta direction with the segmentation varying from $0.087 < \Delta\eta < 0.350$ - larger values of $\Delta\eta$ correspond to towers closer to the beamline. The ϕ segmentation is 5^o for the 5 outermost and 10^o for the 8 innermost towers matching the segmentation of the muon chambers.

The forward HCAL (HF) is located 11.2 m from the interaction point and is used in the reconstruction of very forward jets. It combines 1.6 m of steel absorber with quartz fibres that transfer the emitted Cerenkov light to photomultipliers where it can be detected.

3.2.5 The muon system

There are three subsystems to the CMS muon system. First the barrel drift tube (DT) chambers, then the endcap cathode strip chambers (CSC) and finally the resistive plate chambers (RPC) extending over the whole barrel and part of the endcap.

Drift tube chambers

The barrel region is covered up to $|\eta| < 1.2$ by DT chambers organized in four coaxial cylindrical layers and placed within the magnet return yoke. The four layers are positioned in approximately 1 m radial distance from each other. Each layer is separated into 5 rings in the z-direction and each of the rings consists of 12 30^o-sectors with an aluminium drift tube chamber in each of

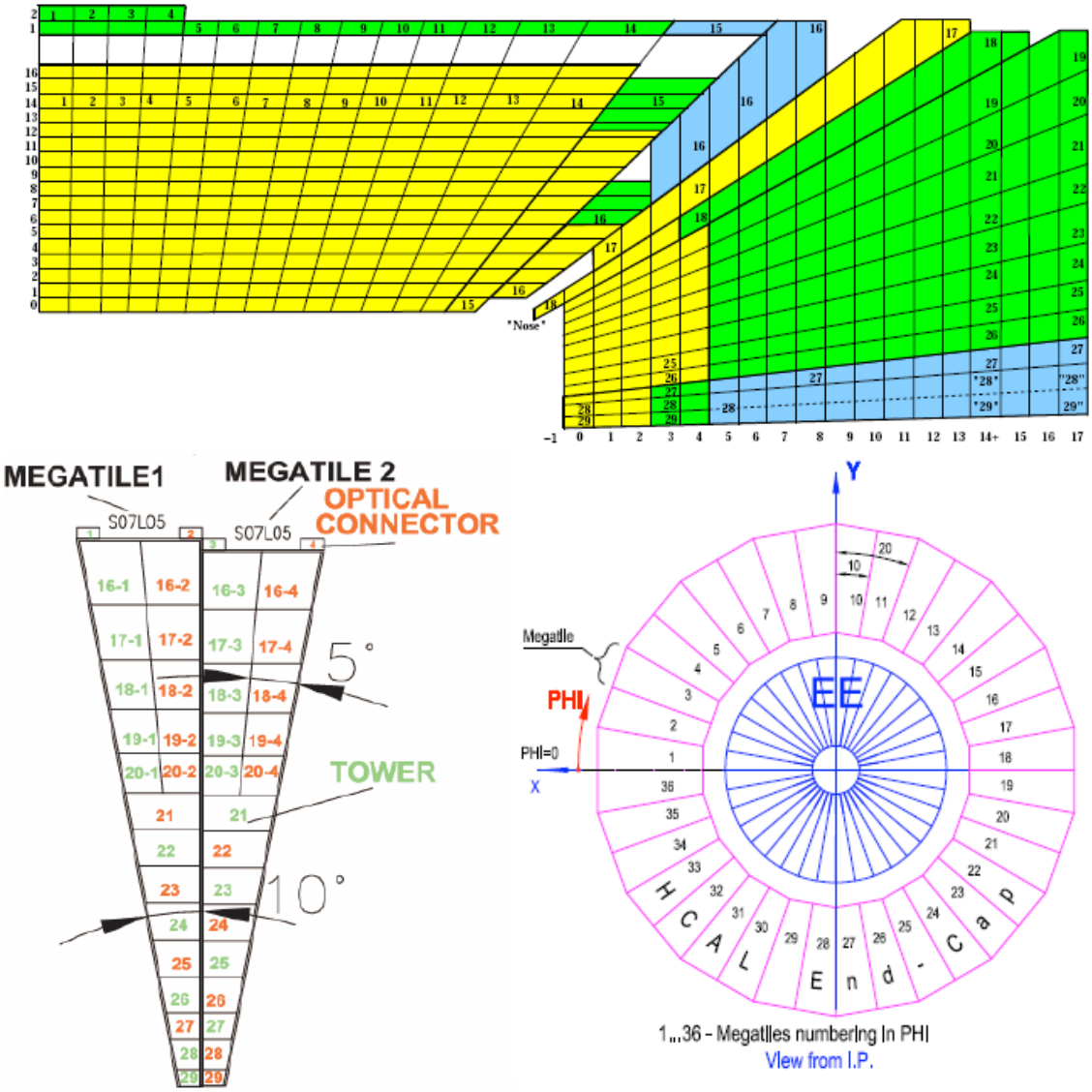


Figure 3.7: Drawings of the HCAL showing the tower segmentation in the barrel and endcaps. Taken from [32] and [20].

them. Two of the sectors of each of the fourth layer rings have two DT chambers instead of one, bringing the total number of chambers to 250.

Every DT chamber in layers 1, 2 and 3 consists of a superlayer of 4 planes of aluminium drift tubes measuring the z-direction, sandwiched between two superlayers of $r - \phi$ measuring planes. Thus, from each chamber we can obtain 12 measurements of the track, resulting in a vector with a precision of less than 100 μ m in position and 1 mrad in direction. The layer-4 chambers do not include z-measuring planes therefore the measured points for a single track crossing all four layers is $36 + 8 = 44$.

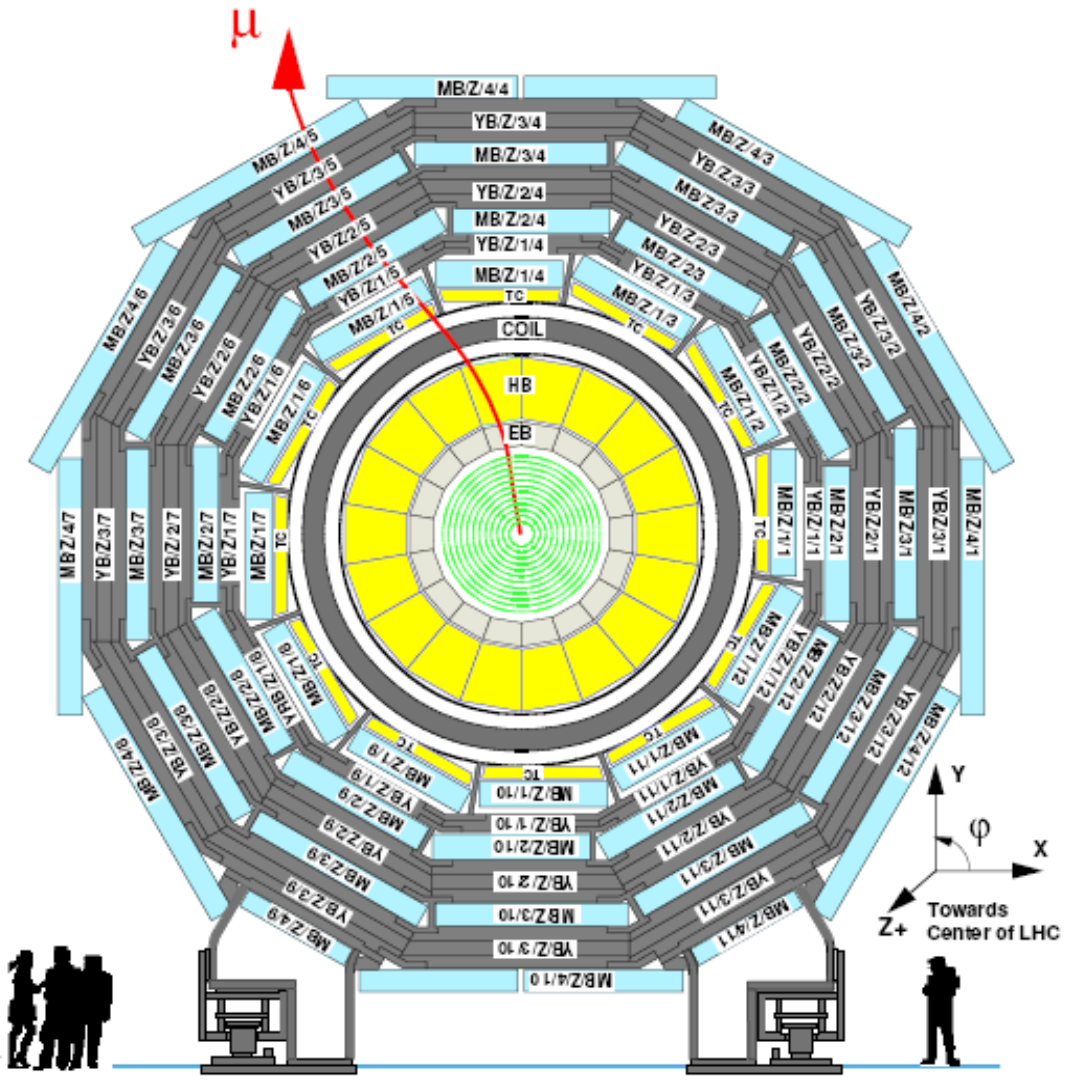


Figure 3.8: The CMS barrel drift tubes installed within the iron return yoke. Taken from [32].

Cathode strip chambers

The endcap muon system consists of 4 disks of CSCs in each endcap. The trapezoid-shaped CSCs are arranged in rings on the disks overlapping so as to avoid gaps in the acceptance. The innermost rings of disks 2, 3 and 4 only have 18 chambers but the rest of the rings including

the innermost of disk 1 have 36 adding up to a total of 468 CSCs. Each chamber comprises 6 gaps. In each of the gaps there is a plane of radial cathode strips and a plane of anode wires positioned perpendicularly to the strips. A muon traversing the gap ionizes the gas and the electrons are collected on one anode wire. The positive charge on the other hand is shared by more than one cathode strips. The wire provides a fast signal that can be used for triggering but the desired high resolution (200 μ m spatial, 10 mrad angular in the ϕ-direction) comes from the charge-weighted mean position of the strips.

Resistive plate chambers

6000 m^2 of RPCs extend both over the barrel and endcap regions up to a pseudorapidity of 2.1. The RPC detectors comprise parallel plates with intermediate 2 mm gaps filled with gas. The main advantage of these detectors is that when operated in avalanche mode (that is, with a low electric field across the gap), they obtain a high rate capability. That, combined with their relatively low cost per unit area, makes them an appropriate choice for CMS to be used for triggering on the muons and accurately determine the corresponding bunch crossing.

3.2.6 The trigger system

As explained in chapter 2, the time interval between two successive bunch crossings at the LHC is 25 ns. Given that the total size of the digitized, zero-suppressed detector signal per bunch crossing is O(1 MB), data at CMS is produced at a rate of tens of GB/s. This data is impossible to transfer and store both because of its sheer volume and of the rate at which it comes. The rate at which data can be reliably recorded by the storage systems available to CMS is of the order of magnitude of 100 MB/s. This translates to a maximum event rate of O(100 Hz) and a required rejection power of O(10^5). Under such a demanding environment, a trigger system must fulfill and balance two basic requirements: a very high processing rate and a complexity high enough to make the selection of events efficient and flexible. The first requirement is fulfilled by the low-level, hardware-based part (Level-1) of the trigger and the second by the higher-level, software-based part.

Level-1 trigger

The Level-1 Trigger relies on hardware components like custom electronic boards and very low-level firmware algorithms to bring the event rate from 40 MHz down to less than 100 kHz. It comprises three parts: the calorimeter trigger, the muon trigger and the global trigger. The time available to the Level-1 Trigger to make the decision of accepting or rejecting an event is approximately 3.2 μs. Signals from the inner tracker, which would require complex and time-consuming processing and large data transfer are thus not used by the Level-1 Trigger.

The calorimeter trigger is based on the trigger tower energy sums computed by the Trigger Primitive (TP) Generator circuits of the three CMS calorimeters. The size of each trigger tower is $0.35\eta \times 0.35\phi$ (in the central region). Trigger towers are grouped into 16-tower (4×4) regions. The TPs are transferred to the Regional Calorimeter Trigger (RCT) which combines them to form candidates for electrons/photons, taus and jets for each of the regions. Isolated and non-isolated electron/photon candidates are seperately constructed. The candidates from each of the regions together with the sum of transverse energy from that region are then forwarded to the Global Calorimeter Trigger (GCT), which sorts them and transmits the leading four together with the total transverse and missing energy to the global trigger. A "quiet" bit and a bit indicating the possible observation of a minimum ionizing particle for each of the regions is forwarded to the global muon trigger, which uses them to determine muon candidate isolation.

Each of the muon subsystems has its own trigger logic. The RPCs have the Pattern Comparator Trigger (PAC), which is based on the spatial and time coincidence of hits in different

muon stations. The spatial pattern of the coinciding hits is compared to a set of pre-defined patterns corresponding to muons of different transverse momenta. A different set of candidate patterns is defined for different muon directions. In this way, the PAC can form candidate tracks of specific transverse momenta and pass a momentum code for each to the global trigger. Contrary to the RPC trigger, the DT and CSC triggers include local, per-chamber processing, which results in a momentum vector (track segment) for each chamber. Vectors from different layers are then combined into tracks of specific transverse momenta. Up to four tracks from each of the two subsystems are selected and ranked based on quality and transverse momentum and sent to the global muon trigger. The Global Muon Trigger (GMT) compares the tracks from the DTs and the CSCs to the tracks from the RPCs and uses any compatible tracks to form a single track of higher quality. The four best tracks at the end of this process are first characterized as isolated or minimum ionizing particles based on the bits passed by the calorimeter trigger for their corresponding calorimeter regions and then forwarded to the Global Trigger (GT).

Based on the information from the calorimeter and the muon triggers, the GT makes the Level-1 trigger decision to accept or reject the bunch crossing and sends it to the Data Acquisition system to initiate a readout of the event if necessary. The GT can apply a range of different criteria, based on the number of Level-1 objects above a configurable energy threshold and/or their absolute and relative positions.

Data acquisition and high level trigger

Once the Level-1 trigger issues an acceptance decision, the Data Acquisition (DAQ) system of the experiment reads out the detector signals through a complex set of readout buffers and switching networks, and builds the "event" for the accepted bunch crossing. This "event" is a software object that can be passed on to the High Level Trigger (HLT). The purpose of the HLT is to further reduce the rate of the events coming out of the Level-1 trigger by O(1000). Processing an event at the HLT takes 40 ms on average. The material part of the HLT is a computer farm. Based on the digitized detector signals included in the event received from the DAQ, the high-level algorithms running on this farm perform the complex task of reconstructing high-level physics objects, like muons, electrons or jets.

The reconstruction algorithms run by the filter farm to create those objects are a subset of the algorithms included in the reconstruction software of the experiment. The details of these algorithms are outside the scope of this chapter. For a very general description of the reconstruction of the physics objects relevant to this analysis, the reader can refer to Chapter 5 of this thesis. A complete account of HLT reconstruction at CMS can be found in the DAQ and HLT Technical Design Report [33] and the Physics Technical Design Report [34]. It is enough to state here that providing algorithms for the HLT farm is a very demanding task from a programming perspective, since the need to reconstruct the physics objects with a high efficiency and resolution usually collides with the need for quick decision making in an online environment.

With the physics objects reconstructed, HLT decisions can be taken based on the trigger table of the experiment. The trigger table is a list of conditions placed on the reconstructed objects of the event. An event is tested for every one of the conditions in the trigger table and marked according to whether it fulfills it or not. This helps subsequent analyses to limit the samples they use by only considering events passing one or more specific triggers. On fulfillment of any of the conditions, the event passes the HLT. It should be noted that the HLT table is the outcome of very detailed study and negotiation between the physics analysis and data acquisition groups. Since it can make the difference between making a new discovery or not, the design of the trigger table is very crucial to the experiment and is certain to change throughout the lifetime of the experiment according to shifting goals, available bandwidth and the luminosity delivered by the accelerator.

Once an HLT decision is made, it is the responsibility of the DAQ system to transfer the event to the storage elements of the experiment, where it becomes available to reconstruction and analysis software.

Chapter 4

Monte Carlo simulation

This work describes a potential analysis to be performed using data from the CMS detector. Since such data does not exist yet, we rely on software tools to simulate it. Based on these tools we attempt to predict the characteristics of the physics (signal and background) events involved and estimate the potential for measuring the quantities of interest at CMS. It is therefore important to discuss a few basic issues concerning Monte Carlo (MC) simulation in high energy physics, so as to understand the merits, and the limitations of this approach.

The current chapter includes two main sections. The first concerns the simulation of the actual physics processes at the generator level. A brief description of the two event generators used, Alpgen [35] and Pythia [36] and a justification for this choice is provided. The second section concerns the simulation of the interaction between the final state products of the physics process and the detector and the production of the detector signals.

4.1 MC generators

An Event Generator is a piece of software that generates simulations of particle physics events. In the case of simulating LHC collisions the basic input to an event generator is an initial state of two protons of opposite momenta of 7 TeV/c and a selection of processes to be simulated. The generator calculates internally the available phase space for each subprocess and the cross-section distribution over this phase space and returns

- the calculated total cross-section,
- the requested number of events, each including a final state (a number of particles with their momenta) along with a history of intermediate particle creations and decays that led to this final state.

The inherent randomness of physical processes is simulated in generators by means of Monte Carlo techniques. The basic principle is that if a process can be parametrized and we know the normalized probability density function (PDF), $f(x)$ for a parameter x in the range $[x_1, x_2]$, we can randomly determine a value for this parameter by generating a random number $R \in [0, \int_{x_1}^{x_2} f(x)dx]$ and solving $R = \int_{x_1}^{x} f(x')dx'$ for x.

The basic part of event generation is the simulation of the "hard process", the main physics process under examination. This is a manageable task, as all single high-energy processes have a well understood structure. The complexity however stems from the importance of simulating the effects surrounding the hard process, which play a very important role in shaping the final state and determining the observability of the event. The most important of those are highlighted below.

4.1.1 Parton distributions

In the simulation of pp collisions we need to take into account that the protons are composite objects made of quarks and gluons. The "hard process" of the event typically takes place between two partons, one from each of the protons. As explained in chapter 1, in order to calculate the effective cross-section of the process we therefore need a function $f_i(x)$ giving us the probability density for finding a parton of kind i carrying a ratio x of the proton's energy. Different choices of PDFs yield slightly different results, this however only introduces a small uncertainty compared to other uncertainties related to the event generation (see below).

4.1.2 Initial and final-state radiation

The partons in the initial state of a process, as well as those in the final state can radiate gluons (when colored) and photons (when charged). Especially in high-energy events this kind of initial or final state radiation can alter the event structure very significantly. The radiated particles can be quite energetic and in turn can branch like their mother particles. In this way, cascades of partons appear. This kind of effects not only changes the experimental signature of the event by adding more final state objects but also the calculated cross-section, by carrying away a portion of the energy that would otherwise be available to the hard process. There are two approaches when it comes to simulating radiative effects.

a) Matrix-element calculation: The straightforward approach involves calculating the Feynman diagrams of the hard radiative processes order by order, as is done for the hard process itself. This is beneficial in that it produces results which are kinematically exact, also taking into account interference effects and the helicity of the involved partons. As a consequence, the number of spatially well-separated particle cascades (jets) and their angular distributions are very well simulated by this method. There are however two drawbacks to the matrix-element treatment of radiative effects. The first is that the complexity and sheer volume of the calculations increases dramatically with the order of the diagrams and soon becomes practically impossible to perform. The second is that the perturbative expansion cannot provide a satisfactory description of multiple soft gluon emission, a very sizeable effect at high energies and one that greatly affects the internal structure of jets.

b) Parton showers: In this approach, each parton in the event branches into two an arbitrary number of times. The martix elements are not calculated. Instead, an approximation is made using simplified kinematics. Thus, each initial parton evolves into a shower of partons, characterized by a virtuality scale Q^2, which decreases with subsequent branchings until it reaches a minimum value at which the parton is no longer allowed to decay. This method has proven successful in simulating the emission of soft radiation at small angles from the direction of the mother particle. It is therefore suitable for describing the internal structure of jets. Furthermore, if tuned to a first-order matrix-element generator, a parton shower generator can also be relied on to describe jet multiplicities.

4.1.3 Fragmentation and the Lund model

QCD perturbation theory can only be applied at short distances. Color carrying quarks and gluons are not to be found free, but always exist within colourless hadrons. What is initially simulated as a shower of partons therefore always evolves into a shower of sometimes decaying hadrons via a process named hadronization or fragmentation and the generator software has to be able to describe this process in a reliable manner. However, there is no theory providing a satisfactory description of fragmentation. The only way to address this issue is to rely on a phenomenological model. Perhaps the most frequently used fragmentation model is called the "Lund model", the details of which are outside the scope of this chapter. The central idea however is that there is a colour flux tube connecting a quark and an antiquark forming a color

singlet. As the q and $\bar{q}$ move away from each other the colour tube behaves like a relativistic string and the partons experience an attractive potential increasing linearly with distance. Eventually, when the potential energy stored in the string increases enough, another $q'\bar{q}'$ pair is formed between the initial quarks and the system splits into two new colour singlets. The process repeats for each of the produced singlets until we arrive to final products that correspond to on-mass-shell mesons. An analogous phenomenological description exists for baryons. The final products of this process are frequently unstable, therefore a generator needs to use the experimental information available so as to properly simulate their subsequent decays into stable particles. It should be noted here that in the context of event generation "stable" refers to particles with lifetime long enough to reach the detector layers. Many of these particles however will decay while travelling through the detector. These "in-flight" decays are handled by the detector simulation software.

4.1.4 Choosing the generator software

The generator-level hard process is the production of $t\bar{t}$ pairs and their subsequent semileptonic decay. Important backgrounds to single-leptonic $t\bar{t}$ decays are the other decays of the $t\bar{t}$ system. More specifically, fully leptonic decays can easily fake single-leptonic decays given the existence of additional jets from initial- or final-state radiation. Additionally, fully hadronic decays can also be taken for single-leptonic decays provided one of the final-state jets fakes a lepton or an isolated lepton occurs from some other source (e.g. b- or c-quark decays). Another significant background to the signal process are W+jets events. The leptonically decaying W can easily be mistaken for the W of the $t \to bW$ decay. Finally, even non-W+jets QCD processes with fake leptons can contribute significantly to the background due to their very high cross-section. The requirement of this analysis regarding generator software is the ability to generate sufficiently big data samples of the above mentioned signal and background processes.

Furthermore, since one of the t-quarks of the signal process decays hadronically, any selection strategy targeting the signal process will unavoidably result in selecting events of high jet multiplicities in the final state. Another requirement from the generator software is therefore to be able to simulate reliably the number and angular distributions of the jets in background and signal events. Additionally, it is always necessary to achieve a good simulation of the fragmentation process and the internal structure of the jets.

a) Pythia

Pythia is one of the most widely used general-purpose event generators. It has a long history of development, which has resulted in a rich selection of available subprocesses. The main reason for the great usefulness of Pythia however, is its capability of addressing all of the issues of event generation highlighted in the previous sections. It includes mechanisms for simulating

- the parton distribution of the colliding hadrons
- the initial- and final-state radiation by the parton shower approach
- the hadronization/fragmentation process by an implementation of the Lund model
- the decays of unstable partons
- the interactions among the beam remnants after the hard process, often referred to as the underlying event.

Inspite of these advantages, Pythia is not an optimal solution for this analysis. The reason is that being a parton shower generator, it is inferior to a matrix element generator in terms of accurately predicting the number of final state jets and their angular correlations.

b) Alpgen

Alpgen is a generator designed for the simulation of Standard Model processes with a fixed number of partons in the final state, resulting from high energy hadron collisions. The final states are produced by exact leading-order calculations of the partonic matrix elements. The description of high jet-multiplicity events with high p_T jets obtained from Alpgen is thus much more precise than that obtained by parton shower generators, as explained in section 4.1.2. This makes it an ideal choice of a generator for the purposes of this analysis.

Although Alpgen produces the final state partons, it does not include the capability of simulating the parton showers resulting from them. To compensate for this, the developers have included an interface to Pythia in the code. This interface can be used to input the Alpgen-produced final state into the Pythia hadronization/fragmentation routines. This process of using a combination of Alpgen and Pythia to generate events is not trivial to carry out for reasons explained in the following section.

c) Combining Pythia and Alpgen - Jet-parton matching

The single-event output of Alpgen is a final state configuration including a number of partons from hard radiation emission. This corresponds to a process of a fixed number of final state jets, the cross-section of which is determined by the program and provided to the user. The danger of using Pythia to perform the hadronization/fragmentation on this output is that Pythia will create additional partons through the parton showering, which might be hard enough and emitted at angles large enough to add new jets to the final-state configuration. The obvious problem is that a sample meant to include Njets events, weighted according to the Njets calculated cross-section, starts including $N + x$ jet events already accounted for in the larger multiplicity samples.

The method used for handling this issue and ensuring that Njets samples produced by the Alpgen+Pythia combination remain exclusive is referred to as jet-parton matching. It is a configurable process, the results of which depend on a number of parameters, but the main principle is simple. If the results of the parton showering are used as input to a jet-reconstructing algorithm, then each of the reconstructed jets has to be matched to one of the partons of the Alpgen final state, otherwise the event is rejected. The matching is performed based on the "angle" $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ between jet and parton.

4.1.5 Generator-level samples

This analysis targets the first few (≈10) pb^{-1} of data to be collected by the detector. However, in order to have enough statistics to evaluate the performance of the event selection and analysis, it was necessary to use data samples corresponding to up to ten times that amount of integrated luminosity. Apart from the $t\bar{t}$ signal events, a number of important sources of background had to be taken into account.

1. $t\bar{t}$: The Alpgen+Pythia solution described above was used by the CMS experiment to produce a large amount of $t\bar{t}$-pair events separated in subsamples according to the number of jets in the event. The breakdown of the $t\bar{t}$ production cross-section across different jet multiplicity bins is shown in Tab. 4.1[1]. An appropriate number of events was taken from each sample, so as to collect a total dataset corresponding to $100\,\text{pb}^{-1}$ of integrated luminosity.

[1]This work targets a rediscovery of the top quark and a measurement of the production cross-section of $t\bar{t}$ pairs at a center of mass energy of 14 TeV, consistent with the original schedule of the LHC. It was recently decided by the CERN management that during the very first period of data taking the LHC will operate at a center of mass energy of 10 TeV. This corresponds to $t\bar{t}$ production reduced by approximately 55% (according to NLO+NLL calculations).

2. W+jets: W+jets production is by far the most important background process for our μ+jets final state. It is a relatively high cross-section process which gets a muon and a neutrino from the decay of the W boson and the jets from radiation. Large samples have been produced by CMS using Alpgen. When selecting our $100\,\mathrm{pb}^{-1}$ dataset out of the sample produced by the experiment, we took advantage of the fact that it was divided into jet multiplicity bins by only selecting events from the 2-jet and higher bins. This allowed us to keep our dataset within a manageable size and was only possible because of the high jet multiplicity in signal events. This means that we can impose a cut on the number of jets in our selection and be confident that the vast majority (if not all) W+jets events of the 0 and 1-jet bins would be cut away anyway.

3. $pp \to \mu X$: QCD multijet background can significantly contribute to a selection targeting single-muonic $t\bar{t}$ decays. The muon required by the typical selection can appear either as a decay product of heavy quarks or as a product of in-flight decays of pions. This background is hard to simulate accurately. The cross-sections for QCD processes are very high and the events selected would only be a very small subset of the total events we would need to produce. It would be very difficult to ensure sufficient CPU time to simulate the large number of events needed to obtain enough statistics of muon-including events. Furthermore, in-flight decays are not simulated at the generator level, but in the later (and much more time-consuming) stage of detector simulation.

   Nevertheless, to ignore this background would be an important mistake, as its contribution can be quite large if care is not taken to reduce it. For this reason, a Pythia-produced sample of muon-including QCD events corresponding to $8.7\,\mathrm{pb}^{-1}$ of integrated luminosity was used in order to at least obtain a rough estimate of how numerous this background will be. Only $\hat{\mathrm{p}}_T > 15\,\mathrm{GeV}$ events with a muon of $p_\mathrm{T} > 10\,\mathrm{GeV}$ are kept. It should be kept in mind that this sample does not give us the complete QCD background as it does not include muons from in-flight decays. However, in-flight decays are only expected to form a small part of the QCD background and should not play a big part in this analysis after muon isolation and p_T cuts.

4. Single t: Single top is not expected to have a major contribution to our selected sample, because its cross-section is not as high as that of the signal ($\frac{\sigma_{t\bar{t}}}{\sigma_t} \approx 3$) and the average number of jets in single top events is lower. However, we do consider t-channel (see Fig. 1.6), the most abundant source of single top, in this analysis. We use a sample produced by the MadGraph/MadEvent [37] Matrix Element generator and corresponding to $122\,\mathrm{pb}^{-1}$ of integrated luminosity. A filter to exclude hadronic top decays has been applied.

4.2 Simulation of the CMS detector

The full CMS simulation software is a data driven [2], realistic and accurate MC program developed and operating within the CMS software framework, with the purpose of predicting the results of the interaction between the products of the high energy physics processes and the detector. The input to the CMS simulation software is the output of the generator software - a number of final state particles. Using a very detailed description of the detector and the GEANT4 [38] toolkit, capable of simulating physics interactions, these particles are propagated through the different layers of the detector and the resulting analogue signals left on the sensitive detector are estimated. As a final step, the treatment of this signal by the data acquisition

[2]The data in this case is of course not collected on the experiment level, since the experiment is not yet in operation. We refer here to subdetector testbeam data, which has been used to tune the simulation to observed results.

Process	Cross-Section at 14 TeV (pb)	N_{evt}
$t\bar{t}$+0j	619	61900
$t\bar{t}$+1j	176	17600
$t\bar{t}$+2j	34	3400
$t\bar{t}$+3j	6	600
$t\bar{t}$+≥4j	1.5	150
W+2j ($p_{\mathrm{T,W}}$ 0-100 GeV)	2500	2.5×10^5
W+2j ($p_{\mathrm{T,W}}$ 100-300 GeV)	225	2.25×10^4
W+3j ($p_{\mathrm{T,W}}$ 0-100 GeV)	590	5.9×10^4
W+3j ($p_{\mathrm{T,W}}$ 100-300 GeV)	100	10^4
W+4j ($p_{\mathrm{T,W}}$ 0-100 GeV)	125	12.5×10^3
W+4j ($p_{\mathrm{T,W}}$ 100-300 GeV)	40	4.0×10^3
W+≥5j ($p_{\mathrm{T,W}}$ 0-100 GeV)	85	8.5×10^3
W+≥5j ($p_{\mathrm{T,W}}$ 100-300 GeV)	40	4.0×10^3
Single t (incl.)	81.7	10^5
$pp \to \mu X$	229600	2×10^6

Table 4.1: Samples (signal and background) used in the semi-leptonic analysis. The cross-sections given are the effective cross-sections after generator-level filters ($p_{\mathrm{T}}(\mu) > 15$ GeV, $\hat{\mathrm{p}}_{\mathrm{T}} > 15\,\mathrm{GeV}$ for the $pp \to \mu X$ and $t \to Wbl\nu$ for the single top) have been applied.

electronics is also simulated. The final output is a collection of digitized signals for each detector subsystem.

Vertex smearing: Generators produce pp interactions at the nominal point (0, 0, 0). However, colliding bunches at the LHC have a finite thickness and length and the interaction can happen anywhere inside the bunch volume. To account for that in the simulation a treatment of the generator-level information is done before any particles are propagated through the detector. This process moves the generator-level vertices from the nominal position to a randomly selected position. This can be done assuming different models for the distribution of particles within a bunch. The default corresponds to a gaussian distribution of $(\sigma_x, \sigma_y, \sigma_z) = (0.0015, 0.0015, 5.3)$ cm around the nominal point.

Magnetic field: Knowledge of the magnetic field at any point within the detector is a crucial aspect of the simulation. CMS uses a simulation of the magnetic field [39] based on the definition of a number of volumes, inside of which the field is taken to be continuous. The boundaries of these volumes correspond to boundaries between materials of different magnetic permeability. Within each of the volumes, a regular 3D grid is defined. The field has been calculated for each of the points in the grid. Whenever the simulation requires the field at a given point a linear interpolation is performed based on the values of the 8 "corners" of the grid "cube" containing the point.

Tracker: The CMS tracker has been described in section 3.2.2. It is a large structure comprising many cylindrical layers of sensitive parts, which lies within the 4 T magnetic field. All the components, active (silicon) and passive (support material, cables, cooling, electronics etc) are simulated in the detector description. The particles traverse this volume affected by multiple scattering and radiating Bremsstrahlung (if charged). Passing through active medium, charged particles lose energy producing charges that are collected and cause a signal in the electronics. Gaussian noise is added and the signal is digitized. The tracker simulation has been validated

with cosmic ray data and found to be satisfactory [40].

Calorimetry: Simulating the electromagnetic and hadronic showering is the most important issue when it comes to generating the response of the calorimeters. This requires a very good description of the physics processes involved. Recent improvements in this area have been achieved by GEANT. This has been verified in extensive test beams including parts of both ECAL and HCAL detectors, comparing measurements to simulations while taking into account the beam shape. It has been shown that the lateral and longitudinal electromagnetic showers are reasonably well described. Some discrepancy between simulation and test beam data remains with regards to the energy deposit of hadronic showers in the ECAL [40].

An accurate description of the geometry of the calorimeters is also necessary to produce reliable simulated results. Indeed the current implementation includes all detector components in the barrel, endcaps and preshower. ECAL Supermodules and supercrystals are included as separate entities, each of which has its own alignment. Supporting structures, cooling components and electronics are also accurately simulated. Finally, noise and cross-talk effects have been included in the simulation and tuned to recent cosmic ray measurements.

Muon systems: The description of muon Bremsstrahlung, muon-nuclear interactions and multiple scattering provided by GEANT 4 has improved compared to earlier versions and has been validated by comparison to data for energies up to the TeV scale. The geometrical description of the muon systems was tested using data taken during the Magnet Test and Cosmic Challenge (MTCC) and was also found to be sufficiently accurate.

Pile-up events: Pile-up events are events produced temporaly close to the signal event firing the detector trigger. The particles produced in these collisions interact with the detector resulting in additional physics objects being reconstructed and "piled-up" on top of those coming from the signal event. There are two types of pile-up events. The first type, referred to as in-time pile-up events, occur between different hadrons of the same bunch crossing as the signal event. The estimated average number of such events during the low-luminosity phase of the LHC is approximately 5 per bunch crossing. The second type, out-of-time pile-up events, come from bunch crossings occuring recently before and after the signal event. The magnitude of the effect of these events depends strongly on the response time of the front end electronics of the different detector subsystems.

CMS software includes a component dedicated to simulating the effect of pile-up. It enters the simulation path after the simulated tracks and hits have been produced for all detectors, adding pre-constructed hits and tracks corresponding to pile-up events simulated in advance. Therefore, the computationally demanding treatment of the additional particles does not hamper the processing of the main event. The generator information is also expanded accordingly, so as to enable subsequent comparisons to the Monte Carlo truth.

Misalignment of the tracking systems: The CMS detector is a very large and complex device consisting of a large number of different components that need to be very precisely placed in the experimental setup. An ideal detector description such as the default one used by the simulation software assumes that the placement precision achievable during the detector construction is infinite. In reality none of the detector parts will have the exact position and orientation dictated by the design, significantly affecting the position resolution of the measurements, particularly those of the tracking systems. This effect can be corrected for on the software level, however, this requires good knowledge of how well the detector components are aligned in reality. This knowledge is gradually obtained in the course of the experiment by applying methods that are outside the scope of this document. However, estimating how much it affects our measurements at any given stage, is a very important requirement of the simulation software.

For this purpose, in addition to the ideal detector description, there are a number of "scenarios", each corresponding to a different point in the lifetime of the experiment and thus to the degree of knowledge the subdetector experts believe obtainable after collecting a given integrated luminosity of data. Available scenarios include the "startup" or "0pb^{-1}" scenario, assuming that all knowledge of the level of misalignment comes from surveys and cosmic ray measurements, as well as scenarios corresponding to 10pb^{-1}, 100pb^{-1} and 1000 pb^{-1} of integrated luminosity. The software has access to these scenarios and uses them to perform the simulation of the detector response.

Calorimeter miscalibration: The issue of calorimeter calibration is more closely related to the reconstruction rather than the simulation. The reason it is included in this chapter is that it bears a strong similarity to the alignment issue described in the previous section and falls under the same label of "detector conditions".

For each energy-measuring device, there exists a multiplication constant reflecting the efficiency of the device, which connects the measured to the actual energy of the particles. In addition, different components of a calorimeter comprising multiple such devices cannot be expected to have identical responses even to identical incident particles. There are differences in shapes and construction details, as well as imperfections that cause the behaviour exposed by those different components to vary. This means that for each energy measured by a single measuring unit in the ECAL and the HCAL there is a predetermined multiplication factor, specific to this unit, that corrects the measurement. Initially, the determination of the calibration constants of the ECAL and HCAL will only rely on testbeams. As more data is collected by the experiment, significant improvements can be achieved by studying specific physics channels, both in the relative calibration of the crystals (intercalibration) and the absolute energy scale. It is necessary that the software can simulate the effect that the miscalibration has, at different points in the lifetime of the experiment. This is the reason why for each misalignment scenario created to be used in simulated analyses, there is also a miscalibration equivalent. This is essentially an assumption on the average relative error of the calibration constants. It can be used to randomly generate deliberately "wrong" calibration constants that vary from one crystal/tower to the other, so as to affect the energy resolution in a way similar to that expected in reality.

4.3 Fast simulation of the CMS detector

As explained above, simulating the propagation of particles in the detector and the detector response in full detail is a very computationally intensive process. The time it takes to fully simulate a multi-particle event in the CMS detector is the main obstacle we are faced with when trying to obtain large enough samples of simulated data to use for analysis. The solution to this problem is to make some simplifying assumptions when simulating the event that will greatly decrease the time needed. Such assumptions inevitably compromise the accuracy of the process. However, by fine-tuning our assumptions so that the results obtained by this fast simulation are as close as possible to those of the full, we can achieve having a very reliable simulation, which is also many times faster than the GEANT-based one, and which we can use for analysis. The advantage of this approach becomes even clearer, if one considers the amount of data required for studies with strict selection cuts, which only select a very small percentage of the events generated. In such cases a very large initial samples might have to be simulated in order to ensure sufficient statistics after the selection. Additionally, the study of systematic uncertainties related for example to detector conditions usually requires multiple simulation of the same events, increasing the required CPU time by many times.

FAMOS [41],[42], the fast simulation software of the CMS experiment, gets the generator-level information as input, performs a much simplified event simulation and then proceeds to produce the reconstructed objects. The CPU time per event achieved for $t\bar{t}$ events in particular

is approximately 300 ms, whereas the corresponding Full Simulation value is two orders of magnitude larger. Where possible, FAMOS will produce the simulated hit collections. However, in some cases, the physics objects are reconstructed more directly, without the intermediate creation of the exact same objects that the full simulation would produce. We discuss here the main differences between the fast and the full simulation in order to give an account of the simplifying assumptions that have to be made to achieve fast simulation of the data samples and their impact on the analysis.

Tracks: Instead of using the complex tracker geometry of the full simulation, FAMOS uses a very simplified one. In this modelling of the tracker all the material is replaced by cylindrical layers of silicon, one for each active layer of the real detector. The thickness of the layers has been determined by comparing the number of Bremsstrahlung photons radiated by electrons to the full simulation. Multiple scattering and ionization effects are also simulated. Hits are reconstructed on the tracker layers at the points crossed by the particle trajectories, with a predetermined efficiency. Gaussian smearing is applied on the hit positions, according to the average resolution functions predicted by the GEANT-based simulation for single muons. Pattern recognition is not applied as it is a very time-consuming process. Instead, all the hits resulting from a particle are used to perform the track fitting, which is identical to that of the standard reconstruction.

Electromagnetic showers: The ECAL is treated by the fast simulation as a homogeneous medium. Electron showers are parametrized using the Grindhammer parametrization [43], which results in thousands of energy deposition spots. The spots are distributed longitudinally according to a Γ function (of flactuating parameters) and laterally by the sum of two functions, one for the core of the shower and one for the tail. The number of spots used is reduced as much as possible, to maximize speed without loss of simulation accuracy. The spots are then transfered to the ECAL and the deposition per crystal in a 7×7 grid is calculated taking into account the effects of rear leakage, gaps between crystals and shower broadening due to the magnetic field (in the barrel only). In case an electromagnetic shower is energetic enough to reach the HCAL this energy will be added to the HCAL hits. Photons are treated as two electron showers starting from a pair creation.

Hadron showers: The simulation of hadron showers in FAMOS is based on the full simulation of charged pions in a 2-300 GeV/c p_T range and with a uniform η-distribution. The resulting reconstructed energies in the ECAL and HCAL are grouped into 0.1 η-bins and each bin is fit using a Gaussian. The result is an $\eta - p_T$ grid of gaussians. Based on that grid the smearing applied on the energy of each hadron to get the energy of the fast simulation shower is determined by interpolation (in the 2-300 GeV/c range) or extrapolation (for higher p_T's). The lateral and longitudinal shower shapes are parametrized in a way that achieves the optimal blance between speed and H/E agreement with the full simulation.

Muons: Muons in the version of FAMOS (1_6_X) used for this analysis are only propagated up to the ECAL entrance. The simulated hits for the muons are therefore not produced. The muon objects are instead reconstructed using parametrizations of the calorimeter response and the muon chambers. The process is tuned to give similar efficiencies and position resolution as the full simulation. This has an impact on the ability to apply misalignment scenarios to the muon reconstruction. Recently, FAMOS has been extended to include the simulation of hits in the muon detectors. In the context of this thesis, work has been contributed to the application of muon misalignment scenarios.

4.3.1 Simulation-level samples

All generator-level files described in Sec. 4.1.5 were part of centrally produced, fully simulated and reconstructed datasets, distributed in various storage elements around the world. However, this analysis targets the first data to be collected by the CMS experiment. These ready-made datasets were only simulated under detector conditions corresponding to a later stage of the experiment, when the detector is expected to be better calibrated and aligned.

For this analysis, aiming at the very early stage of data taking, it was decided to access the available samples using the grid technology developed by the experiment, strip them of the simulation and reconstruction parts keeping only the generator-level information and copy them locally. It was then necessary to run the detector simulation and the reconstruction using the available local resources, assuming the detector conditions expected at startup. Given the large size of data required to perform the analysis, this was achieved using the Fast Simulation. Furthermore, the Fast Simulation offered the very important possibility of producing additional samples with the same generator information, but with changing conditions, so as to provide an estimate of the systematic uncertainties related to the analysis. Thus, the datasets from each of the generator-level samples were produced assuming:

- a detector under startup miscalibration and misalignment conditions,
- an ideal detector with no miscalibration and misalignment,
- a detector under startup miscalibration and misalignment conditions and an average of five pileup events per bunch crossing,
- a detector under startup miscalibration and misalignment conditions with jet energies rescaled by 10% (see Sec. 5.2.3 on jet calibration).

These tasks were performed exclusively in the context of this thesis work.

Chapter 5

Reconstruction of single-leptonic $t\bar{t}$ events

The purpose of this chapter is to provide a short overview of the reconstruction of the basic physics objects that bear relevance to a single-leptonic $t\bar{t}$ analysis. A much more detailed description of the reconstruction software of the CMS experiment can be found in [32], chapters 9-12.

5.1 Final state objects

As explained in Ch. 1, the single-leptonic decay of a $t\bar{t}$ pair results in the production of a lepton, its neutrino and a b-quark from the leptonically decaying t-quark and two light quarks and a b-quark from the hadronically decaying t-quark (Fig. 5.1, left). For this analysis in particular, signal decays involve a muon as the lepton. The interaction of the muon and the neutrino with the detector and the experimental signatures they result in have been covered elsewhere (Sec. 3.1). The quarks hadronize and produce particle showers, typically concentrated in cone-like volumes, many of which reach the calorimeters and deposit their energy there. The signature left in the detector is therefore a concentrated energy deposit in the electromagnetic and hadronic calorimeters combined with tracks originating from a common vertex reconstructed in the inner tracker. A calorimeter-based object which is used as the observable equivalent of the shower-generating parton is called a jet. We therefore expect at least four jets in a single-leptonic $t\bar{t}$ decay (provided that all four showers were within the calorimeter acceptance). Additional ones frequently occur from other sources, for example initial and final state gluon radiation or pileup events.

In conclusion, when looking for single-muonic $t\bar{t}$ decays, one can expect to reconstruct the following objects

- at least four jets representing the two light and the two b-quarks,
- a muon from the leptonically decaying W,
- missing E_T ($\not{E}_T$) representing the neutrino.

Furthermore, the ability to reconstruct electrons also plays an important role e.g. in separating single-muonic events from the di-lepton background (Fig. 5.1, right), which can mimic our signal if additional jets are present. A graphical representation of a signal event in the CMS detector can be seen in Fig. 5.2.

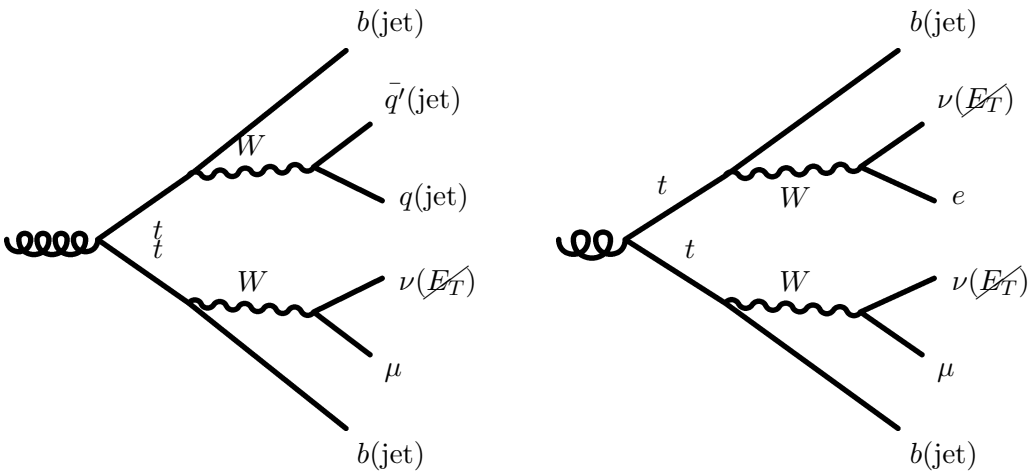


Figure 5.1: The final-state objects of a single-muonic $t\bar{t}$ decay (left) are four jets corresponding to the two b and the two light quarks, a muon and missing transverse energy corresponding to the neutrino. Electrons are also relevant, as we can use them to identify dilepton ($t\bar{t} \to b\bar{b}e\nu_e\mu\nu_\mu$) background events (right), with e.g. an additional radiated gluon (not shown) replacing the missing jets.

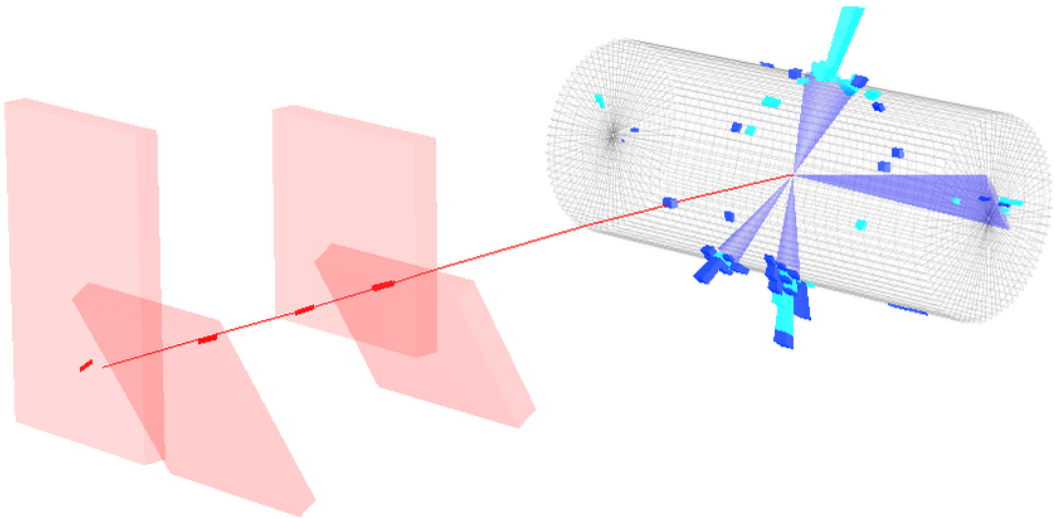


Figure 5.2: A typical $t\bar{t} \to b\mu\nu bq\bar{q}'$ event with four jets and an isolated muon. The 3D image was produced using the Fireworks event display. The calorimeter energy depositions and reconstructed jet cones appear in shades of blue, whereas the isolated muon track and the hit chambers are in red.

5.2 Jet reconstruction

5.2.1 Calorimeter towers

The basic building block for a jet is the calorimeter tower. Since the ECAL has a finer granularity than that of the HCAL, a calorimeter tower consists of an HCAL cell plus the ECAL cells covering the same $\eta - \phi$ area. The energy of the tower is therefore defined as the sum of the energies of these cells. For the purposes of jet reconstruction, each of the towers in the calorimeters is treated as a massless particle of an energy equal to the tower energy and direction from the nominal vertex to the center of the tower. Noise-reducing E_{T} cuts are applied at both the individual cell and the tower level. After the towers are reconstructed they are input to the jet reconstruction algorithms.

5.2.2 The iterative cone algorithm

There are several jet algorithms implemented and studied for the CMS experiment. For the purpose of this analysis the algorithm used is the iterative cone algorithm, therefore it is briefly covered here.

The calorimeter towers are ordered by E_{T}. Starting from the highest E_{T} tower, we define a cone in the $\eta - \phi$ space around it and combine all the objects included in that cone to determine E_{T} and direction for the jet. Following that, the cone is redefined around the new jet direction and the process is repeated until the E_{T} and direction of two subsequent jets remains practically unchanged. The way of combining multiple towers into a single E_{T} and θ direction is by equating the jet E_{T} with the sum of the E_{T} of the individual towers and calculating θ from the formula $\sin\theta = \frac{\sum E_{\mathrm{T}}}{\sum E}$. After a jet is fixed, the constituent towers are removed from the list and the process is repeated for the highest E_{T} tower in the updated list.

5.2.3 Jet calibration

When reconstructing jets we normally hope to end up with objects that we can associate with the partons resulting from the hard process of the event. Simply taking the energy of the jet reconstructed with the iterative cone algorithm as the energy of that parton is naive. In reality, the way the jet energy is associated with the energy of the corresponding parton depends on

- the algorithm employed for the jet reconstruction and its parameters,
- the kinematic properties of the parton (obviously, when reconstructing real data, we only have access to the reconstructed object),
- the type of the parton. This affects the composition and the structure of the resulting shower and therefore the efficiency of the total energy recovery of the reconstruction algorithm.

For these reasons, the energies of the reconstructed jets have to be corrected. The method used to that end consists of three steps.

First, the reconstructed energy of the jet is reduced by a fixed offset. This is meant to account for energy deposited in the calorimeters as a result of in- and out-of-time pile-up and the underlying event and for the effects of electronic noise.

Second, the particle-level calibration is applied. This is done based on the E_{T} and η of the jets and uses the scale factors obtained from Monte Carlo studies. The scale factors have been determined by running the jet algorithm on large samples of dijet events, with both the calorimeter towers and the stable generator-level particles as input. Reconstructed and "generated" jets are produced respectively and a matching based on distance in the η, ϕ-space follows. The ratio of the generated over the reconstructed jet energy is parametrized by jet E_{T} and η

and the energy rescale factors are obtained. As a result of applying these factors, the energy of an "average" QCD jet can be scaled back to the energy of the corresponding generator-level particle jet.

The third and final step of the jet energy correction accounts for the fact that different kinds of partons result in jets of different composition and structure. Gluon and b-quarks in particular tend to result in lower energy response than light quarks. Therefore, if we want to correct the energy scale to the parton level, we have to hypothesize on a particular flavor for the jet and apply the corresponding factor.

5.3 Muon reconstruction

Muon reconstruction in CMS is based on the principle of regional reconstruction. Reconstructing a muon basically amounts to reconstructing a track. When performed across all of the tracking devices of the detector, this task is very CPU-intensive. The principle is thus to restrict the area considered for the muon track, based on the measurements made by the muon systems alone.

The starting points for the muon reconstruction are the hits in the muon systems, DTs, CSCs and RPCs (described in Sec. 3.2.5). Track segments reconstructed in the innermost chambers, constitute the seeds of the tracks. Starting from these seeds, trajectories are built by extrapolation to the outside layers using the Kalman filter procedure. "Bad" hits are rejected by a cut on their χ^2 value. When all hits for the track are selected the procedure is reapplied from the outside in, so as to achieve an accurate estimation of the track parameters for the innermost hit. Based on this result the track is propagated to the nominal interaction point and a vertex-constrained fit is performed. The result of this is referred to as a stand-alone muon, as no inner tracker information is used to reconstruct it.

The reconstruction of a global muon is done by extrapolating the innermost hit of the standalone muon to the tracker layers. As a result of this process a region of interest is defined on each of those layers, the size of which is based on the track parameters and their uncertainties. A pair of hits belonging to different tracker layers within the region of interest constitutes a seed/track candidate. A relaxed beam spot constraint is applied to obtain initial trajectory parameters and then a track is reconstructed from each of the seeds by means of the Kalman filter technique. Each compatible hit on each of the tracker layers (from the inside out) results in a new trajectory corresponding to the original seed and each of the new trajectories is propagated to the next layer. At the end of this process a single trajectory is selected for each of the seeds, based on its χ^2 and number of hits. A final fit using the hits from the tracker trajectory and the corresponding muon system trajectory is performed, unconstrained by the beam spot. Any persisting ambiguities are resolved by a final cleaning step selecting tracks based on the χ^2 value of this final fit. The tracks passing this final selection constitute the global muon objects.

5.3.1 Muon isolation

Muon isolation is a tool used to distinguish the muons that come from heavy particle decays from those that appear within jets. This is very important because most of the muons in the QCD background are produced in b and c quark decays and come as part of a jet, contrary to the muon from the W boson in signal events, which is usually isolated. We can therefore cut on the isolation to reduce the contribution of this source of background. Calculating the muon isolation is based on the definition of a cone in the η-ϕ space, around the momentum of the muon at vertex (Fig. 5.3). We then define the following two quantities:

- The "track isolation", $E_{\text{iso}}^{\text{trk}}$, which is the scalar sum of the p_{T} of all the reconstructed tracks included in the isolation cone, except for the track of the muon itself,

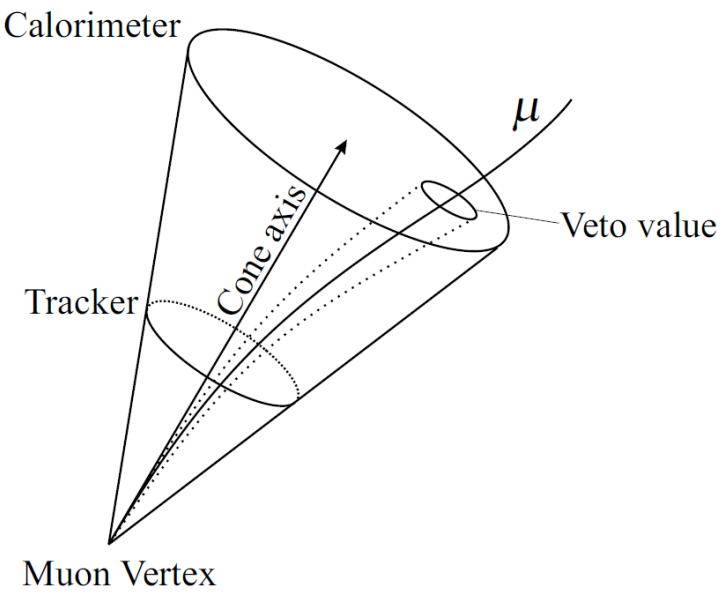


Figure 5.3: The cone in the η-ϕ space is centered around the momentum of the muon at vertex. The tracker isolation is calculated from the p_{T} of the tracks inside the cone. The calorimeter isolation is calculated from the calorimeter energy deposited inside the cone. The "veto value" is excluded from the sum.

- The "calorimeter isolation", $E_{\mathrm{iso}}^{\mathrm{calo}}$, which is the weighted sum of the transverse energy deposited in the electromagnetic and the hadronic calorimeters within the isolation cone. The energy deposited in a small area around the muon (called the veto value) is attributed to the muon itself and therefore excluded.

$$E_{\mathrm{iso}}^{\mathrm{calo}} \equiv \alpha(E_{\mathrm{T}}^{\mathrm{ECAL}} - E_{\mathrm{T}}^{\mathrm{veto,ECAL}}) + (E_{\mathrm{T}}^{\mathrm{HCAL}} - E_{\mathrm{T}}^{\mathrm{veto,HCAL}}), \text{ where } \alpha = 1.5 \tag{5.1}$$

The value of α reflects the better performance of the electromagnetic calorimeter. The largest $E_{\mathrm{iso}}^{\mathrm{trk}}$ and $E_{\mathrm{iso}}^{\mathrm{calo}}$ are, the more likely the muon is to belong to a jet.

5.4 Electron reconstruction

An electron created in the center of the CMS detector follows a curved trajectory towards the ECAL under the influence of the 4 T magnetic field. In the absence of the tracker material a single electron would lead to an energy cluster in the calorimeter, extending over a small number of adjacent crystals. In reality, an electron will radiate Bremsstrahlung photons that will carry away some of its energy and potentially deposit it as separate clusters in the ECAL. What will finally reach the ECAL is a spray of energy, spread mostly in the ϕ direction. There are therefore two initial steps to be taken when reconstructing an electron: To group adjacent crystals into clusters and to find out which separate clusters correspond to the same initial electron.

The Island clustering algorithm addresses the first issue by starting from the local maxima of energy deposition, progressing outwards in all directions and adding crystals of decreasing energy to the cluster. The process stops when a rise in energy or an energy "hole" is encountered. A position is then calculated for the cluster using an algorithm that takes into account the orientation of the calorimeter crystals and their individual energy contributions to the clusters as well as the typical shape of an electromagnetic shower. After the clustering is complete, the Bremsstrahlung Recovery algorithm sorts the clusters by energy and starts from the most

energetic, adding the clusters included in a fixed-size array around it. The array is wide in ϕ and narrow in η, to reflect the shape of the typical shower of Bremsstrahlung energy. The resulting object is a supercluster. The position of a supercluster is defined as the energy-weighted mean of the positions of the constituent clusters and corresponds to the position of impact that the initial electron would have in the absence of Bremsstrahlung.

The next step in reconstructing the electron is to propagate the supercluster position through the magnetic field, back to the layers of the pixel detector, assuming that it has come from the nominal vertex. Using a window wide in z and narrow in ϕ we look for matching hits in the layers of the pixel detector starting from the innermost layer. When a compatible hit is found it is used to constrain the search in the ϕ dimension and a second compatible hit is looked for in the next tracker layers. Any combination of two such compatible hits and the corresponding supercluster constitutes an electron seed. From this seed, tracks are reconstructed in the tracker using the Kalman filter technique. The final product, consisting of a supercluster and an associated track is the electron object.

The definition of the electron isolation is analogous to that of the muon. The main difference is in the calculation of the calorimeter isolation, where we do not define a "veto area" around the electron. Instead, we subtract the energy of the corresponding supercluster from the total calorimetric deposition in the isolation cone.

5.5 A note on b-tagging and missing transverse energy

So far in this chapter, we have covered all the types of reconstructed objects used in this work. The CMS reconstruction software, however, is capable of providing additional information, that could be of use to an analysis of single-leptonic $t\bar{t}$ decays. This information comes in the form of b-jet tags and the quantity commonly refered to as missing transverse energy ($\not\!\!E_T$). It is thus useful to briefly describe these objects of the reconstruction, so as to justify the decision not to use them.

5.5.1 b-tagging

Being able to identify jets originating from b-quarks can be important for a t-quark related analysis, as the t-quark decays almost exclusively into a b-quark and a W. As will be shown in Sec. 9.1.3 the selection and jet-parton assignment can be significantly improved by using such methods. Since many other physics processes contain b-quarks as well, CMS has implemented several algorithms capitalizing on the unique properties displayed by b-hadrons in order to tag the resulting jets as b-jets. Such properties include the relatively large lifetimes and the large semileptonic branching ratios.

More specifically, if very precise tracking is available, it can be used to efficiently reconstruct the primary vertex of the event. If the origins of tracks associated with a particular calorimeter jet are found to be significantly displaced from the primary vertex, this is a strong indication of the jet originating from a B-hadron. Some b-tagging algorithms are also capable of reconstructing the "secondary vertex" of the event and calculating the distance from the primary vertex. The decision to tag the jet can then be taken based on this distance.

As explained in Sec. 4.2, during the initial phase of data taking, the CMS tracker will not be very well aligned. This will have a substantial impact on the quality of the reconstructed tracks. It would therefore not be very safe to assume that the b-tagging algorithms will perform robustly and design the analysis relying on this assumption. This is not to say that use of b-tagging as a possibility should be outright excluded. In Sec. 9.1.3 we describe a first look into one of the b-tagging algorithms that shows some promise.

5.5.2 Missing transverse energy

Owing to momentum conservation, the vectorial sum of all the transverse components of the momenta of the particles produced in a hadron collision should be almost zero (the momentum of the initial partons might have a small transverse component). Since the CMS calorimetry covers the area around the interaction point almost hermetically, summing all the individual E_{T} contributions in the calorimeters leads to an estimation of the total transverse component of all the objects "invisible" to them.

The missing transverse energy vector in CMS is the vectorial sum of the transverse energy measured in the calorimeter towers

$$\overrightarrow{\not{E}}_T = \sum E_n \sin\theta_n \cos\phi_n \overrightarrow{i} + E_n \sin\theta_n \sin\phi_n \overrightarrow{j} \tag{5.2}$$

This has to be corrected in the presence of muons, by subtracting their calorimeter deposition (approximately 2 GeV for a typical muon) and adding their energy as estimated by their tracks. Corrections to the jet energy should also be taken into account to more accurately determine $\not{E}_T$.

Missing transverse energy is arguably the most difficult quantity to reconstruct. Pile-up effects, differences in the calorimeter response to different particles, the bending of tracks by the strong magnetic field, the calorimeter miscalibration, the tracking systems misalignment, and the jet energy and position resolution uncertainties, all play a part in degrading the measurement. Consequently, this object is less suited to analyses performed with early data obtained from a poorly understood detector. It is for this reason that this analysis does not make use of $\not{E}_T$. It is nevertheless certain, that a study performed at a later time in the lifetime of the experiment will utilize this quantity as a powerful discriminating tool between signal and background.

Chapter 6

Selection

The purpose of this analysis is to estimate the amount of data that CMS needs to collect so as to claim an early observation of $t\bar{t}$ pairs decaying semileptonically and to make a first measurement of the cross-section for $t\bar{t}$ production. We are targeting an integrated luminosity of a few pb^{-1}. As already stated in the previous chapter, during this initial phase of operation, the detector will not be well understood and energy and position of the reconstructed objects will be measured at a reduced accuracy due to misalignment of the tracking systems and miscalibration of the calorimeters. This reality imposes the following prerequisites on the analysis (see also Sec. 5.5):

- Missing transverse energy should not be used. $\not\!\!E_T$ measurement requires good knowledge of the detector and is particularly sensitive to calorimeter calibrations. It therefore cannot be relied upon for an early analysis.

- The method should not rely on b-tagging algorithms. The performance of b-tagging depends on the precision of the track and vertex reconstruction. The misalignment of the inner tracker during the first period of data-taking makes their usefulness uncertain.

Given those two requirements, the purpose of the studies detailed in this chapter was:

- To define a selection strategy that ensures the collection of a $t\bar{t}$-rich sample, keeping the contribution of the background processes to a minimum.

- To investigate ways to efficiently assign the jets in a selected event to the final state partons of the $t\bar{t} \to b\mu\nu bqq\bar{}'$ decay.

6.1 Preselection

A basic set of very loose cuts is first applied on the signal and background samples, with the aim of reducing the number of events to be analyzed to a manageable size:

- at least two jets with $E_{\mathrm{T}} > 10\,\mathrm{GeV},\ |\eta| < 2.4$,

- at least one muon with $p_{\mathrm{T}} > 20\,\mathrm{GeV/c},\ |\eta| < 2.1$.

The second of these requirements also serves to fulfill the High Level Trigger (HLT) acceptance. HLT is not explicitly included in the analysis, however, it has been shown to have a flat efficiency of approximately 90% for muons in that range (as a function of p_{T}). This means that the HLT efficiency can be taken to be 90% for all events satisfying the preselection conditions, without large error.

	$t\bar{t}$ signal	other $t\bar{t}$	single t	$W + jets$	QCD	total B	S/B
preselection	8844	8317	15500	63460	3.4×10^6	3.4×10^6	2.6×10^{-3}
di-lepton veto	7932	3602	13780	60491	455874	533747	1.5×10^{-2}
loose μ iso	5030	2070	7751	35931	5425	51177	9.8×10^{-2}
loose $\mu + E_{\mathrm{T}}^{\mathrm{jet}}$	1542	238	77	538	115	967	1.6
tight μ iso	3646	1529	5477	25716	356	33078	1.1×10^{-1}
tight $\mu + E_{\mathrm{T}}^{\mathrm{jet}}$	1070	155	55	396	< 11	606	1.8

Table 6.1: Number of events selected after the background-reducing cuts described in section 6.2. The signal and the different sources of background are displayed in separate columns. "Other $t\bar{t}$" corresponds to $t\bar{t}$ events that do not decay by the single-muonic channel. "Single t", "$W + jets$" and "QCD"($p\bar{p} \rightarrow \mu X$) refer to the samples described in Sec. 4.1.5. The last two columns display the total number of background events and the signal to background ratio. Each line corresponds to the selected sample after each of the selection cuts of the analysis.

6.2 Background reduction

This first step of the selection has a high efficiency for the semi-leptonic $t\bar{t}$ signal, however this signal is overwhelmed by the very large background. The various sources of background can be grouped into the following three categories based on the muons included in the event:

- Events with one muon and an additional lepton, in which neither the muon nor the additional lepton belongs to a jet. We refer to such muons as *isolated*. Z+jets events with the Z decaying into muons as well as $t\bar{t}$ di-leptonic decays with at least one muon are the most important contribution to this type of background.
- Events with muons which belong to jets (non-isolated). QCD events with muons coming from b- or c-quark decays form the bulk of this background.
- Events with a single, isolated muon and no other isolated leptons. This background mostly comes from W+jets events with the W decaying into a muon and a neutrino.

Various cuts are introduced to suppress events from each of these different background sources, while keeping as much of the $t\bar{t}$ signal as possible. A separate study has been performed to determine how to reduce each different type of background, while keeping as much of the signal as possible. These studies are detailed in the following three sections.

6.2.1 The di-lepton background

The background from di-lepton events is relatively easy to suppress. Whether they come from di-leptonic $t\bar{t}$ decays or Drell-Yan/Z+jets, it is sufficient to require that there is no second isolated lepton in the events. Yet, it is important to pick a sensible definition of the isolation requirements. In the following we study the efficiency and rejection as a function of the isolation variables in order to make an optimum choice.

The definition is a tradeoff between the need for high efficiency, so that the isolated leptons are not missed and the need to not mistake non-isolated muons for isolated. The consequence of the latter would be to mistake signal events for di-lepton events, as the b-jets in the $t\bar{t} \rightarrow bW^+\bar{b}W^-$ often contain leptons. It is therefore useful to know how the efficiency of identifying really isolated leptons relates to the efficiency of misidentifying non-isolated leptons as isolated.

To understand this, the following exercise is performed. We select all reconstructed muons in the $t\bar{t}$ sample, separating those that come from W-bosons (approximately 10^4) and those that come from b-jets (approximately 1.7×10^4). We refer to these muons as "good" and "bad"

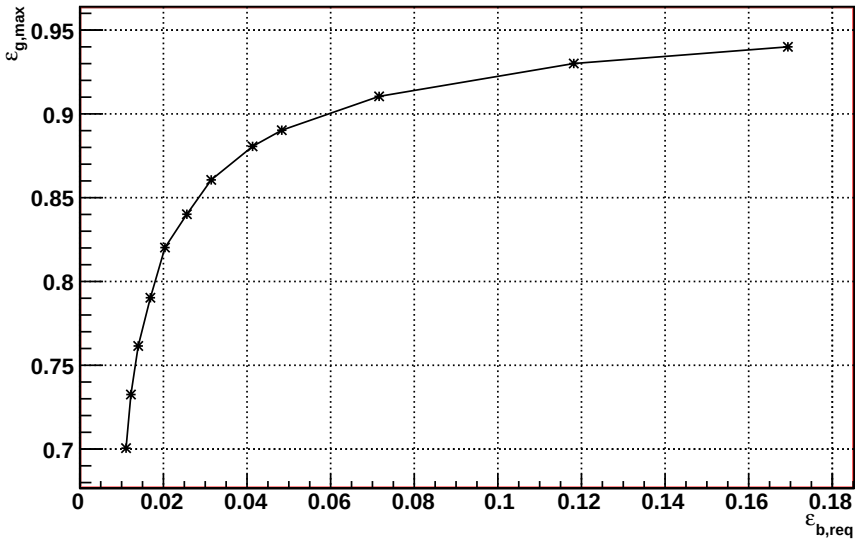


Figure 6.1: Maximum efficiency on muons from W-bosons $\varepsilon_{\mathrm{g,max}}$ as a function of the required efficiency on the "bad" muons $\varepsilon_{\mathrm{b,req}}$.

respectively. Our definition of the isolated muon will use cuts on three variables, one on muon p_{T}, one on calorimeter isolation $E_{\mathrm{iso}}^{\mathrm{calo}}$ and one on tracker isolation $E_{\mathrm{iso}}^{\mathrm{trk}}$. Each of these cuts is allowed to vary by small incremental steps inside a given range, independently from the other two:

- $p_{\mathrm{T},\mu} > p_{\mathrm{T,min}}$, with $10\,\mathrm{GeV} < \mathrm{p_{T,min}} < 20\,\mathrm{GeV}$,
- $E_{\mathrm{iso}}^{\mathrm{calo}} < E_{\mathrm{iso,max}}^{\mathrm{calo}}$, with $0\ \mathrm{GeV} < E_{\mathrm{iso,max}}^{\mathrm{calo}} < 10\,\mathrm{GeV}$,
- $E_{\mathrm{iso}}^{\mathrm{trk}} < E_{\mathrm{iso,max}}^{\mathrm{trk}}$, with $0\ \mathrm{GeV} < E_{\mathrm{iso,max}}^{\mathrm{trk}} < 10\,\mathrm{GeV}$.

In this way, a grid in the three dimensional space of cuts is defined. Each point in the grid corresponds to a different definition of muon isolation. For each of these definitions, the efficiency on the "good" and the "bad" samples can be calculated. If then we require a (preferably low) efficiency on the "bad" muons, $\varepsilon_{\mathrm{b,req}}$, we can run over the points in the grid and choose the one that fulfills this requirement and results in the maximum efficiency on the "good" muons $\varepsilon_{\mathrm{g,max}}$. Doing this for different values of $\varepsilon_{\mathrm{b,req}}$, we can thus find the corresponding $\varepsilon_{\mathrm{g,max}}$ and draw the one versus the other (Fig. 6.1). Selecting the cuts defining the isolated muon is then reduced to selecting a point in the $\varepsilon_{\mathrm{b,req}}$-$\varepsilon_{\mathrm{g,max}}$ curve.

In order to make sure that very few di-leptonic events remain in the selection, it is sufficient to keep the efficiency for isolated muons in $t\bar{t}$ events higher than 90%. The probability to identify a di-muon event as such is then at least 81% (in fact it is slightly higher because it is also possible to mis-identify correctly by mistaking a soft muon for isolated). We conclude on the following definition of the isolated muon:

- $p_{\mathrm{T},\mu} > 10\,\mathrm{GeV}$.
- $E_{\mathrm{iso}}^{\mathrm{calo}} < 6\,\mathrm{GeV}$.
- $E_{\mathrm{iso}}^{\mathrm{trk}} < 5\,\mathrm{GeV}$.

which result in $\varepsilon_{\mathrm{g}} = 90.7\%$ and $\varepsilon_{\mathrm{b}} = 6.9\%$ (see Fig. 6.1). This is our way to count isolated muons when it comes to rejecting events with more than one lepton. Electrons play a less important role in this analysis, since they are only useful in supressing an already small source of background, it was therefore decided not to repeat the optimization study for electrons. Instead, the isolation definition recommended by the CMS top physics analysis group for dilepton analyses ($p_{\mathrm{T}} > 10\,\mathrm{GeV}$, $|\eta| < 2.4$, $E_{\mathrm{iso}}^{\mathrm{calo}} < 6\,\mathrm{GeV}$, $E_{\mathrm{iso}}^{\mathrm{trk}} < 3\,\mathrm{GeV}$) has been used without further investigation. Given these definitions of the isolated muon and the isolated electron, we require that there is exactly one isolated muon in the event and no other isolated lepton.

The effect of the cut can be seen in Tab. 6.1. We see that indeed, 90% of the signal survives the cut, whereas the dileptonic $t\bar{t}$ ($t\bar{t} \to bl\nu bl\nu$) are drastically reduced, leading to a 57% reduction in the "other $t\bar{t}$" background. The QCD background is also reduced by an order of magnitude, but it should be noted that this is because of the tighter cut on the first muon, not the veto on the second (a targeted cut to reduce the QCD background will be explored in Sec. 6.2.2). The overall signal to background ($\frac{S}{B}$) ratio is increased from 2.6×10^{-3} to 1.5×10^{-2}, but the really important effect of the cut is the $\frac{S}{B_{\mathrm{other}t\bar{t}}}$ increase from 1.1 to 2.2.

Drell-Yan/Z+jets background is not considered for this analysis. The cross-section for these processes is much smaller than that of the other backgrounds considered and the addition of the dilepton veto can be relied on to make their contribution insignificant.

6.2.2 The QCD background

After the preselection, the QCD background is overwhelming (three orders of magnitude larger than the signal, see Tab. 6.1). Useful as it is in reducing dilepton events, the definition of the isolated muon detailed in the previous section is not effective enough to control the QCD. If however, we use a tighter definition, we can exploit the fact that muons in QCD events are usually parts of jets and are therefore not isolated. This time we use the $p\bar{p} \to \mu X$ sample (see Sec. sec:gendatasets) to get the undesired ("bad") muons. Additionally, we do not only include the p_{T} of the muon, $E_{\mathrm{iso}}^{\mathrm{calo}}$ and $E_{\mathrm{iso}}^{\mathrm{calo}}$ in our scan of the cuts, but also the minimum distance of the muon from the closest jet in the $\eta - \phi$ space, $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$. We only consider jets of $E_{\mathrm{T}} > 15\,\mathrm{GeV}$ to calculate this distance.

The required efficiency on background muons is now dictated by the size of the QCD background - in order to bring it to the same order of magnitude as the signal, we need to reduce it by at least three orders of magnitude (given that our cuts will also somewhat reduce our signal) [1]. The number of muons in both the "good" and the "bad" samples roughly correspond to the number of signal and QCD events respectively (since there is only one muon in the majority of the events). We can therefore expect our cuts to reduce the QCD background by approximately the same fraction as they reduce the "bad" muon sample. This time, for different required "bad" muon efficiencies we find the maximum achievable "good" muon efficiency and plot the latter versus the former (Fig. 6.2). We also plot the expected signal to background ratio as a function of the former.

There are two ways to proceed. The first one is to assume that the shapes of different variables (e.g. jet multiplicity, jet transverse energy) of the QCD background can be estimated reasonably well in some way. In this case we do not need to get rid of QCD background completely. Bringing it to the same order of magnitude as the signal suffices. The second approach does not assume any knowledge of the shapes of QCD background and tries to practically eliminate it by reducing it to a level one order of magnitude smaller than the signal. We thus define two points in the curves of Fig. 6.2, one that corresponds to expected signal to background ratio of 1 and one to 10. The corresponding sets of cuts are:

[1] We consider muons passing the preselection stage, so as to keep the size of the "good" muon sample as high as possible, we therefore refer to the desired reduction in terms of the size of the preselection samples, not the samples surviving the dilepton veto

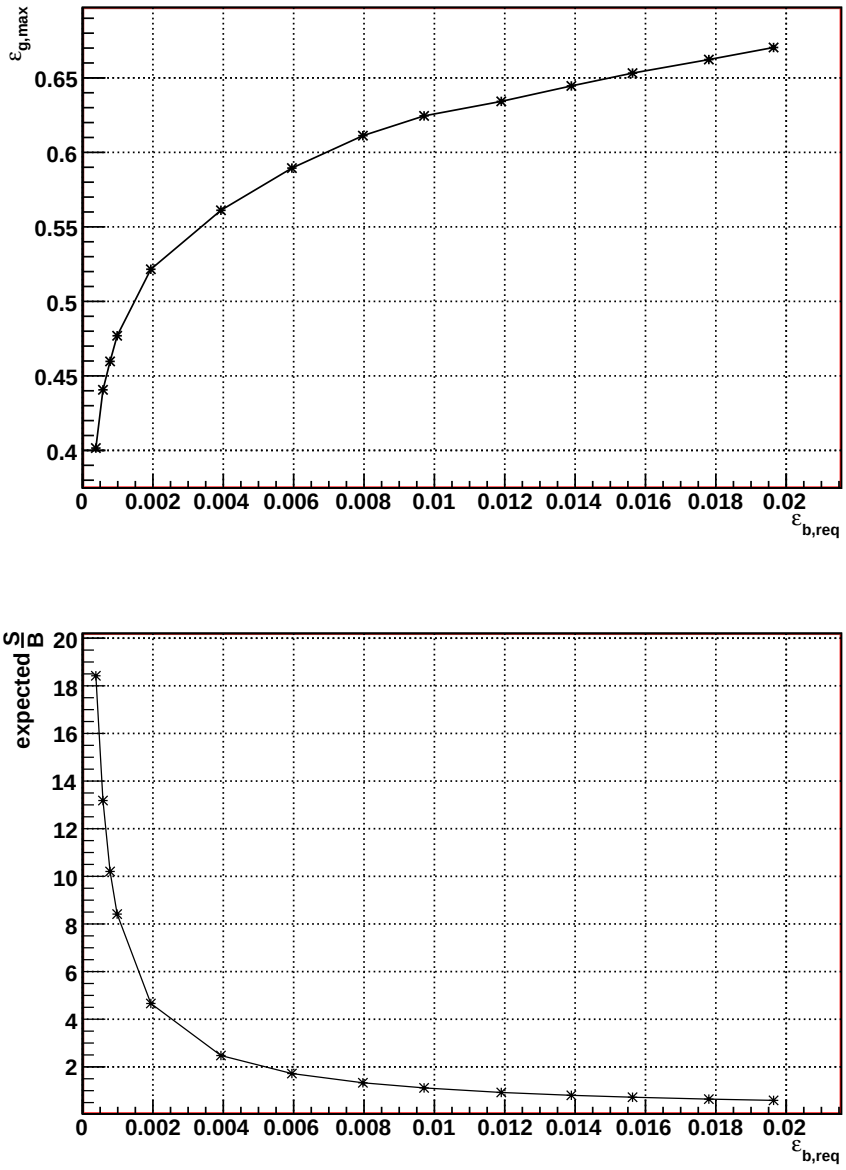


Figure 6.2: Maximum efficiency on "good" muons from W-bosons (top) and expected signal to QCD background ratio (below) as a function of the required efficiency on the muons in the QCD sample.

1. **Loose cuts**, corresponding to $\varepsilon_{\rm g} \approx 63\%$, $\frac{S}{B} \approx 1$ (before the final jet $E_{\rm T}$ cut):
   - $p_{T,\mu} > 32$ GeV,
   - $E_{\rm iso}^{\rm calo} < 2$ GeV,
   - $E_{\rm iso}^{\rm trk} < 2.75$ GeV,
   - $\Delta R_{min} > 0.075$.
2. **Tight cuts**, corresponding to $\varepsilon_{\rm g} \approx 46\%$, $\frac{S}{B} \approx 10$ (before the final jet $E_{\rm T}$ cut):
   - $p_{T,\mu} > 37$ GeV,
   - $E_{\rm iso}^{\rm calo} < 1.0$ GeV,
   - $E_{\rm iso}^{\rm trk} < 1.5$ GeV,
   - $\Delta R_{min} > 0.15$.

6.2.3 The W+jets and single top background

We cannot expect to use the muon to reduce W+jets background, because the muon of this background is also isolated. In fact, the W+jets background is difficult to reduce drastically. However, the jets in the single leptonic $t\bar{t}$ signal have a larger multiplicity than in W+jets background, as we start from a minimum of four jets even without any additional jets from gluon radiation. The high invariant mass of the $t\bar{t}$ system also means that jets in $t\bar{t}$ events will tend to have higher energies. We can therefore hope to somewhat reduce the W+jets background by placing an $E_{\rm T}$ cut on the fourth most energetic jet in the event.

Single top is also a source of background that enters our selection by a isolated muon. However, when a single top event does have this isolated muon it can only mean that it does not have a large number of jets, as the only top quark in the event has decayed leptonically. This is the reason why the jet $E_{\rm T}$ cut can be relied on to eliminate this source of background too.

In order to place the cut in the most economic way possible, we start by placing it at a very low value (15 GeV), with the purpose of tightening it until the desired fraction of signal over background is achieved. Therefore, the summary of our selection requirements up to this point is:

- At least one muon (isolated or non-isolated) of $p_{\rm T} > 20$ GeV and $|\eta| < 2.1$, to fulfill the HLT acceptance.
- Exactly one isolated muon ("loose" or "tight", as defined in 6.2.2) in $|\eta| < 2.4$ [2].
- No second isolated lepton (as defined in 6.2.1) in $|\eta| < 2.4$.
- At least 4 jets of $E_{\rm T}$ >15 GeV.

The shape of the $E_{\rm T}$ of the fourth leading jet [3] for signal and background after these cuts is shown in Fig. 6.3, 6.4. Predictably, the fourth leading jet tends to be more energetic in signal than in background events. The background distribution falls sharply already from 15 GeV, whereas the signal distribution does not display a clear drop until the region of 35 GeV. This confirms that a cut on the fourth leading jet $E_{\rm T}$ can be very effective in improving the purity of the selected sample.

[2] At first sight, this cut may seem to contradict the cut introduced earlier in the selection to ensure a flat High Level Trigger efficiency (requiring a muon of $|\eta| < 2.1$). However, the muon that fires the trigger on a signal event is not necessarily the isolated muon from the top decay that we are trying to identify here. It might well be a non-isolated muon from a b-quark decay. Increasing the allowed $|\eta|$-range here does not mean moving to a lower HLT efficiency region, as the event has already passed the requirement for an "HLT" muon at this stage.

[3] By fourth leading jet, we mean the fourth jet in order of transverse energy. A cut on the $E_{\rm T}$ of the fourth leading jet thus also affects the three leading jets as well.

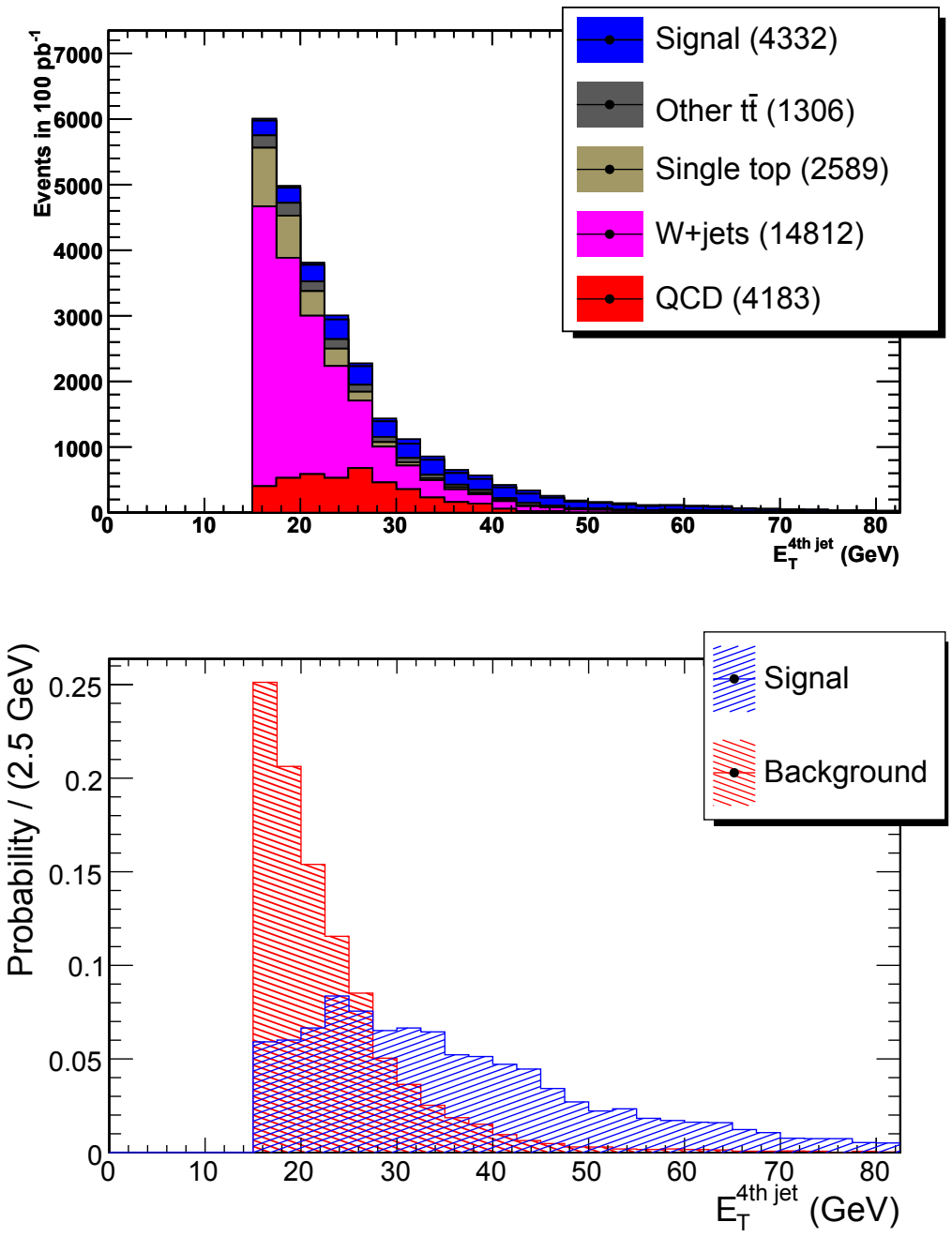


Figure 6.3: E_{T} distribution of the fourth leading jet for signal and background after the "loose" muon cuts with the additional requirement for at least four jets with $E_{\mathrm{T}} > 15$ GeV. Contributions from the signal and the different sources of background are stacked and displayed in different colors (top). The signal and total background distributions, each normalized to unit area, are displayed superimposed (bottom).

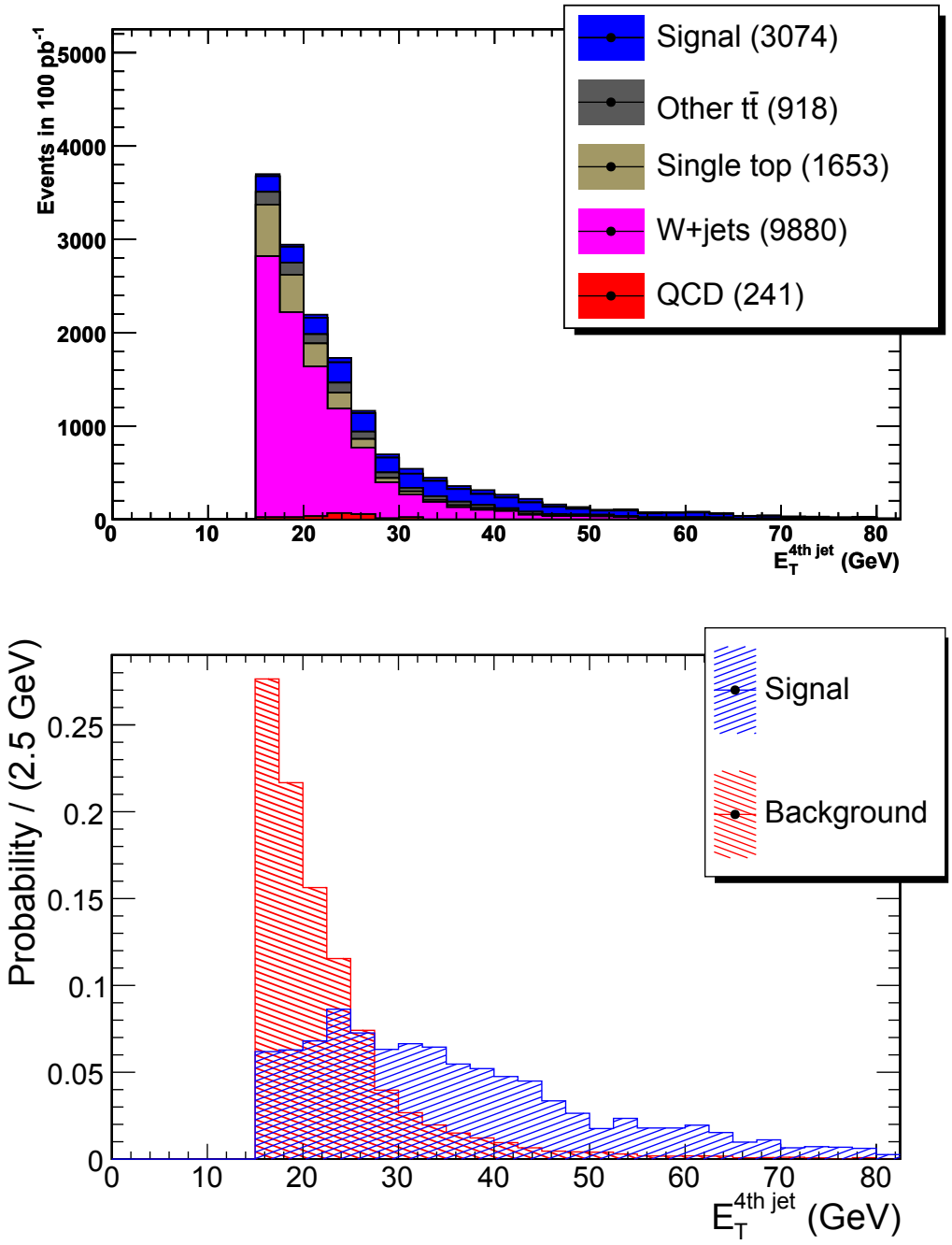


Figure 6.4: E_{T} distribution of the fourth leading jet for signal and background after the "tight" muon cuts with the additional requirement for at least four jets with $E_{\mathrm{T}} > 15$ GeV. Contributions from the signal and the different sources of background are stacked and displayed in different colors (top). The signal and total background distributions, each normalized to unit area, are displayed superimposed (bottom).

A commonly used quantity in particle physics is the *significance*. The significance is defined as the number of observed excess events, S, which we attribute to the signal process, divided by the uncertainty, δB, on the expected number of background events, B. Disregarding systematic uncertainties on the background for the moment, the statistical uncertainty on the background is $\sqrt{B}$. We may thus define a significance-like quantity, $\frac{S}{\sqrt{B}}$. This is a measure of how clear the indication of a signal is, when the method is limited to counting events.[4] We select the cut in such a way that this quantity is maximized. We plot this "significance" and the signal to background ratio versus the position of the cut for the "loose" and "tight" muon samples (Fig. 6.5 and Fig. 6.6). It is interesting to note that even though the signal to background ratio increases with a tighter cut, the benefit to be gained in terms of "significance" has a maximum. This reflects the fact that, after some point, the achievable background suppression does not any longer justify the losing any more signal by further tightening the cut. We decide on a requirement of 40 GeV on the fourth jet, as, at that point, we have already achieved the maximum "significance" and only the tail of the background distribution remains in our selected sample.

The breakdown of the selected sample (normalized to 100/pb integrated luminosity) after each of the cuts described so far is given in Tab. 6.1. It can be seen that a high purity sample has been identified. Particularly in the sample selected with the stricter muon isolation cut, QCD background has been completely eliminated and the purity reaches 64%, corresponding to a signal to background ratio of 1.8. The possibility of a looser muon selection is maintained as an option, in case we require a greater number of signal events in our sample and will be examined in the next chapter, as part of an effort to improve the cross-section measurement.

6.3 Jet-parton assignment

The selection described in the previous sections results in a high-purity $t\bar{t}$ sample. When working on Monte Carlo data this is very easy to verify, by checking with the generator-level information of each selected event. However, if this selection was applied on real data, we would only observe an excess in the number of events compared to expectations from the backgrounds. Attributing the observed excess to top-quark pair events, would have to somehow be justified. A convincing way to do that would be to plot the invariant mass of the three jets perceived to originate from the hadronically decaying top quark. We would expect this distribution to have a clear maximum around the mass of the top-quark.

In order to obtain this distribution, we first need to make an assignment of selected jets to quarks of the final state of the $t\bar{t}$ decay (see Fig. 5.1). The initial assumption that we make is that three of the four leading jets in the event will originate from the three quarks in the single-leptonic $t\bar{t}$ final state. The reason for this assumption is that, due to the boost and high mass of the top-quark, the jets it produces in its decay tend to be more energetic than the "unwanted" jets from initial and final state radiation. By selecting the four highest-E_{T} jets, we thus hope to select the three jets from the hadronically-decaying top-quark (we need to consider a fourth jet, because E_{T} does not distinguish the hadronic-side b-quark from its leptonically decaying counterpart). This fails if at least one of the following is true:

- One or more of the three jets corresponding to the hadronic-side partons are outside the considered η-acceptance. (12% of the events without an existing good Monte Carlo match [5] satisfy this condition).

[4] For a more detailed account of the significance and how it is calculated and used to claim a discovery the reader is referred to Ch. 7.

[5] We say that a jet "matches" a parton if the angle between the momentum of the jet and that of the (generator-level) parton in the $\eta - \phi$ space is smaller than 0.4. If the jet is assigned to a mathing parton we refer to it as a "matched" jet. If all three partons of the hadronic-side of the $t\bar{t}$ decay ($t \to bq\bar{q}'$) have a matching jet within the

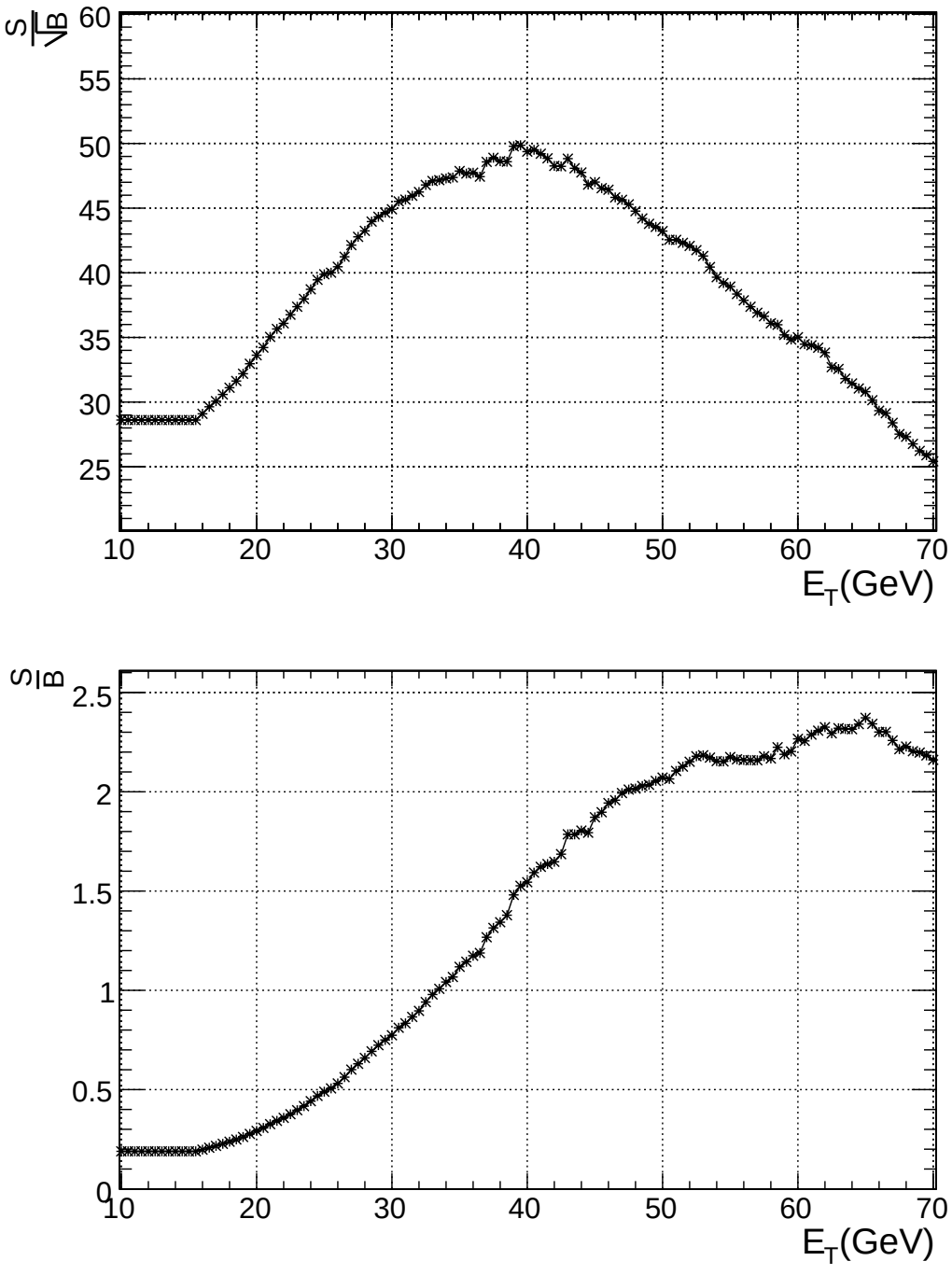


Figure 6.5: "Significance" (top) and signal to background ratio (bottom) as a function of the E_{T} cut on the four leading jets. The selection requires a "loose" muon.

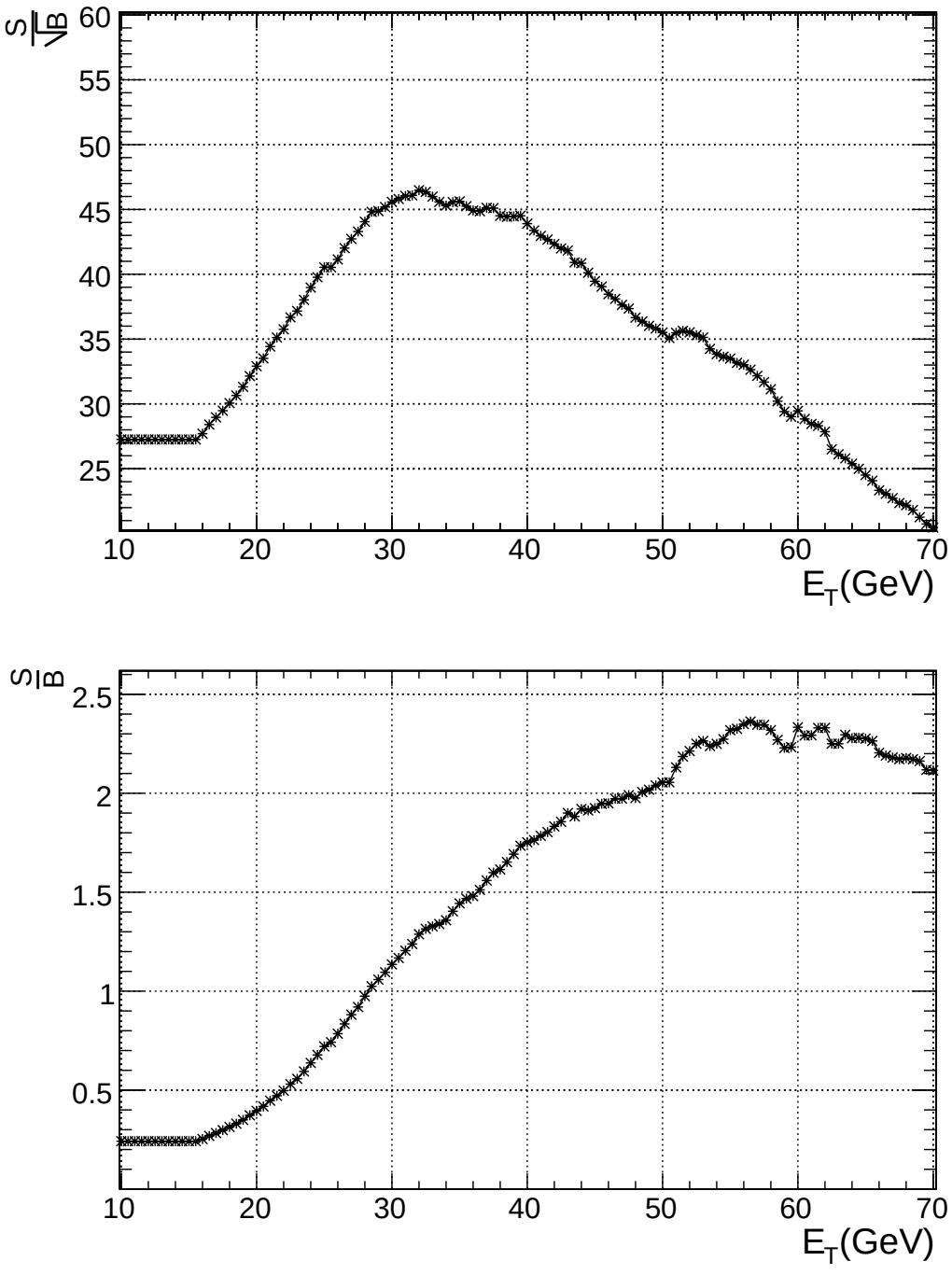


Figure 6.6: "Significance" (top) and signal to background ratio (bottom) as a function of the E_{T} cut on the four leading jets. The selection requires a "tight" muon.

- One or more of the three jets corresponding to the hadronic-side partons does not pass the jet E_{T} cut (86%). There is an overlap of 5% between this and the previous category.

- None of the previous two is true, but a high-E_{T} jet from another source has higher transverse energy than one or more of the jets from the hadronic-side top-quark partons. (7%)

We refer to those events as "unmatchable MC" events. But even for events for which none of the above is true and the four jets do include the three necessary for the top mass reconstruction, there is still a chance that we use the wrong fourth instead of one of the correct three. We refer to this as a wrong "solution". All these "unmatchable MC" and incorrectly "solved" constitute the "combinatorial background" for this analysis, which broadens the resulting top mass distribution and disguises our signal. The following subsections describe how we can try to reduce the impact of the combinatorial background, by either excluding some of the possible solutions from the start or selecting efficiently among the solutions we do consider.

6.3.1 Choosing the three hadronic-side jets

The transverse momentum spectrum of the top quarks produced in $t\bar{t}$ pairs peaks at approximately $m_t/2$ (Fig. 6.7(a)) and the top quark decay products will be boosted in its original direction. As a result of this, the angular separation of the three hadronic-side jets from each other will be smaller than from the leptonic-side b-jet. We therefore expect that:

- The vectorial sum of their momenta, will have a larger transverse component than that of any other jet combination. This can be confirmed by selecting MC-matching assignments and calculating

$$c_1 \equiv |\sum_{i=1}^{3} \vec{p}_{\mathrm{T}}| \tag{6.1}$$

  for all possible combinations:

  - A : hadronic-side b-jet + light jet 1 + light jet 2
  - B : hadronic-side b-jet + light jet 1 + leptonic-side b-jet
  - C : hadronic-side b-jet + light jet 2 + leptonic-side b-jet
  - D : leptonic-side b-jet + light jet 1 + light jet 2

  If our criterion indeed helps to assign the correct three jets to the hadronically decaying top-quark, then c_1^A will be higher than c_1^B, c_1^C, c_1^D for more than 25% of the events (which would be the probability of randomly selecting the correct jet assignment). A plot of $c_1^A - max(c_1^B, c_1^C, c_1^D)$ can be found in Fig. 6.8(a). A positive value signifies a case where the criterion would indeed point us to the correct assignment. Thus, the bins corresponding to values higher than zero divided by the total area of the histogram is the efficiency of our selection method. We find for the loose (tight) selection:

$$\varepsilon = 59.5\% \ (63.1\%)$$

- The sum of the angles between the leptonic-side b and the remaining jets,

$$c_2 \equiv \angle \vec{p}_{b,l}\vec{p}_q + \angle \vec{p}_{b,l}\vec{p}_{\bar{q}'} + \angle \vec{p}_{b,l}\vec{p}_{b,h}, \tag{6.2}$$

four leading jets of an event, we say that *a good Monte Carlo match exists.* Given a specific jet-parton assignment, if the three jets attributed to the hadronic side are matched to the corresponding partons (in other words, if the jet-parton assignment does not contradict the Monte Carlo truth) the event is refered to as a "matched" event.

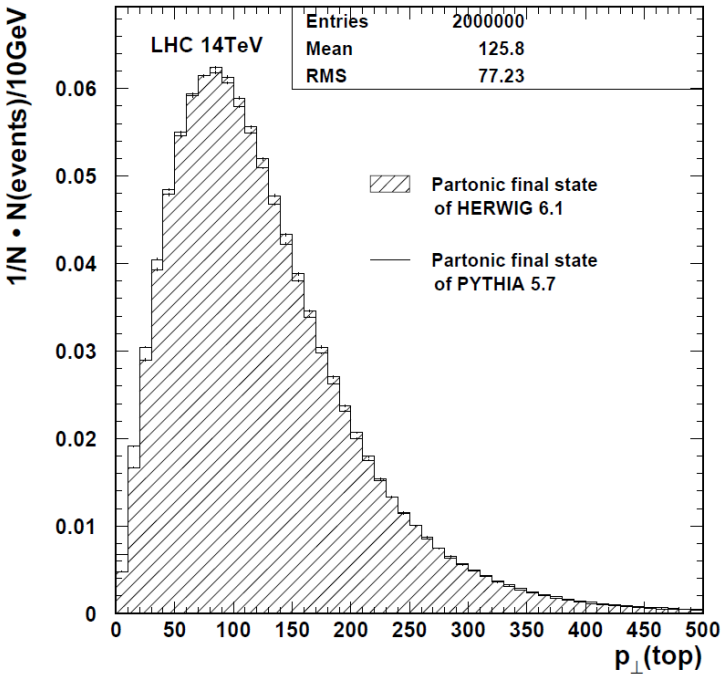


Figure 6.7: Transverse momentum spectrum of the top quark at the LHC, according to the Pythia and Herwig event generators. Taken from [44].

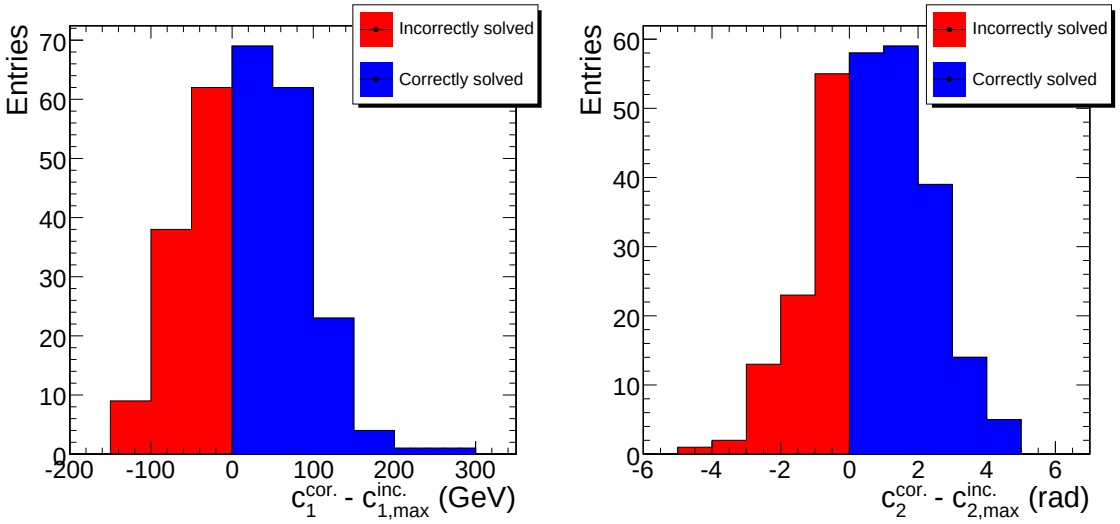


Figure 6.8: Difference of the value of the jet assignment criterion (Eq. 6.1,6.2) for the correct assignment from the maximum of the values for incorrect assignments $c^A - max(c^B, c^C, c^D)$. A positive value (blue area) indicates a case for which using the criterion results in a correct assignment. The selection here requires a "loose" muon.

will be larger than that between any of the other three jets and the remaining jets. Again, we calculate the variable for all possible combinations and plot $c_2^A - max(c_2^B, c_2^C, c_2^D)$ in Fig. 6.8(b). We find:

$$\varepsilon = 65.1\% \ (67.7\%)$$

In addition to the somewhat higher efficiency, this method has the advantage that it is not susceptible to jet miscalibration.

The top mass distributions resulting from the two methods, based on the selection described above, i.e. choosing the solution with largest c_1 or the largest c_2, are shown in Fig. 6.9 and 6.10.

6.3.2 Solution pruning

Some solutions can easily be excluded from the start, thus enhancing the success probability of the jet-parton assignment. We refer to the exclusion of possible solutions as "solution pruning". Applying solution pruning also has the desirable side-effect of improving the S/B, by sometimes cutting away all the possible solutions in a background event (and thus the event itself). However, these cuts are placed not to optimize their efficiency (in terms of signal to background or combinatorial background), but to make sure they only exclude a minimal amount of correct solutions. Therefore, their values only depend on the distributions obtained from "correct" solutions, not on the method by which a single event solution is picked among all possible ones. Three such possibilities have been investigated.

PTM pruning

It is possible to cut away some wrong solutions based on the $\sum_{i=1}^{3} p_{\mathrm{T}} - M_3$ of the three jets assigned to the hadronic side of the $t\bar{t}$ decay [45]. This is referred to as "PTM pruning". Figure 6.11 shows the MC matching solutions as well as the combinatorial background. The solutions corresponding to the points in the plot have been selected based on the assumption that the three hadronic-side jets will have the highest vectorial sum of transverse momenta (see section 6.3.1 for details). The shape of the background is obviously dependent on the jet-parton assignment method, therefore, this should not be taken as an optimization plot, just as an indication of PTM pruning being useful in excluding some wrong solutions and keeping the correct ones. We place it at $\sum_{i=1}^{3} p_{\mathrm{T}} - M_3 > -40\,\mathrm{GeV}$ (the dashed line in Fig.6.11).

Angle pruning

Another way to exclude wrong solutions is to cut on the angle between the two light jets, $\angle \vec{p_p}\vec{p_q}$, in the rest frame of the hadronically decaying top quark. In this reference frame, the W-boson is produced back to back with the b-jet and the light jets it decays into cannot come at any angle to each other. This angle, reconstructed for correct solutions is shown in Fig. 6.12(a), fit with with a Breit-Wigner convoluted with a Gaussian. According to this plot, we reject almost none of the good solutions by only keeping those corresponding to angles between 2.0 ± 0.6.

W mass pruning

The third way to exclude wrong assignments of jets to final state partons is by the invariant mass of the two light jets. Figure 6.12(b) shows the mass reconstructed from solutions that

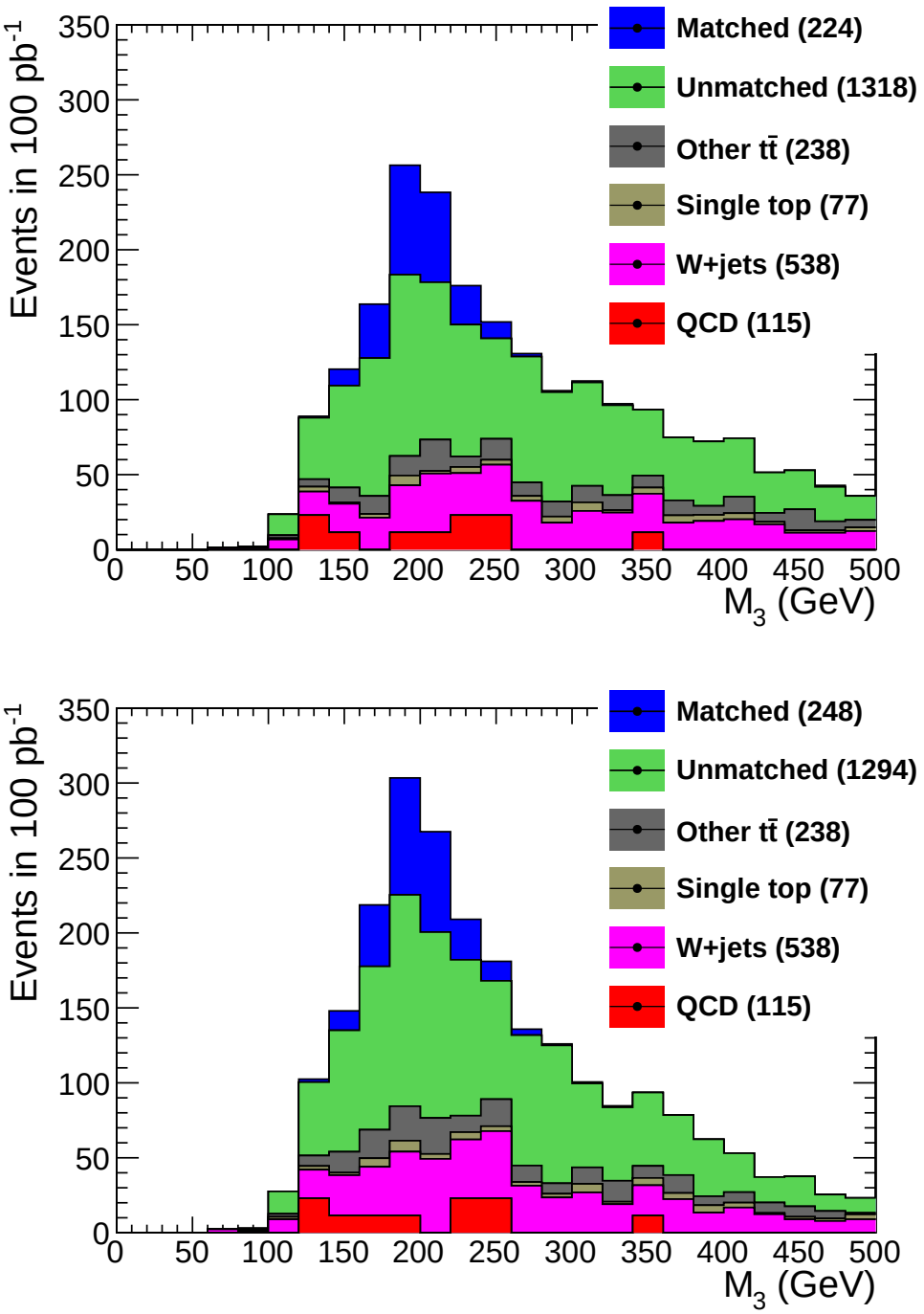


Figure 6.9: Invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system by the c_1 (top) and c_2 (bottom) criteria (defined in Sec 6.3.1). The selection requires a "loose" muon. The "matched" category corresponds to events for which the MC matching solution has been selected. The "unmatched" category corresponds to events for which a MC matching solution does not exist, or exists and has not been selected due to the inefficiency of the criterion used.

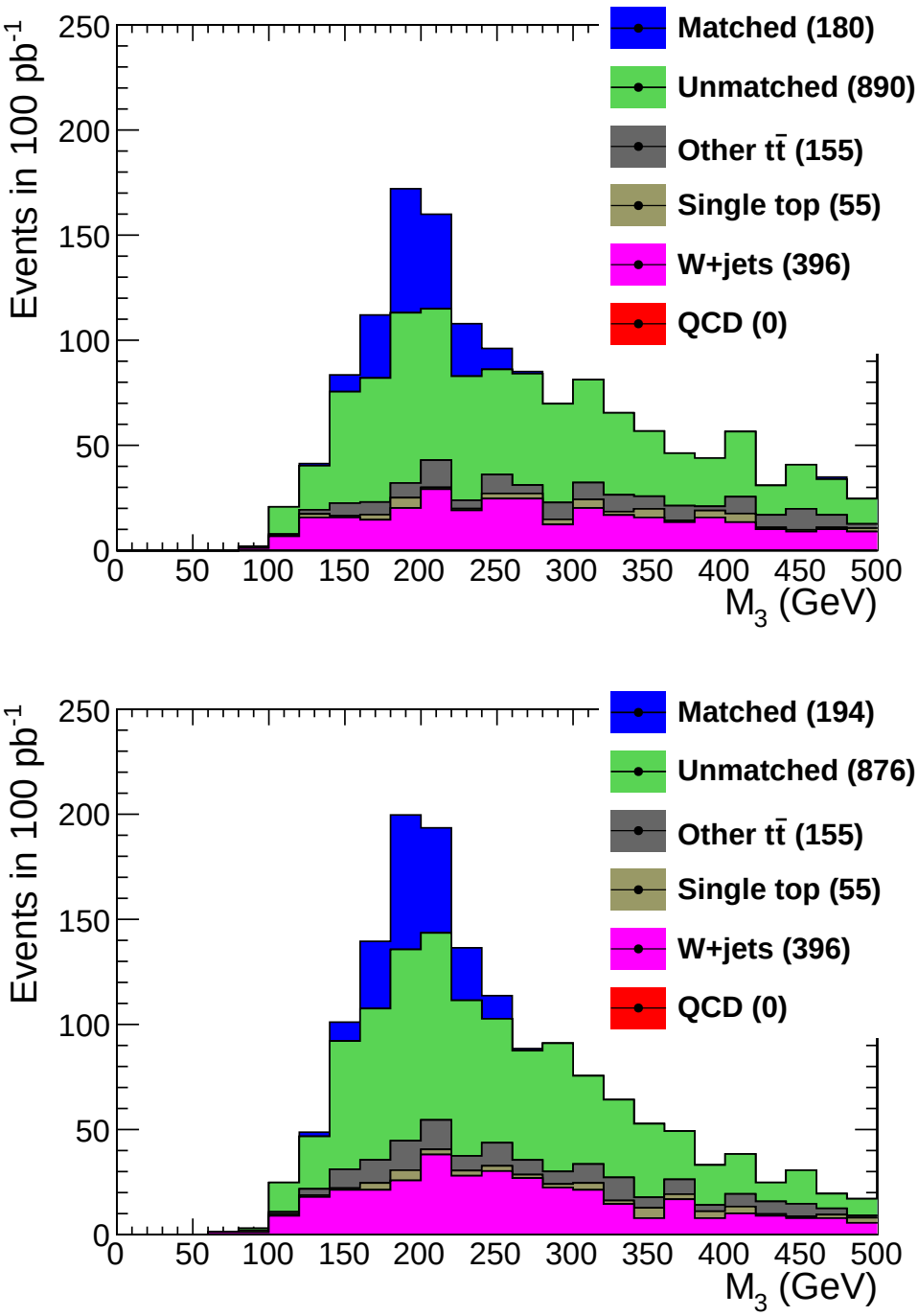


Figure 6.10: Invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system by the c_1 (top) and c_2 (bottom) criteria (defined in Sec 6.3.1). The selection requires a "tight" muon. The "matched" category corresponds to events for which the MC matching solution has been selected. The "unmatched" category corresponds to events for which a MC matching solution does not exist, or exists and has not been selected due to the inefficiency of the criterion used.

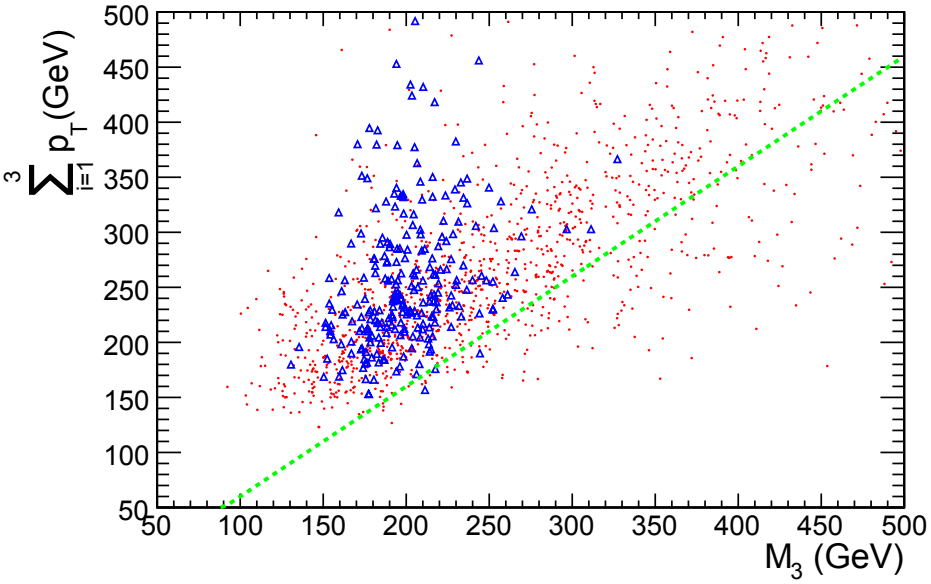
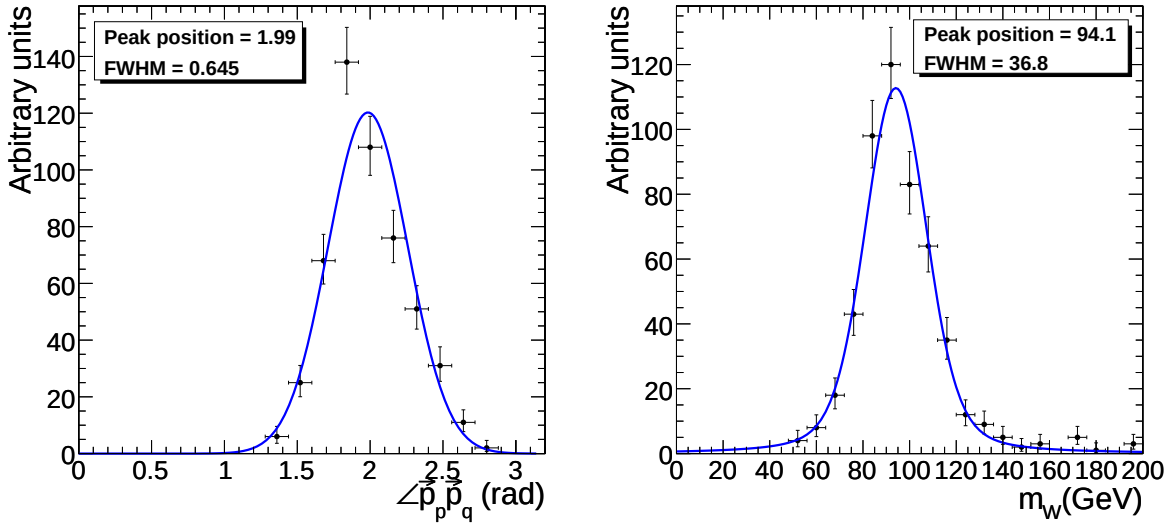


Figure 6.11: Algebraic sum of the p_{T} of the three jets assigned to the hadronic decay of the top-quark vs the 3-jet invariant mass, M_3, for $t\bar{t}$ events. Correct solutions are represented by the blue triangles and combinatorial background by the red dots. The signal-poor area below the dashed line is excluded by the PTM cut.



(a) Angle between the two light jets in the rest frame of the hadronically decaying top quark.

(b) Invariant mass of the two light jets.

Figure 6.12: The two variables used to cut away wrong jet-parton assignments. The solutions matched to the Monte Carlo truth have been used and a Breit-Wigner convoluted with a Gaussian is used for the fit.

Loose muon selection							
	c_1, Matched	c_1, Unmatched	c_2, Matched	c_2, Unmatched	B	$\frac{S}{B}$	$\frac{S}{\sqrt{B}}$
No pruning	224	1318	248	1294	967	1.59	49.6
PTM	263	1133	258	1138	797	1.72	49.4
$\angle\vec{p_p}\vec{p_q}$	239	1294	248	1076	957	1.60	49.5
m_W	221	1103	257	1276	754	1.78	48.2
All but m_W	269	1052	262	1059	714	1.76	49.4
All	259	815	268	806	510	2.10	47.5
Tight muon selection							
	c_1, Matched	c_1, Unmatched	c_2, Matched	c_2, Unmatched	B	$\frac{S}{B}$	$\frac{S}{\sqrt{B}}$
No pruning	180	890	194	876	606	1.77	43.5
PTM	209	755	202	762	496	1.94	43.3
$\angle\vec{p_p}\vec{p_q}$	191	872	195	733	597	1.78	43.5
m_W	178	750	200	863	451	2.06	43.7
All but m_W	211	708	204	715	448	2.05	43.4
All	212	543	203	552	293	2.57	44.1

Table 6.2: Number of signal and background events surviving different pruning methods (see text for details). The table is normalized to $100\,\mathrm{pb}^{-1}$ of integrated luminosity. The first two columns list the number of events selected after the c_1 criterion has been applied to select the three hadronic-side jets. The next two correspond to the c_2 criterion. The different criteria do not change the number of signal and background events selected, but lead to a different fraction of correct solutions for signal events and also respond differently to different solution pruning.

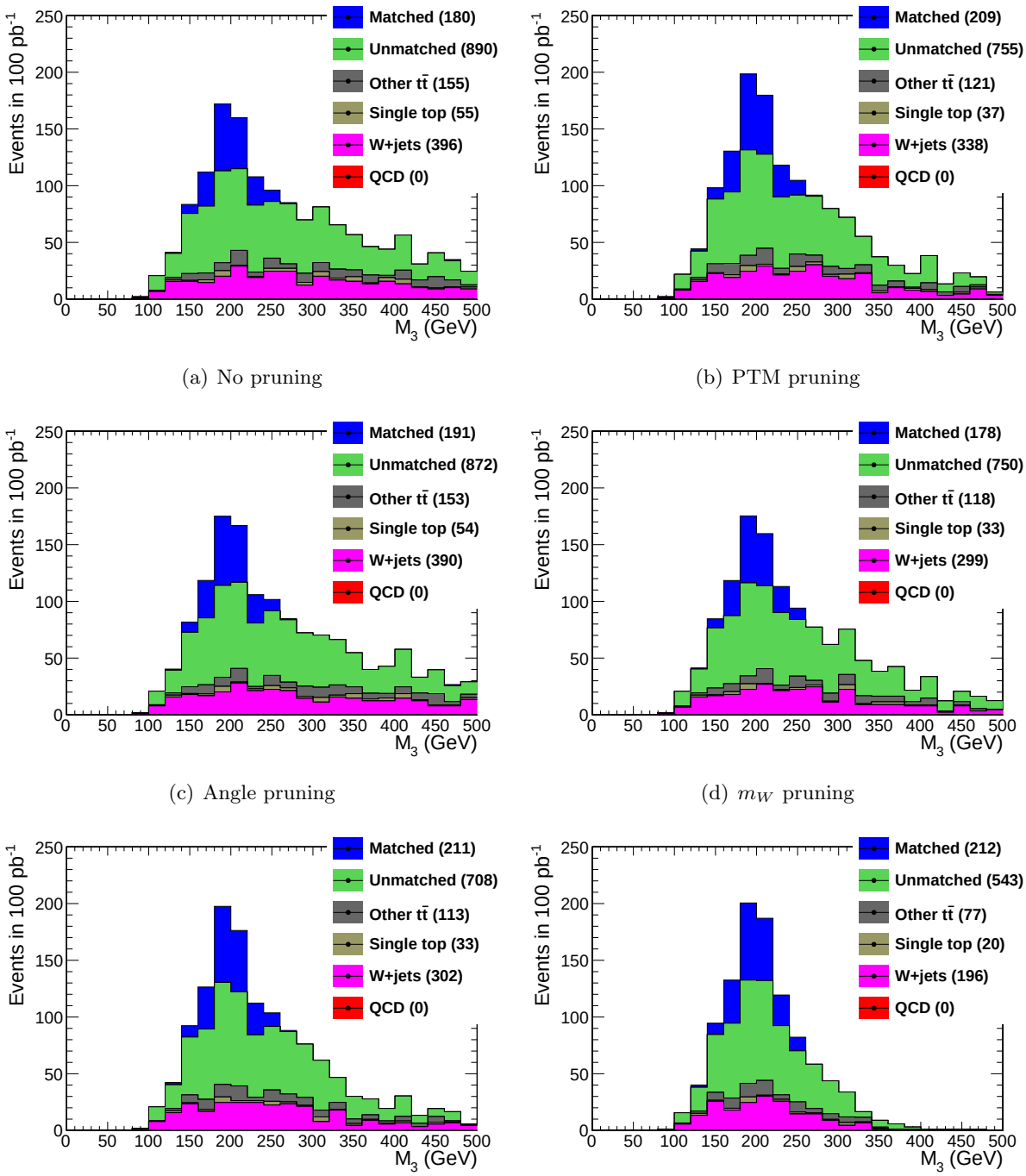
match to the Monte Carlo truth. Due to the imperfect jet calibration [6], the peak is not centered at the W-boson mass but shifted towards higher values. Based on this plot, we conclude that solutions resulting in invariant masses outside the region 94 ± 35 GeV can be safely rejected.

Effects and comparison of the solution pruning methods

The effects of each of the solution pruning methods on the three-jet invariant mass distribution can be seen in Fig. 6.13 and Tab. 6.2. Figure 6.13 shows the invariant mass of the three jets assigned to the hadronic side of the $t\bar{t}$ decay, resulting from the requirement for a tight muon and the use of the c_2 criterion (the invariant mass plots corresponding to different selection and jet-parton assignment strategies can be found in Appendix A). Each of the distributions presented corresponds to a different pruning strategy, from no pruning at all (Fig. 6.13(a)) to all pruning methods applied together (Fig. 6.13(f)). Table 6.2 summarizes the number of events selected as a result of applying each pruning method separately or in combinations. There are a number of useful observations to be made on these results.

- Predictably, even though solution pruning naturally reduces the total amount of signal, it almost invariably increases the number of signal events for which the jet-parton assignment has been successful ("matched" events). The number of "matched" events when using the c_1 criterion with the loose (tight) muon selection is enhanced by up to approximately 20% (18%). The c_2 criterion however only benefits by 8% (5%). This almost eliminates the difference in the performance of the two criteria discussed in Sec. 6.3.1. The effect of this improvement in the jet-parton assignment is reflected by the visible reduction in the tail

[6] The calibration used in this analysis has been derived from Monte Carlo studies of the CMS Jet and Missing E_{T} group. The jets used for these studies include gluon jets, which tend to result in lower energy response than that of light quark jets and therefore will have higher correction factors.



(a) No pruning

(b) PTM pruning

(c) Angle pruning

(d) m_W pruning

(e) PTM and angle pruning

(f) All three pruning methods applied

Figure 6.13: Invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, after pruning. The plots correspond to the tight selection and the c_1 criterion has been used to select the hadronic-side jets. The "matched" category corresponds to events for which the MC matching solution has been selected. The "unmatched" category corresponds to events for which a MC matching solution does not exist, or exists and has not been selected due to the inefficiency of the criterion used. The bottom right figure corresponds to the highest purity (72%).

of "unmatched" events that can be seen in Fig. 6.13(b), 6.13(c), 6.13(d), 6.13(e) and, most of all, 6.13(f) in comparison to Fig. 6.13(a). As the tail decreases, the invariant mass maximum becomes sharper, indicating a more efficient assignment.

- One of the main arguments against using solution pruning to reduce the combinatorial background would be that it could significantly reduce the total amount of signal by excluding "unmatchable MC" events with no correct solutions. In a "cut and count" rediscovery analysis that would only be acceptable if this signal reduction was accompanied by a sufficient reduction of the background. Indeed, Tab. 6.2 clearly shows how large the impact on the background is. The signal to background ratio substantially increases with solution pruning; and in such a way, that the significance-like quantity, $\frac{S}{\sqrt{B}}$, remains almost constant. The reduction of the background is also visible in Fig. 6.13. When all pruning methods are applied, the background tail is eliminated, further contributing to obtaining a sharper invariant mass peak.

- The m_W cut in particular is more effective in increasing the signal to background ratio than it is in enhancing the number of correct solutions. When the PTM and angle cut have already been applied, it is futile to try to increase the number of "matched" events by also applying m_W-based pruning. The result instead, is the exclusion of substantial numbers of "unmatched" and background events, without any important gain in significance.

- There is another good reason why leaving the m_W cut out of our selection could be useful. If we are interested in using the distribution of the invariant mass of the jets assigned to the W-boson (e.g. to correct our jet calibration for other analyses), it is important not to use the m_W cut, since that would severely bias the distribution and imply that we already know exactly where the W mass peak is. For this reason, an additional line corresponding to the application of all but the m_W cut has been included in the tables.

6.4 Conclusions

This concludes our discussion of the event selection. The goal of identifying a sample rich in $t\bar{t} \to bq\bar{q}'b\nu\mu$ has been achieved. The signal to background ratio resulting from the tight (loose) muon selection strategy is approximately 1.8 (1.6) corresponding to a purity of 64% (61%). "Pruning" cuts applied to reduce the combinatorial background not only maintain this signal to background ratio but further increase it up to 2.6 (2.1) corresponding to a purity of 72% (68%). The result is a clear maximum appearing in the invariant mass distribution of the three jets attributed to the hadronic side of the decay of the $t\bar{t}$ system. We summarize here the selection and jet-parton assignment strategy, which offers the best result in terms of purity:

1. Tight selection strategy:
   - At least one muon (isolated or non-isolated) of $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.1$, to fulfill the HLT acceptance.
   - Exactly one isolated muon ("loose" or "tight", as defined in 6.2.2), $|\eta| < 2.4$.
   - No second isolated lepton (as defined in 6.2.1), $|\eta| < 2.4$.
   - At least 4 jets of E_{T} >40 GeV.
2. Jet-parton assignment strategy:
   - PTM solution pruning: $\sum_{i=1}^{3} p_{\mathrm{T}} - M_3 > -40\,\mathrm{GeV}$.
   - Angle solution pruning: $\angle \vec{p_p}\vec{p_q} = 2.0 \pm 0.6$.

- W mass solution pruning: $m_W = 94 \pm 35$ GeV).
- c_1 criterion to select the hadronic side jets. In terms of purity, the choice between c_1 and c_2 does not make any difference, but c_1 results in a slightly better signal to combinatorial background ratio after all pruning cuts have been applied (39% as opposed to 37%).

This selection and jet-parton assignment strategy corresponds to the final line in Tab. 6.2 and to Fig.6.13(f). In the next chapter we will estimate how soon this analysis can give us evidence of $t\bar{t}$ events and how we can use our selection of events in order to measure the $t\bar{t}$ cross-section.

Chapter 7

t-Rediscovery and cross-section measurement

This chapter is divided into several parts. We first try to estimate the amount of data that CMS will need to collect in order to claim a "rediscovery" of the top quark signal. Next, we concentrate on a potential measurement of the cross-section using early data. A basic method is proposed and the expected uncertainties are calculated. This is followed by an evaluation of this method and an investigation of possible improvements that could help obtain a more accurate result and better understanding of the uncertainties.

7.1 Rediscovery of the top quark

In order to claim rediscovery of the top quark in the $t\bar{t} \to bq\bar{q}'b\nu\mu$ channel, we need a clear maximum in the region of the top quark mass in the three-jet invariant mass distribution and a significant excess of counted events over the expectation on the amount of background in a scenario without top quarks. Naturally, the more uncertain we are on our background expectation, the less significant an excess becomes. To express and treat the problem in more quantitative terms, the usual practice is to make the hypothesis that there is no signal and treat the number of measured events as a Poisson variable with a mean equal to the expected number of background events, ν_b. That way, supposing we perform the experiment and measure n_m events, with $n_m > \nu_b$, we can calculate the probability that this excess could be caused by just a statistical fluctuation of the background. This probability is commonly referred to as the P-value:

$$P(n \geq n_m) = 1 - \sum_{n=0}^{n_m-1} \frac{\nu_b}{n!} e^{-\nu_b}. \tag{7.1}$$

A very low P-value is a strong indication of the existence of an effect on top of the background. Instead of the P-value it is also common to refer to the *significance*, that is, the observed excess divided by the standard deviation of the background (we are approximating the Poissonian with a Gaussian),

$$S \equiv \frac{n_m - \nu_b}{\sqrt{\nu_b}} \tag{7.2}$$

and consider a significance of 3 as evidence and a significance of 5 as a clear observation of a signal.

We attempt a first prediction of the significance of the signal in our analysis by assuming that our number of measured events will be proportional to the integrated luminosity and the proportionality factor will be given by the number of signal events we have predicted in the

previous chapter, for an integrated luminosity of $100\,\mathrm{pb}^{-1}$:

$$n_m(\int_0^t Ldt) = \frac{n_m(100\,\mathrm{pb}^{-1})}{100\,\mathrm{pb}^{-1}} \times \int_0^t Ldt \tag{7.3}$$

The same assumption needs to be made for the expected number of background events.

$$\nu_b(\int_0^t Ldt) = \frac{\nu_b(100\,\mathrm{pb}^{-1})}{100\,\mathrm{pb}^{-1}} \times \int_0^t Ldt \tag{7.4}$$

This allows us to calculate the expected significance as a function of the integrated luminosity.

$$S = F\sqrt{\int_0^t Ldt} \tag{7.5}$$

Where we have defined:

$$F \equiv \frac{n_m(100\,\mathrm{pb}^{-1}) - \nu_\mathrm{b}(100\,\mathrm{pb}^{-1})}{\sqrt{100\mathrm{pb}^{-1}\nu_\mathrm{b}(100\,\mathrm{pb}^{-1})}} \tag{7.6}$$

The uncertainties need to be considered in two different contexts.

7.1.1 Systematic uncertainty on ν_b

.

It is common practice to treat the systematic uncertainties as if they were Gaussian in nature, and add them in quadrature to the statistical uncertainty due to Poissonian fluctuations in the number of background events. This leads to the following definition of the significance:

$$S \equiv \frac{n_m - \nu_b}{\sqrt{\nu_b + \delta\nu_b^2}} \tag{7.7}$$

Where $\delta\nu_b$ is the systematic uncertainty (which typically does not scale down when the integrated luminosity increases). Effectively, we are assuming that our estimate of the systematic uncertainty (which more often than not is an arbitrary, if educated, guess), has the effect of broadening the poisson distribution, changing its σ from $\sqrt{\nu_b}$ into $\sqrt{\nu_b + \delta\nu_b^2}$, without changing its expected value, ν_b.

With this approach, the significance is no longer proportional to the square root of the integrated luminosity as the relative statistical uncertainty becomes small, the significance approaches an asymptotic value set by the systematic uncertainty. Assuming a total relative systematic uncertainty, η_{sys}, we have:

$$S = \frac{n_m - \nu_b}{\sqrt{\nu_b + (\eta_{sys}\nu_b)^2}} \tag{7.8}$$

Defining:

$$\alpha \equiv \frac{n_m(100\,\mathrm{pb}^{-1})}{100\,\mathrm{pb}^{-1}}, \beta \equiv \frac{\nu_b(100\,\mathrm{pb}^{-1})}{100\,\mathrm{pb}^{-1}}, x \equiv \int_0^t Ldt \tag{7.9}$$

and substituting equations 7.3 and 7.4 into 7.8 we get:

$$S(x) = \frac{\alpha - \beta}{\sqrt{\beta + (\eta_{sys}\beta)^2 x}}\sqrt{x} \Longrightarrow \lim_{x\to\infty} S(x) = \frac{\alpha - \beta}{\eta_{sys}\beta} \neq \infty \tag{7.10}$$

To use this approach in the context of this analysis, it is necessary to determine exactly what to choose as the systematic uncertainty that we quadratically add to the standard deviation. We

list here the most important sources of systematic error.

- Uncertainties in the generator predictions for the cross-sections of the signal and background processes (we use the next to leading order Alpgen cross-sections given in Tab. 4.1.) These uncertainties can be considerable but are difficult to quantify. One way to do that is to produce the same samples using different generators and use the differences between generators as an estimate of the uncertainty. This would require a lot of storage space and CPU time and none of these was available for this work. However, the best solution is to avoid this uncertainty by estimating the background from data (see Sec. 9.1).

- Trigger and lepton reconstruction efficiency uncertainties. The trigger and reconstruction efficiencies can be taken from the detector simulation. It is easy to calculate the efficiency on a certain type of events by dividing the number of selected by the number of generated events of that type, but we cannot get the uncertainty on this number. Alternatively, there are also methods to estimate the efficiencies from data. Either way, the related uncertainty is expected to be very small in comparison to other systematic uncertainties considered in this analysis and can be safely ignored.

- The use of samples of limited statistics. This is a systematic uncertainty of statistical origin. For the "loose" muon selection for example, we expect we would select 2509 events (Tab. 6.2). This expectation really corresponds to an estimate of the selection efficiency, which we made using our $100\,\mathrm{pb}^{-1}$-equivalent of signal and W+jets events, our $122\,\mathrm{pb}^{-1}$-equivalent of t-channel single top events and our $8.7\,\mathrm{pb}^{-1}$-equivalent of QCD events. Using larger samples we would have arrived at a different and more accurate number of expected events for $100\,\mathrm{pb}^{-1}$ of expected luminosity. Our current estimate is only as good as the statistical uncertainty on our numbers of selected events implies. Treating the number of selected events for each contributing process as a Poisson variable, we can calculate this uncertainty. To continue with the example of the "loose" muon selection with no solution pruning, we have:

  - for the signal, 1542 ± 39 events,
  - for other $t\bar{t}$, 238 ± 15 events,
  - for single top, 77 ± 8 events,
  - for W+jets, 538 ± 25 events,
  - for QCD, 115 ± 36 QCD events.

  The statistical uncertainties on the different sources of background are added quadratically to calculate the uncertainty on the total background, so, in the end, we have a prediction of 1542 ± 39 events for the signal and 967 ± 47 events for the background or a relative systematic uncertainty of 2.5% and 5% for signal and background respectively.

- Other sources of systematic uncertainties. The need to estimate these systematic uncertainties was the motive for using the Fast Simulation in this analysis. As mentioned in Sec. 4.3.1, three additional datasets corresponding to different detector conditions were produced, which allow us to make an estimate on how much the amount of background can change due to these changes in conditions. Executing our selection over these samples, we find:

  - An uncertainty of $\pm 5\%$ on the background, due to the misalignment of the tracker and the miscalibration of the calorimeters expected at startup.
  - An uncertainty of $+15\%$ on the background, due to the possibility of an average of 5 pileup events per bunch crossing. This uncertainty only works in the direction

of increasing the background with regards to our expectation value that we derive from our reference sample. The reason is that as more energy is deposited in the calorimeters, more jets pass the E_{T} threshold and more background events satisfy the selection requirements.

– An uncertainty of +55% on the background, due to a possible over-scaling of the jet energy by 10%. This uncertainty only works in the direction of increasing the background with regards to our expectation value that we derive from our reference sample. The uncertainty due to a possible under-scaling of the jet energy scale is calculated to be -43%.

The errors given here still correspond to the "loose" muon selection without solution pruning, but their values for the different selection strategies proposed in Chapter 6 are in the same order of magnitude. Evidently, the contribution of 10% uncertainty in the jet energy scale makes up most of the systematic uncertainty of the expected background.

We can now add the uncertainties in quadrature:

$$\eta_{b,sys} = \frac{\delta\nu_b}{\nu_b} \approx^{+58\%}_{-36\%} \tag{7.11}$$

It is important to stress that the positive-sign uncertainty has a different consequence from the negative-sign uncertainty. We are worried about overestimating our background (negative-sign) because it would prevent us from claiming a discovery as soon as we could. On the other hand, the danger of underestimating the background is also serious, as it would lead to an erroneous claim. In the calculation of the significance we take the background to be $\nu_b(100\mathrm{pb}^{-1}) = 967$, and $\frac{\delta\nu_b}{\nu_b} = 58\%$ (the positive-side uncertainty) as the related uncertainty.

We can calculate the systematic uncertainty for any selection method the same way we did for the loose selection with no solution pruning. Table 7.1 shows the results for the tight and loose selection and different combinations of solution pruning methods. We also include the tight selection results for a jet energy scale uncertainty of 5% (instead of 10%) to demonstrate the impact of this source of systematic uncertainty on the measurement. Indeed, the achievable significance nearly doubles in this case.

7.1.2 Discovery potential

Using the analysis described in the previous section, we can attempt to predict the significance that we expect to observe for a given luminosity. The "expected" significance is calculated using the total number of events we expect to select under reference conditions (startup misalignment and miscalibration, no jet energy rescale, no pileup), ν_m, normalized to the appropriate integrated luminosity.

In addition to this "expected" scenario, it is useful to consider the effect of a 1σ systematic error on ν_m. We can calculate the *systematic bias* on the total number of selected events in the same way we calculated the systematic uncertainty on the number of background events. Keeping to the example of the loose selection with no solution pruning, we find:

$$\nu_m(100\ \mathrm{pb}^{-1}) = 2509 \tag{7.12}$$

$$\eta_{m,sys} = \frac{\delta\nu_m}{\nu_m} \approx^{+36\%}_{-26\%} \tag{7.13}$$

It is important to stress that $\eta_{m,sys}$ is not the uncertainty on a measured number of events, n_m - if we measure n_m events, our result is not $n_m{}^{+36\%}_{-26\%}$. Instead, it is the uncertainty on the number of events we expect to measure, ν_m.

We thus define an "optimistic" ($\nu_m + 36\% \times \nu_m$) and a "pessimistic" ($\nu_m - 26\% \times \nu_m$) scenario. In the "optimistic" scenario the systematic effects worked towards increasing the number of events and in the "pessimistic" scenario towards decreasing it, leading to a higher and a lower significance respectively.

For any given integrated luminosity, we thus can:

1. Find our expected number of measured events, ν_m by normalizing the reference sample to the integrated luminosity.

2. Define an optimistic ($\nu_{m,opt}$) and a pessimistic ($\nu_{m,pes}$) scenario corresponding to a 1σ systematic error for each of the two sides of ν_m.

3. Calculate an optimistic and a pessimistic significance using $\nu_{m,opt}$ and $\nu_{m,pes}$ instead of n_m in Eq. 7.8.

Repeating these steps for many different values of integrated luminosity, we can draw the optimistic, expected and pessimistic scenario curves versus the integrated luminosity. This was done for all selection and jet-parton assignment strategies presented in the previous chapter. The full results can be found in Appendix A. Table 7.1 shows the systematic uncertainties on ν_b, ν_m and the maximum obtainable significance for the "expected scenario".

We do include here the significance curves corresponding to the tight selection with no pruning cuts (Fig. 7.1(a)), with angle and PTM pruning (Fig. 7.1(c)) and with all three pruning methods applied (Fig. 7.1(b)). We see that placing the "tight" selection requirements on the muon without any pruning can only lead to a "rediscovery" ($S > 5\sigma$) in an optimistic scenario. The expected scenario only leads to a maximum significance, $\lim_{x\to\infty} S(x)$, of 4.19σ (Tab. 7.1). However, if all three pruning methods are applied, the expected scenario significance reaches 5σ at approximately 7pb^{-1} of integrated luminosity (Fig. 7.1(b)) and the maximum significance becomes 5.68σ. It might be desirable to avoid using W mass pruning, so as to not bias the W mass distribution. Figure 7.1(c) shows us that a rediscovery is still possible if only PTM and angle pruning are applied, as $\lim_{x\to\infty} S(x) = 5.0$. It is interesting to note how much the jet energy scale uncertainty affects the significance. Figure 7.1(d) shows what the significance curves would be if the jet energy scale uncertainty was reduced to its half (5%). Even without pruning, it is only in a pessimistic scenario that the rediscovery is not guaranteed - the expected scenario maximum significance is 7.62. The third subtable of Tab. 7.1 shows how significant the signal could be expected to be for all selection strategies, if the jet energy scale uncertainty was halved.

7.1.3 Conclusions

In this section we have seen how we can evaluate our signal to determine if it is significant and thus claim a "rediscovery" of the top quark. We also investigated how likely it is to achieve the rediscovery using the selection strategies covered in the previous chapter, considering the systematic effects we expect to be important.

We conclude that by placing the "tight" selection requirements on the muon and by "pruning" the solutions using all the methods described in Sec. 6.3.2, we can expect to get a significant signal (Fig. 7.1(b)). Getting a 5σ significance is also a possibility with any of the "tight muon" strategies examined. A notable case is the tight selection with PTM and angle pruning (Fig. 7.1(c)). Evidently, it is possible to get a significant signal even without the use of m_W pruning.

On the other hand, as we can see from the "pessimistic" scenario curves, there is also a considerable chance that we will not be able to ensure the rediscovery, if the systematic effects work towards reducing the number of events. This can be excluded by reducing the systematic uncertainties on the background. The jet energy scale has been shown to be by far the most important uncertainty considered (see Fig. 7.1(d) and the third subtable of Tab. 7.1). There is

however no easy way to control the jet energy scale uncertainty before isolating a $t\bar{t}$ sample with hadronic W decays. The best way to address this issue is thus to incorporate the systematic effect into our estimate of the background, by obtaining it from data. More on this can be found in Sec. 9.1.2. An alternative is to rely on the MC simulation to obtain the shape of the distribution of an observable quantity for signal and background. These shapes can then be normalized to fit the distribution obtained from the selected dataset. In this way, our MC estimate is still subject to the systematic uncertainty but using real data as input reduces the effect this uncertainty has on the measurement. This possibility is investigated in Sec. 8, in the context of a cross-section measurement.

7.2 Cross-section measurement

After observing the $t\bar{t}$ signal, the next goal is to estimate the cross-section. The cross-section can be extracted from the number of observed events using the formula:

$$\sigma \times BR \equiv \frac{n_m - \nu_b}{\varepsilon \times \int Ldt} \tag{7.14}$$

Where BR is the branching ratio of the single-muonic decay (12/81, see Tab. 1.1) and ε is the total efficiency for signal events, calculated using the Monte Carlo simulation. When performing our counting experiment, we can use this formula to obtain the cross-section. The question we need to answer now is what uncertainty we can expect on this result.

7.2.1 Statistical uncertainty

The relative statistical uncertainty is easily calculated as:

$$\frac{(\delta\sigma)_{st}}{\sigma} = \frac{\frac{\partial\sigma}{\partial n_m} \times (\delta n_m)_{st}}{\sigma} = \frac{\sqrt{n_m}}{\varepsilon \times \int Ldt} \times \frac{\varepsilon \times \int Ldt}{n_m - \nu_b} = \frac{\sqrt{n_m}}{n_m - \nu_b} \tag{7.15}$$

And by the definitions of Sec. 7.1.1:

$$\frac{(\delta\sigma)_{st}}{\sigma} = \frac{\sqrt{\alpha}}{\alpha - \beta} \times x^{-\frac{1}{2}} \tag{7.16}$$

We can use the "expected scenario" to calculate α and β and obtain the expected statistical uncertainty on our cross-section measurement for any integrated luminosity, x.

7.2.2 Systematic uncertainty

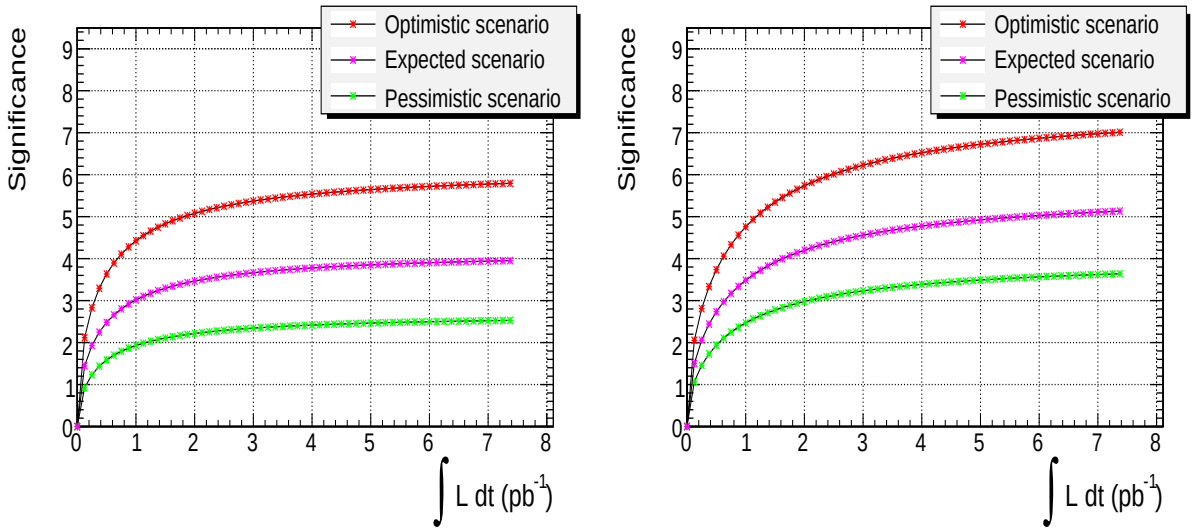
An exhaustive estimation of the systematic uncertainties is more difficult for the cross-section measurement than for the rediscovery, as the cross-section formula also includes the efficiency and the integrated luminosity. We will work out the tight case with the PTM and angle pruning applied as an example.

In our measurement we will use the efficiency predicted by the Monte Carlo for the reference sample (startup scenario), no matter how many events, n_m, we select.

$$\varepsilon = \frac{\nu_{s,selected}}{\nu_{s,generated}} = \frac{919}{83650} \approx 0.011 \tag{7.17}$$

We will also use the expected number of background events predicted by the simulation for the reference sample.

$$\nu_b = 448 \tag{7.18}$$



(a) No pruning

(b) All three pruning methods applied.

(c) PTM and angle pruning

(d) No pruning, 5% JES.

Figure 7.1: Significance as a function of integrated luminosity for different selection strategies. If solution pruning (see text, Sec. 6.3.2) is not applied (top left), a 5σ significance is only possible in an optimistic scenario, where the systematic effects work towards amplifying the observed excess. The application of all three solution pruning methods (top right) offers the best "rediscovery" potential. However, it is also possible to have a significant signal without using W mass pruning (bottom left). Pruning would not be required at all, if it were possible to reduce the jet energy scale (JES) uncertainty from 10% to 5% (bottom right).

Loose muon selection			
	ν_b	ν_m	$\lim_{x\to\infty} S(x)$
No pruning	$967^{+58\%}_{-36\%}$	$2509^{+36\%}_{-26\%}$	2.77
PTM	$797^{+54\%}_{-35\%}$	$2193^{+33\%}_{-25\%}$	3.24
$\angle\vec{p_p}\vec{p_q}$	$957^{+58\%}_{-36\%}$	$2281^{+36\%}_{-26\%}$	2.76
m_W	$754^{+61\%}_{-37\%}$	$2287^{+33\%}_{-24\%}$	2.87
All but m_W	$714^{+55\%}_{-35\%}$	$2035^{+33\%}_{-25\%}$	3.37
All	$510^{+61\%}_{-38\%}$	$1584^{+31\%}_{-24\%}$	3.46
Tight muon selection			
	ν_b	ν_m	$\lim_{x\to\infty} S(x)$
No pruning	$606^{+42\%}_{-29\%}$	$1676^{+30\%}_{-23\%}$	4.19
PTM	$496^{+40\%}_{-28\%}$	$1460^{+28\%}_{-22\%}$	4.83
$\angle\vec{p_p}\vec{p_q}$	$597^{+43\%}_{-29\%}$	$1525^{+30\%}_{-23\%}$	4.17
m_W	$451^{+45\%}_{-29\%}$	$1514^{+26\%}_{-21\%}$	4.59
All but m_W	$448^{+41\%}_{-29\%}$	$1367^{+29\%}_{-22\%}$	5.00
All	$293^{+45\%}_{-30\%}$	$1048^{+26\%}_{-21\%}$	5.68
Tight muon selection, 5% E_j scale uncertainty			
	ν_b	ν_m	$\lim_{x\to\infty} S(x)$
No pruning	$606^{+23\%}_{-15\%}$	$1676^{+15\%}_{-12\%}$	7.62
PTM	$496^{+22\%}_{-14\%}$	$1460^{+14\%}_{-11\%}$	8.93
$\angle\vec{p_p}\vec{p_q}$	$597^{+23\%}_{-15\%}$	$1525^{+15\%}_{-12\%}$	7.59
m_W	$451^{+27\%}_{-15\%}$	$1514^{+14\%}_{-11\%}$	7.58
All but m_W	$448^{+22\%}_{-14\%}$	$1367^{+13\%}_{-11\%}$	9.28
All	$293^{+26\%}_{-15\%}$	$1048^{+15\%}_{-11\%}$	9.94

Table 7.1: Systematic uncertainties on the background expectation (first column), uncertainty on the expected number of events (second column) and maximum significance that can be obtained by increasing the integrated luminosity for the "expected scenario" corresponding to startup detector conditions (third column). The top and middle subtables correspond to the loose and tight selection strategy respectively. The best "rediscovery" potential is offered by the tight muon selection with all pruning methods applied (see text, Sec. 6.3.2). The bottom subtable corresponds to a case of reduced jet energy scale uncertainty (5% instead of 10%).

If the detector conditions are exactly as we expect them, we would be able to measure the cross-section without systematic error. In reality, however, they will be quite different and this will affect the number of events measured, n_m. How can this affect the measured cross-section? Suppose a systematic effect causes us to measure n'_m instead of n_m. This would result in us calculating a cross-section, σ' quite different from the one we would calculate in the absence of that effect. The relative uncertainty we attach to that result should therefore be:

$$\eta_\sigma = \frac{\sigma - \sigma'}{\sigma'} = \frac{n_m - \nu_b - (n'_m - \nu_b)}{n'_m - \nu_b} \tag{7.19}$$

We now go through the systematic effects and calculate the related uncertainties.

- Due to the miscalibration and misalignment of the detector, we have an uncertainty of $\eta_{\sigma,startup} \approx \pm 1.3\%$.
- Due to the possibility of having an average of 5 pileup events per bunch crossing, $\eta_{\sigma,pileup} \approx -0.8\%$.
- Due to the possibility of over-scaling the jet energy scale by 10%, $\eta_{\sigma,+10\%} \approx -30\%$.
- Due to the possibility of under-scaling the jet energy scale factor by 10%, $\eta_{\sigma,-10\%} \approx +42\%$.
- Due to a 10% uncertainty on the integrated luminosity, $\eta_{\sigma,lumi} = \pm 10\%$.

Adding the systematic uncertainties quadratically, we get:

$$\eta_{sys,-} = -\sqrt{\eta^2_{\sigma,startup} + \eta^2_{\sigma,pileup} + \eta^2_{\sigma,-10\%} + \eta^2_{\sigma,lumi}} = -32\% \tag{7.20}$$

$$\eta_{sys,+} = +\sqrt{\eta^2_{\sigma,startup} + \eta^2_{\sigma,+10\%} + \eta^2_{\sigma,lumi}} = +44\% \tag{7.21}$$

Now that the systematic uncertainties are estimated we can calculate:

$$\alpha = \frac{n_m(100\text{ pb}^{-1})}{100\text{ pb}^{-1}} = 13.67, \beta = \frac{\nu_b(100\text{ pb}^{-1})}{100\text{ pb}^{-1}} = 4.48 \tag{7.22}$$

And for any integrated luminosity x (in pb^{-1}), the uncertainties are:

$$\frac{\delta\sigma}{\sigma} = \pm\frac{\sqrt{\alpha}}{\alpha - \beta} \times x^{-\frac{1}{2}}(\text{stat.})^{+44\%}_{-32\%}(\text{sys.}) \approx \pm 0.40 \times \text{x}^{-\frac{1}{2}}(\text{stat.})^{+44\%}_{-32\%}(\text{sys.}) \tag{7.23}$$

This calculation has been repeated for all different selection strategies. The results can be found in Tab. 7.2.

7.2.3 Conclusions

The conclusion drawn from Tab. 7.2 is that the accuracy of the event counting method will not be limited by the available statistics - even 10 pb^{-1} is enough to reduce the statistical uncertainty to less than 15%. In that sense, the best selection strategy for this cross-section measurement method is the tight muon selection with all pruning methods applied, because even though it has the highest statistical uncertainty (14%) it is the least prone to systematic effects ($^{+38\%}_{-29\%}$). The dominating source of statistical uncertainty is the uncertainty on the jet energy scale, which is in all cases more than 29%. It is therefore essential to either come up with an alternative method of measuring the cross-section, which will be less prone to the jet energy scale, or find a way to estimate the number of background events from data, as systematic effects such as pileup or jet energy miscalibrations are naturally accounted for in such an estimate.

Loose muon selection							
	$\frac{\delta\sigma_{stat}}{\sigma} \times x^{\frac{1}{2}}$	$\frac{\delta\sigma_{stat}(10\text{pb}^{-1})}{\sigma}$	$\frac{\delta\sigma_{startup}}{\sigma}$	$\frac{\delta\sigma_{pileup}}{\sigma}$	$\frac{\delta\sigma_{+10\%}}{\sigma}$	$\frac{\delta\sigma_{-10\%}}{\sigma}$	$\frac{\delta\sigma_{sys.}}{\sigma}$
No pruning	0.32	10%	±4%	−10%	−36%	+57%	$^{+58\%}_{-39\%}$
PTM	0.34	11%	±5%	−8%	−34%	+51%	$^{+52\%}_{-37\%}$
$\angle\vec{p_p}\vec{p_q}$	0.33	10%	±4%	−11%	−36%	+57%	$^{+58\%}_{-39\%}$
m_W	0.34	11%	±6%	−12%	−33%	+50%	$^{+51\%}_{-37\%}$
All but m_W	0.34	11%	±6%	−9%	−33%	+49%	$^{+51\%}_{-36\%}$
All	0.37	12%	±7%	−8%	−31%	+45%	$^{+47\%}_{-34\%}$
Tight muon selection							
	$\frac{\delta\sigma_{stat}}{\sigma} \times x^{\frac{1}{2}}$	$\frac{\delta\sigma_{stat}(10\text{pb}^{-1})}{\sigma}$	$\frac{\delta\sigma_{startup}}{\sigma}$	$\frac{\delta\sigma_{pileup}}{\sigma}$	$\frac{\delta\sigma_{+10\%}}{\sigma}$	$\frac{\delta\sigma_{-10\%}}{\sigma}$	$\frac{\delta\sigma_{sys.}}{\sigma}$
No pruning	0.38	12%	±2.0%	−1.3%	−32%	+46%	$^{+48\%}_{-34\%}$
PTM	0.40	13%	±1.6%	−1.0%	−30%	+43%	$^{+44\%}_{-32\%}$
$\angle\vec{p_p}\vec{p_q}$	0.38	12%	±2.3%	−1.6%	−32%	+47%	$^{+48\%}_{-34\%}$
m_W	0.40	13%	±2.4%	−2.7%	−28%	+39%	$^{+40\%}_{-30\%}$
All but m_W	0.40	13%	±1.3%	−1.0%	−30%	+42%	$^{+44\%}_{-32\%}$
All	0.43	14%	±1.0%	−1.6%	−27%	+36%	$^{+38\%}_{-29\%}$

Table 7.2: Systematic uncertainties on the cross-section. The first column holds the value of $\frac{\sqrt{\alpha}}{\alpha-\beta}$ and the second the relative statistical uncertainty for an integrated luminosity of 10pb^{-1}. The uncertainties due to different systematic effects (misalignment & miscalibration, pileup, jet energy scale overestimation and underestimation by 10%) occupy the next four columns. The total systematic uncertainty can be found in the last column.

Chapter 8

Cross-section measurement with the maximum likelihood method

The basic method of measuring the $t\bar{t}$ production cross-section by counting the number of observed events and subtracting the expected background is straightforward and simple. However, as explained in the previous section, without a way of measuring the background from data, we have to rely on the MC prediction on the background cross-section. This induces an uncertainty which is difficult to quantify. In this section, we present an alternative measurement, which uses the Maximum Likelihood (ML) method to remove the uncertainty of the MC prediction on the total background cross-section.

In order to use the method of maximum likelihood for the purpose of the cross-section measurement, we formulate the problem as follows [1]. Each of the n events passing the selection cuts is treated as an independent measurement of a variable x, which can be any measurable quantity (e.g. the jet multiplicity). We can obtain the signal and background PDFs, $f_s(x)$ and $f_b(x)$, from the Monte Carlo simulation but, in order to get the total PDF, we need to sum the two in the correct proportion. We thus write:

$$f(x) = \frac{\nu_s}{\nu_s + \nu_b} f_s(x) + \frac{\nu_b}{\nu_s + \nu_b} f_b(x). \tag{8.1}$$

Where the number of signal and background events are both Poisson variables and ν_s, ν_b are their mean values that we need to estimate. Since our data is binned, the extended log-likelihood function that we have to maximize is:

$$\log L(\nu_s, \nu_b) = -\nu_s - \nu_b + \sum_{i=1}^{N} \log(\nu_s \nu_{s,i} + \nu_b \nu_{b,i}). \tag{8.2}$$

The sum is over the N bins and $\nu_{s,i}, \nu_{b,i}$ are the expected signal and background entries in bin i:

$$\nu_{s/b,i} \equiv \nu_{s/b} \int_{x_i^{min}}^{x_i^{max}} f_{s/b}(x) dx. \tag{8.3}$$

The advantage of the maximum likelihood approach is that we do not have to rely on the Monte Carlo prediction for the background cross-section to obtain the number of signal events, as we did for the event counting method. We only have to know the shape of the background and signal distributions to make an estimate of ν_s and ν_b.

The accuracy of this estimate strongly depends on the choice of the variable x - the larger the difference between the shapes of the signal and background distributions, the easier it is

[1] For a more detailed explanation of the Maximum Likelihood method and the technical details related to the fits presented in this chapter the reader is referred to Appendix B.

to determine the signal and background proportions from a finite data sample. We examine two discriminating variables. The first is the invariant mass of the three jets assigned to the hadronically decaying top quark ($t \to bq\bar{q}'$), M_3, and the second is the jet multiplicity.

8.1 M_3 fit

The first variable to examine is the invariant mass of the three jets assigned to the hadronic side of the decay. This is a natural choice, as a clear maximum in the region of the top mass is expected for signal but not for background events. The distribution of the variable for all events in the selected sample as well as the signal ($f_s(M_3)$) and background ($f_b(M_3)$) PDFs for an example selection strategy can be found on Fig. 8.1. Indeed, $f_s(M_3)$, displays the expected maximum.

Having obtained the signal and background PDFs from the full MC sample, we can produce an example fit to 10 pb^{-1} of data. We first sample the signal and background distributions to create a "data" sample corresponding to 10 pb^{-1}. The number of signal and background events in the "data" are allowed to vary following a Poisson distribution around the expected value as estimated from the Monte Carlo. We then perform an extended binned ML fit of the signal and background PDFs to this "data". As an interface to the MINUIT [46] minimization library we use the RooFit toolkit [47] within the ROOT data analysis framework [48], which helps automate the necessary tasks. The fit returns an estimate of the number of signal events and of the error on this number. Such an example fit can be found on Fig. 8.2 [2]. It is accompanied by the distributions (obtained from one thousand similar MC experiments) of:

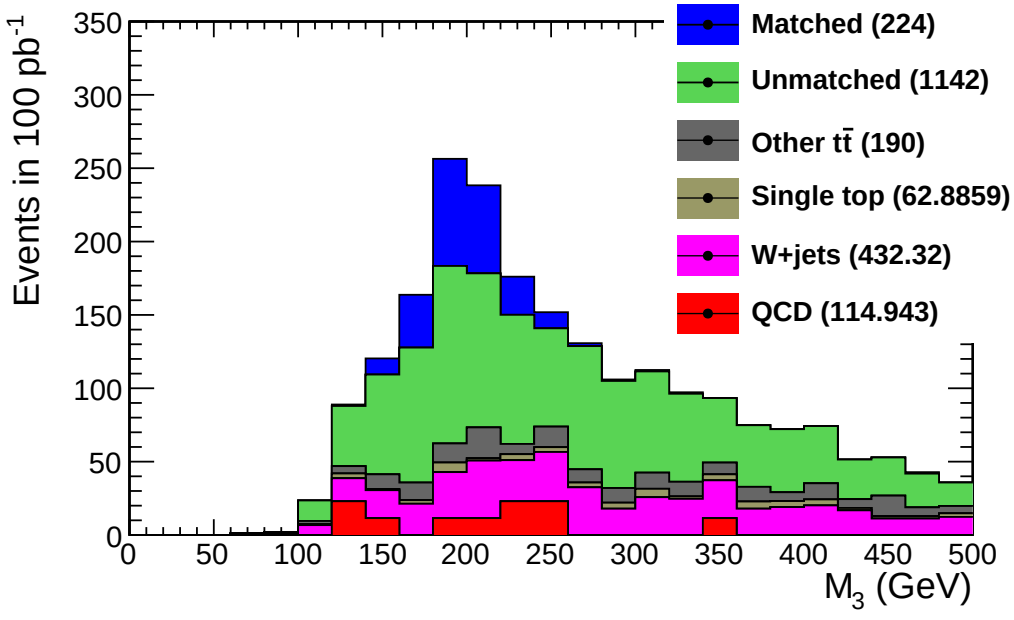
- The number of signal events estimated from the fit, n_s.
- The fit error (as calculated by MINUIT), δn_s.
- The *pull*, defined as the difference of a fit result, n_s, from ν_s divided by the fit error returned by MINUIT.

This MC study not only estimates the uncertainty of our result by the distribution of the fit result and the MINUIT error, but also confirms that our fit is stable and unbiased. We find that n_s is distributed around the expected value in an approximately Gaussian shape and that the standard deviation of this distribution is consistent with the mean MINUIT error. Furthermore, the pull distribution has a mean of approximately zero and a standard deviation of one, indicating that the fit is not systematically biased towards overestimating or underestimating the result and confirming the credibility of the error estimate returned by MINUIT.

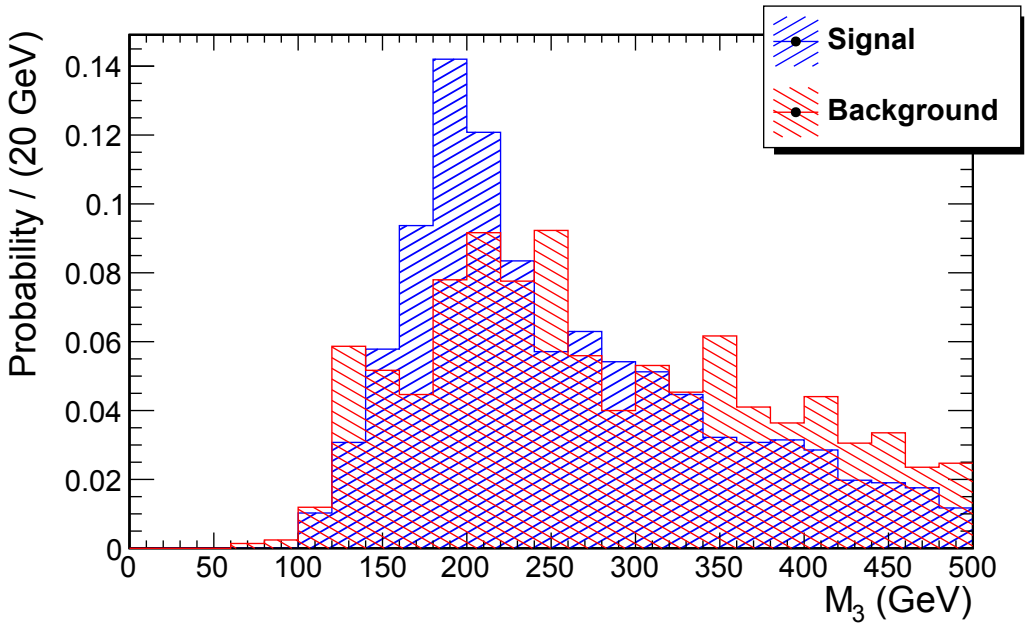
8.2 Systematic effects of the M_3 fit

Having established the good behaviour of the ML fit, it is important to investigate the influence of systematic effects to its stability and its result. The stability of the fit has been successfully checked under all considered systematic effects. This was achieved by a MC study including a large number of fits of the reference PDFs to 10 pb^{-1} datasets produced under each effect. A systematic effect can bias the result of the fit in two different ways. The first is by changing the number of signal events in the selected sample. As a result, the fit might still guess the correct number of events in the dataset, but this number of events will correspond to an efficiency

[2] The signal and background PDFs displayed on Fig. 8.2 are organized in bins of a larger size compared to those of Fig. 8.1. There are two reasons for this. The first is to follow the binning selected for the 10 pb^{-1} data and the second is to eliminate statistical fluctuations from the PDFs. For more on this subject, the reader is referred to Appendix B

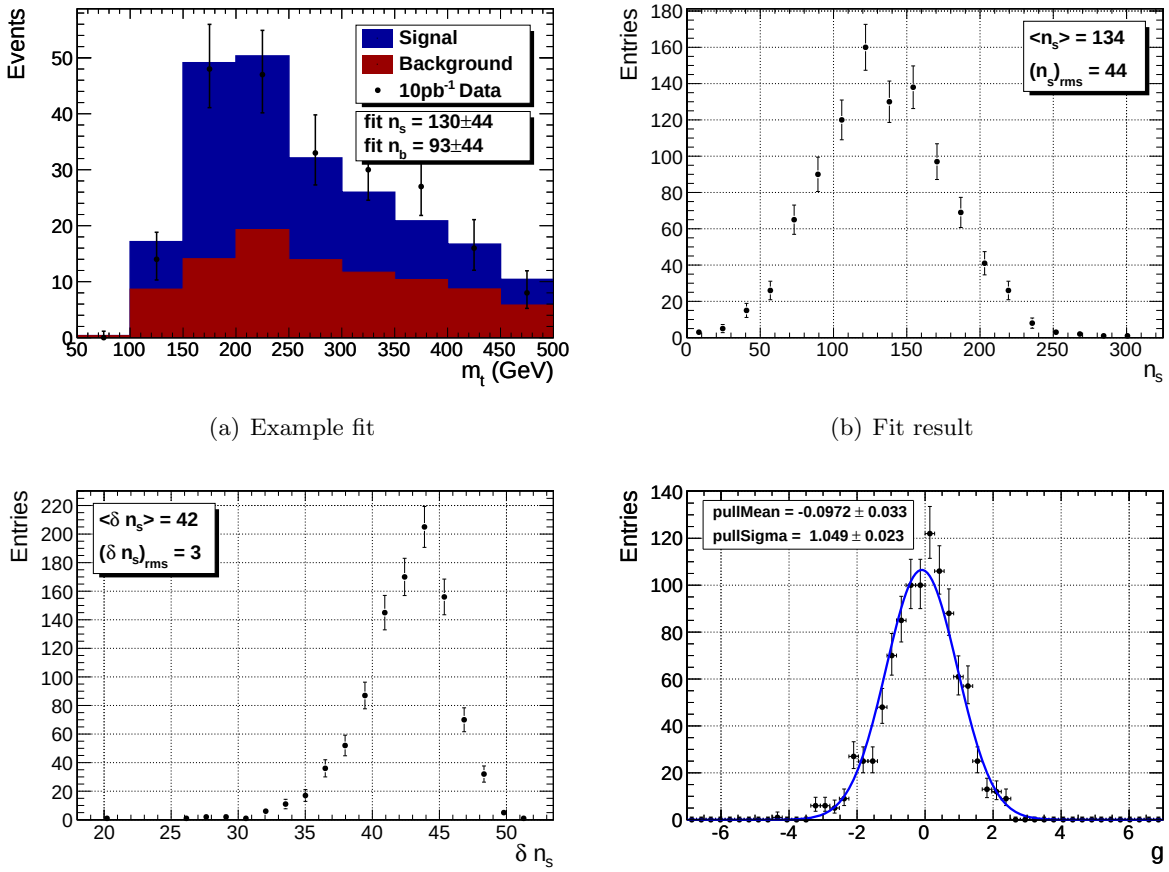


(a)



(b)

Figure 8.1: Invariant mass distribution of the three hadronic-side jets obtained from the full MC sample (100 pb^{-1}). The selection strategy here corresponds to a "loose" muon and a jet-parton assignment by the c_1 criterion with no solution pruning (see text). The stack plot of the different types of signal and background events is displayed above and the normalized signal ($f_s(M_3)$) and background PDFs ($f_b(M_3)$) are displayed below.



(a) Example fit

(b) Fit result

(c) Error

(d) Pull

Figure 8.2: ML fit of the invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, M_3. The plots correspond to the loose selection, the c_1 criterion has been used for the jet-parton assignment and no solution pruning has been applied (see text). An example fit on toy MC data (equivalent to 10pb^{-1}) is shown in the top-left figure. The distributions of the estimated number of signal events (top-left), the fit error (bottom-left) and the pull (bottom-right) obtained from one thousand toy MC experiments are also shown. The bin width is 50GeV.

different from the one extracted from the reference sample and used to calculate the cross-section. The second way is by changing the shape of the M_3 signal and background distributions.

The change in the number of events due to a systematic effect can be predicted using the systematics-affected MC samples. All that is required is to count the number of events passing the cuts and falling within the M_3 fit range. The reasons why a systematic effect biases the number of signal and background events in a certain way are in general easy to understand or can become clear with some investigation. On the other hand, changes in the shapes of the PDFs are difficult to quantify and the way they affect the ML fit is very hard to predict or even explain in retrospect. Each of the numerous steps of any selection strategy introduces a different kind of bias into the selected sample. These sometimes correlated biases result in a M_3 distribution of a unique shape. The exact way a systematic effect changes each of these distributions and how this change is propagated to the fit result is also unique to the distribution. As a result, general trends are not easy to identify and explain. In the following three paragraphs, we look into each source of systematic bias separately and, where possible, explain the effect it has on the fit result.

8.2.1 Misalignment and miscalibration

As we have seen in Sec. 7.2.2 (Tab. 7.2 has been calculated for all events passing the selection and not those within the 0 GeV $<$ M_3 $<$ 500 GeV fit range, but is still accurate enough), misalignment and miscalibration has a relatively small effect on the number of background events (up to 6%). The number of signal events is less affected (up to 2%).

A larger bias in the result of the fit has to be attributed to a difference in the PDF shapes. We obtain the M_3 PDFs for the signal and the background from the reference sample ("reference PDF") and from the sample produced under the systematic effect under examination ("systematic PDF"). Figures 8.3(a) and 8.3(b) show a comparison of the signal and background PDFs respectively for ideal and startup conditions detector. Evidently, the signal PDF does not change significantly. The change in the background seems much more important, as the distribution corresponding to the ideal detector appears to be narrower than that corresponding to the misaligned and miscalibrated detector. A possible explanation would be that the miscalibration results in a lower jet energy resolution leading to a wider distribution. It should however be noted that the uncertainty on the shape of the background is quite large, due to the small amount of simulated data. Particularly the QCD component, which is quite significant in the lower M_3 bins is made up of only fourteen (weighted) events. Furthermore, this distribution-widening effect is not observed for the signal distribution and, if we try to fit a continuous Landau shape to the background PDF (Fig. 8.3(d)), it disappears even for the background [3].

8.2.2 Pileup

The signal shape extracted from the pileup-including sample has a wider peak, slightly shifted in the direction of higher energies. This can be seen on Fig. 8.4. The number of background events increases as a result of the pileup, whereas the number of signal events remains practically stable in the "loose" muon selection (Figure 8.4(e) should be compared to Fig. 8.1(a)) and even slightly decreases in the "tight" muon selection (913 signal events as opposed to 950 for the reference sample). The factors contributing to these changes are the following:

- Pileup events add to the energy of the signal and background jets leading to an increase in their invariant mass.
- The extra energy is distributed inhomogeneously among the jets, thus smearing the invariant mass of the jets correctly assigned to the hadronic side of the decay.

[3] More information on how the continuous M_3 PDFs appearing on Fig. 8.3(f), 8.4(f) and 8.5(f) have been constructed can be found in App. B

- A number of events which did not fulfill the jet requirements, now fulfill them. These are mostly background events, since the number of high E_{T} jets in signal is much larger. The result of this is a relative increase in the number of selected background events.

- The added activity in the tracker and the calorimeter interferes with the muon isolation. This effect works in the opposite direction of the previous one, as some events no longer survive the muon cuts. It is interesting to note that in the case of the "tight" muon selection, the effect of the jet energy increase for signal events is completely overturned by the deterioration of the muon isolation, leading to a reduced number of signal events.

We see that the exact way in which pileup affects the fit result is determined by a number of different effects of variable importance, not all of which work in the same direction. A shifted or widened peak for example, would probably lead to an underestimation of the number of signal events, as the relatively narrow peak of the signal PDF becomes harder to fit to the data. We therefore expect the signal to be systematically underestimated by the fit to first approximation. This, however, might be masked by an increasing number of events passing the jet cuts, like we see in the loose muon selection.

8.2.3 Jet energy scale

Finally, a jet energy scale increased by 10% effectively loosens the jet E_{T} cut, resulting in a significant increase in both signal (of the order of 20%) and background (50%). For this reason, a fit to a dataset produced under this effect is more likely to be biased towards higher values of n_s.

The signal and background shapes are also affected. Figure 8.5 shows how the peak of the signal distribution is clearly shifted towards the higher values of M_3, because of the additional jet energy. This was expected, since all three jets coming from the top quark and participating in the computation of M_3 have their energy increased by 10%. The shift of the M_3 peak causes the PDF peak to become more difficult to fit to the data, leading to a reduction in the n_s estimated by the fit. This works towards countering the effect of the additional signal events entering the dataset. The background is not subject to this shift. The reason for this is that the maximum in the background distribution does not correspond to a decaying particle of a specific invariant mass. Even though all events in the reference have the energy of their jets increased, there is also a large number of new events of less energetic jets entering the selection. These new events preserve the shape of the background M_3 distribution. There are of course new events entering the signal distribution as well. Those events, when not correctly solved, are the reason why the "unmatched" component of the signal distribution is not shifted to the right as much as the peak of "matched" events is.

8.2.4 Estimating the systematic uncertainty

To quantify the systematic uncertainties we keep the Monte Carlo templates obtained from the reference samples fixed. One by one, each of the systematic effects described above is applied to the simulation of the generated samples and the reference templates are fit to the resulting dataset. The difference between the result of the fit to the reference and the systematic-affected dataset is taken as the systematic bias due to the effect. Figures 8.3(f), 8.4(f) and 8.5(f) show examples of such fits to systematic-affected datasets. The systematic uncertainty estimates for all possible combinations of selection strategy and jet-parton assignment are summarized in Tab. 8.1.

The merit of this method is that most of the events in the reference sample are also in the systematic-affected sample and vice versa. This means that almost all of the difference between the two samples is due to the systematic effect. On the other hand, when using datasets of a small

size we run the risk of only getting a small, inadequate subset of all the possible event movements between bins, which is not really representative of the total change that the systematic effect brings to the distribution. This makes our results inaccurate. As already stated, the size of the MC dataset for the background is limited. In particular for the loose muon selection, there is a significant QCD component, which is only represented by a few events of high weight[4]. These few events make our estimates of the systematic bias very uncertain. It has been checked that by removing them from both the reference and the systematic-affected dataset we obtain very different results.

This is not to say that the systematic biases calculated for the tight muon selection are completely reliable. An attempt to repeat the estimation using a different bin width led to considerably different results on the systematic bias. This shows that a very reliable and accurate estimation of the systematic uncertainty of the M_3 fit requires larger MC samples. This is also the main reason why general trends in the way the same systematic effect affects different samples are very hard to identify. However, the results obtained in this section can still serve as an estimate of at least the order of magnitude of the systematic uncertainties. Comparing with the uncertainties calculated for the event-counting method (Tab. 7.2), we see that even though the misalignment & miscalibration and pileup uncertainties are at least as large, the jet energy scale uncertainty is significantly smaller. This results in a total systematic uncertainty consistently lower than 20% for the tight selection (for all jet-parton assignment strategies). In comparison, the minimum total systematic uncertainty achieved by the event-counting method was $(^{+38}_{-29})$.

8.3 Jet multiplicity fit

The same method applied to the M_3 distribution can also be used with other discriminating variables. As explained in Sec. 6.2.3, signal events are expected to include a larger number of jets than events coming from any of the sources of background. Indeed, Fig. 8.6 shows that the signal and background not only include a different average number of jets, but also have jet multiplicity distributions of a very different shape. This is the reason why using the jet multiplicity as the observable in a ML fit can be very advantageous.

In order to have a highly populated, background-rich region in the jet multiplicity fit, we relax the event selection to include two- and three-jet events. In an event-counting measurement, this is particularly useful, because we can easily compare the Monte Carlo prediction for the background-rich region to data and check that the prediction is properly normalized before counting the events in the signal region. In the case of a ML fit this is automatically achieved by simply including the background-rich region in the fit range. In our case, this part is played by the two-jet bin.

Figure 8.7 shows an example fit to 10 pb^{-1} of data sampled from the "tight" selection PDFs. The results of a Monte Carlo study to verify that the fit functions properly are also displayed. The pull distribution (Fig. 8.7(d)) shows that this is indeed the case. The large number of events, since the two and three-jet bins are included, and the clear difference between the signal and background shapes lead to a relatively small fit error ($\approx 11\%$), which means that a good fit behaviour can be ensured even with little data.

8.3.1 Systematic uncertainties for the jet multiplicity fit

As for the M_3 fit, systematic effects bias the result of the fit by changing the number of signal events in the selected sample and the shapes of the signal and background distributions. Figure 8.8 shows how the number of signal and background events and the distribution shapes

[4]The reader is reminded that the QCD dataset corresponds to 8.7 pb^{-1} of data, the weight for a 100 pb^{-1} plot is therefore 11.5.

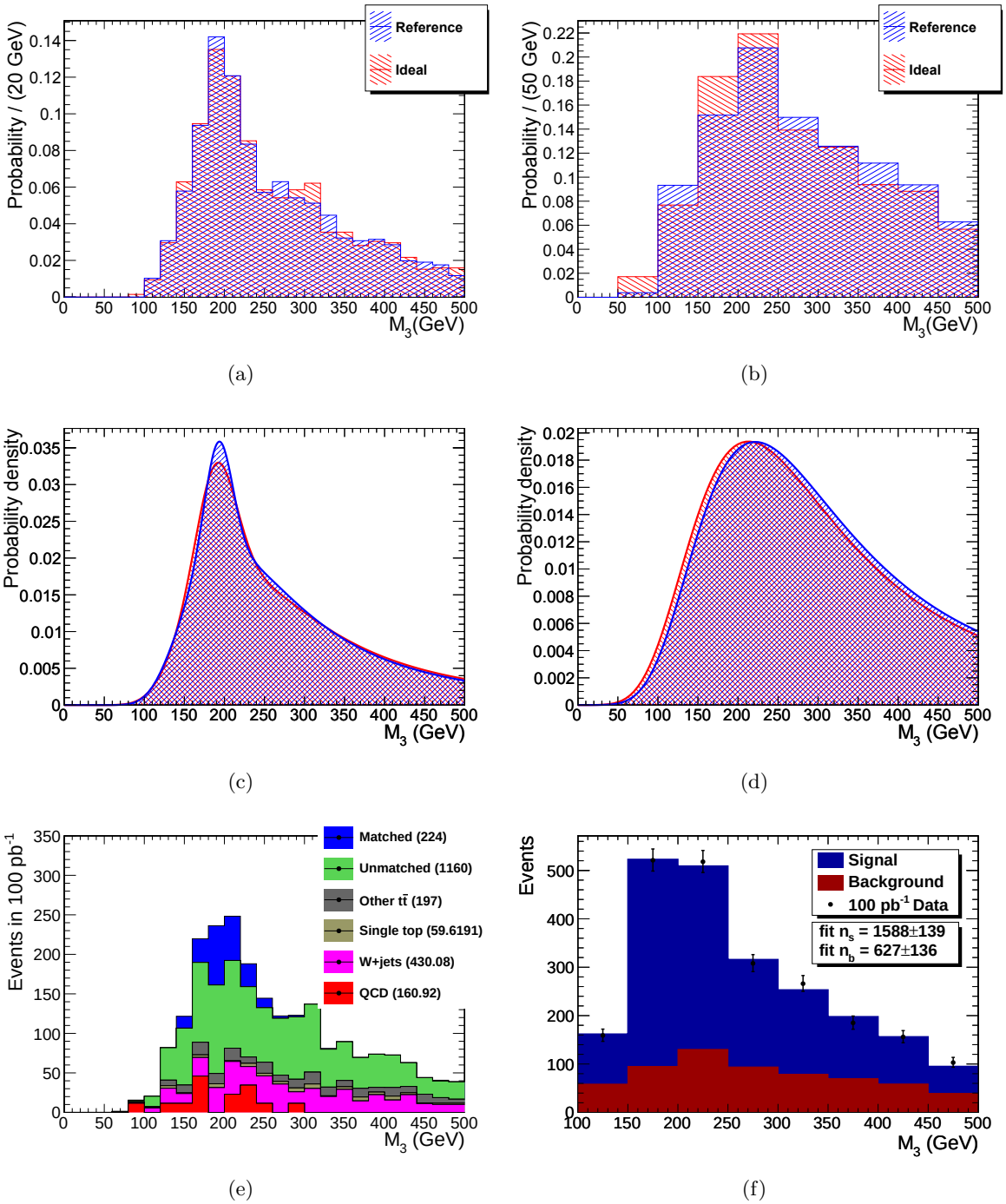


Figure 8.3: Effect of misalignment and miscalibrations on the M_3 distribution (loose muon selection, c_1 jet-parton assignment, no solution pruning - see text). The signal (top left) and background (top right) discrete PDFs are presented with (red) and without (red) pileup. Continuous shapes that help to visualize the PDFs have been fit to the histograms in the top two figures and are also shown (middle). The stack plot of the different types of signal and background events in the ideal scenario sample are displayed below left. The fit of the reference PDFs to the ideal scenario data is also shown (bottom right).

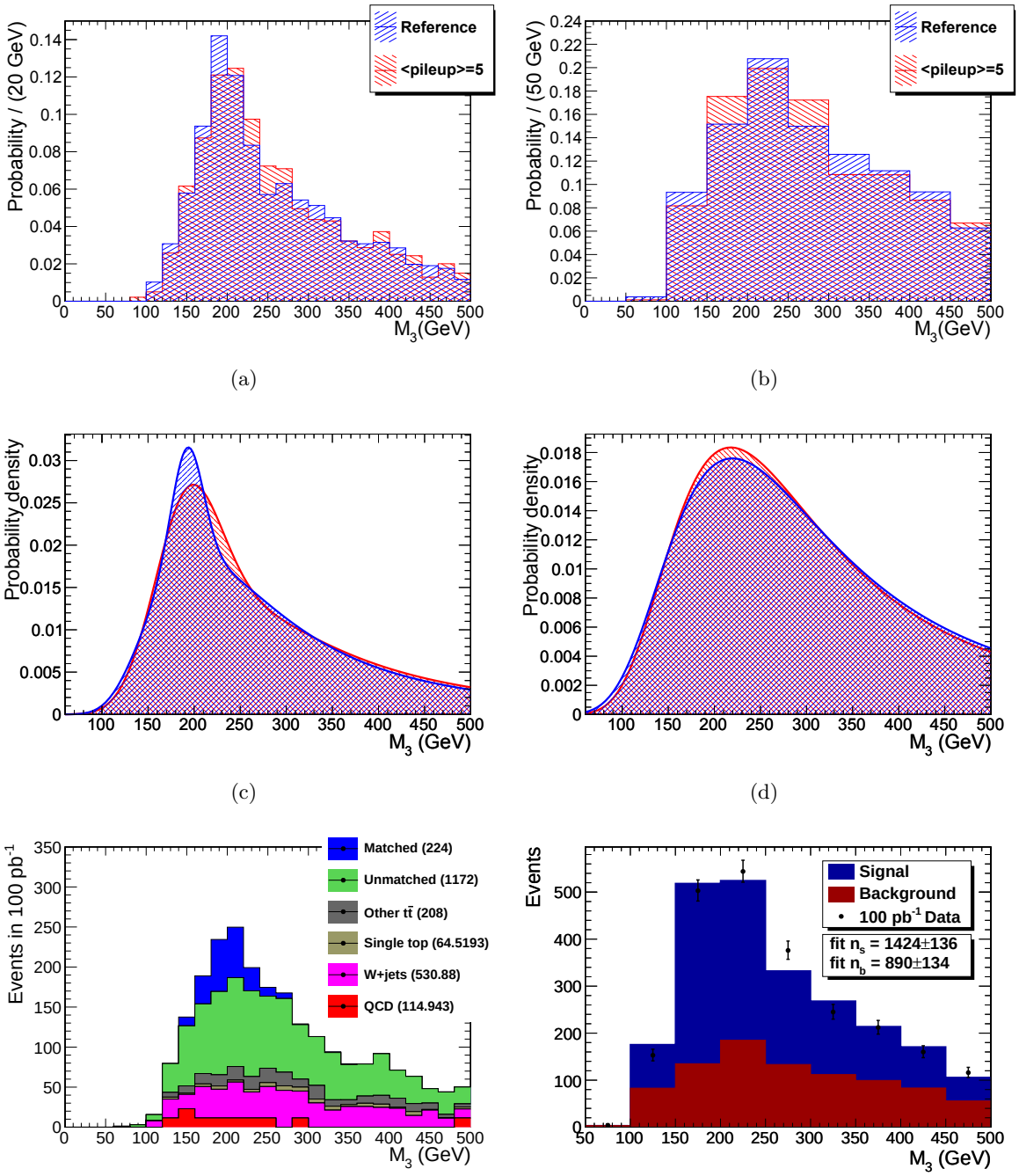


Figure 8.4: Effect of pileup on the M_3 distribution (loose muon selection, c_1 jet-parton assignment, no solution pruning - see text). The signal (top left) and background (top right) discrete PDFs are presented with (red) and without (red) pileup. Continuous shapes that help to visualize the PDFs have been fit to the histograms in the top two figures and are also shown (middle). The stack plot of the different types of signal and background events in the pileup-affected sample are displayed below left. The fit of the reference PDFs to the pileup-affected data is also shown (bottom right).

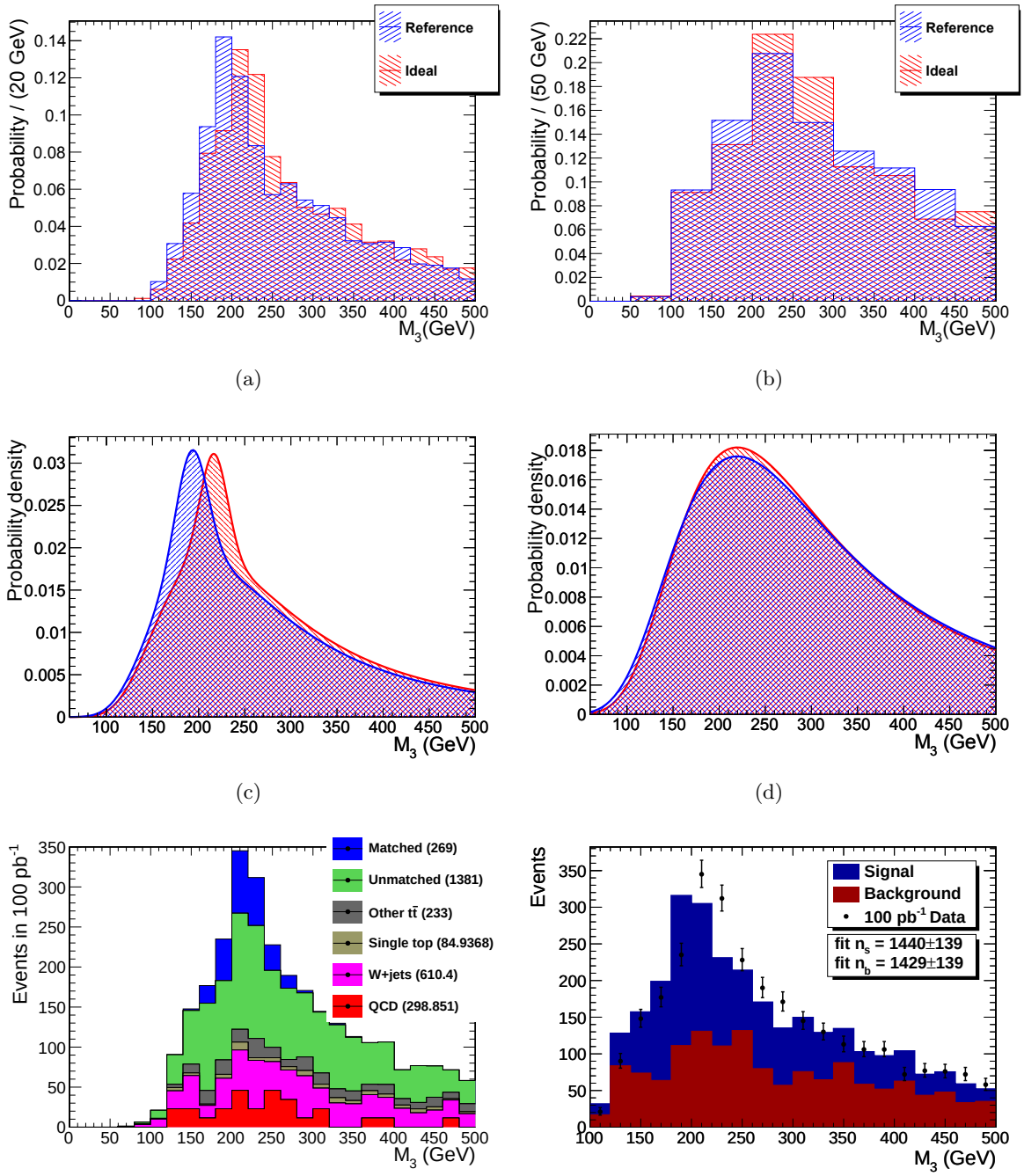
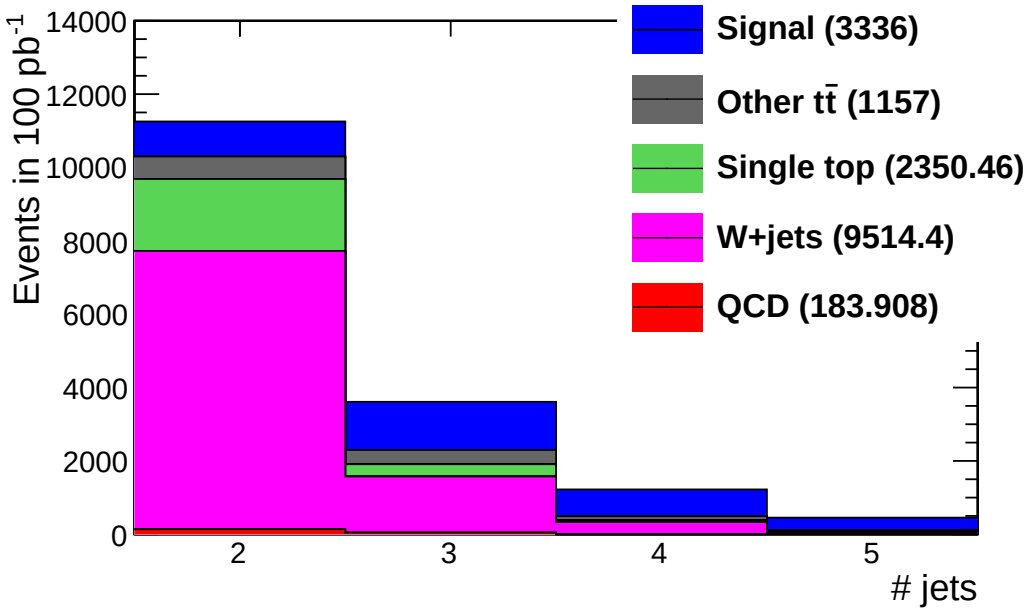


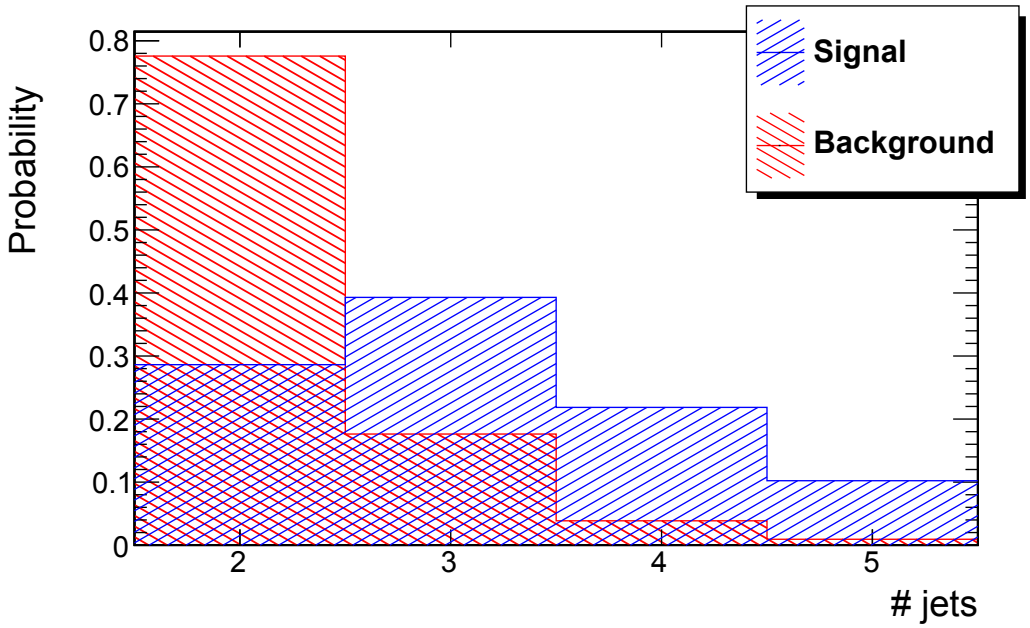
Figure 8.5: Effect of a 10% increase on the jet energy scale on the M_3 distribution (loose muon selection, c_1 jet-parton assignment, no solution pruning - see text). The signal (top left) and background (top right) discrete PDFs are presented in the presence (red) and absence (blue) of the systematic effect. Continuous shapes that help visualize the PDFs have been fit to the histograms in the top two figures and are also shown (middle). The stack plot of the different types of signal and background events in the systematic-affected sample are displayed below left. The fit of the reference PDFs to the systematic-affected dataset is also shown (bottom right).

M_3 fit, loose μ	$\frac{\delta\sigma_{\text{stat}}(10\text{pb}^{-1})}{\sigma}$	$\frac{\delta\sigma_{startup}}{\sigma}$	$\frac{\delta\sigma_{pileup}}{\sigma}$	$\frac{\delta\sigma_{+10\%}}{\sigma}$	$\frac{\delta\sigma_{-10\%}}{\sigma}$	$\frac{\delta\sigma_{sys.}}{\sigma}$
no pruning, c_1	32%	$\pm$7%	-14%	-15%	+18%	$^{+22\%}_{-20\%}$
PTM, c_1	35%	$\pm$4%	-12%	-26%	+35%	$^{+37\%}_{-28\%}$
$\angle\vec{p_p}\vec{p_q}$, c_1	31%	$\pm$11%	-22%	-14%	+16%	$^{+33\%}_{-20\%}$
m_W, c_1	41%	$\pm$11%	-19%	-25%	+34%	$^{+35\%}_{-30\%}$
All but m_W, c_1	32%	$\pm$7%	-15%	-24%	+31%	$^{+51\%}_{-27\%}$
no pruning, c_2	35%	$\pm$9%	-15%	-24%	+32%	$^{+34\%}_{-28\%}$
PTM, c_2	37%	$\pm$2%	-9%	-33%	+50%	$^{+51\%}_{-35\%}$
$\angle\vec{p_p}\vec{p_q}$, c_2	31%	$\pm$13%	-18%	-23%	+30%	$^{+34\%}_{-28\%}$
m_W, c_2	39%	$\pm$15%	-15%	-32%	+47%	$^{+51\%}_{-37\%}$
All but m_W, c_2	33%	$\pm$8%	-5%	-27%	+38%	$^{+40\%}_{-30\%}$
M_3 fit, tight μ	$\frac{\delta\sigma_{\text{stat}}(10\text{pb}^{-1})}{\sigma}$	$\frac{\delta\sigma_{startup}}{\sigma}$	$\frac{\delta\sigma_{pileup}}{\sigma}$	$\frac{\delta\sigma_{+10\%}}{\sigma}$	$\frac{\delta\sigma_{-10\%}}{\sigma}$	$\frac{\delta\sigma_{sys.}}{\sigma}$
no pruning, c_1	28%	$\pm$13%	+18%	-9%	+10%	$^{+19\%}_{-19\%}$
PTM, c_1	32%	$\pm$8%	+4%	-6%	+6%	$^{+14\%}_{-14\%}$
$\angle\vec{p_p}\vec{p_q}$, c_1	27%	$\pm$10%	+16%	-10%	+11%	$^{+18\%}_{-17\%}$
m_W, c_1	36%	$\pm$8%	+10%	-10%	+11%	$^{+17\%}_{-16\%}$
All but m_W, c_1	32%	$\pm$12%	+18%	-8%	+8%	$^{+17\%}_{-17\%}$
no pruning, c_2	32%	$\pm$1%	+16%	-1%	+1%	$^{+10\%}_{-10\%}$
PTM, c_2	38%	$\pm$2%	+9%	-11%	+13%	$^{+16\%}_{-15\%}$
$\angle\vec{p_p}\vec{p_q}$, c_2	28%	$\pm$6%	+11%	-4%	+5%	$^{+12\%}_{-12\%}$
m_W, c_2	37%	$\pm$2%	+10%	-13%	+15%	$^{+19\%}_{-17\%}$
All but m_W, c_2	31%	$\pm$4%	+13%	-14%	+16%	$^{+20\%}_{-18\%}$
#jets fit	$\frac{\delta\sigma_{\text{stat}}(10\text{pb}^{-1})}{\sigma}$	$\frac{\delta\sigma_{startup}}{\sigma}$	$\frac{\delta\sigma_{pileup}}{\sigma}$	$\frac{\delta\sigma_{+10\%}}{\sigma}$	$\frac{\delta\sigma_{-10\%}}{\sigma}$	$\frac{\delta\sigma_{sys.}}{\sigma}$
loose μ	9%	2%	$-$12%	$-$30%	+43%	$^{+43\%}_{-32\%}$
tight μ	11%	2%	$-$3%	$-$25%	+33%	$^{+33\%}_{-25\%}$

Table 8.1: Uncertainties on the cross-section measured by the ML fit. The first column holds the value of the relative ML fit error for an integrated luminosity of 10pb^{-1}. The uncertainties due to different systematic effects (misalignment&miscalibration, pileup, jet energy scale overestimation and underestimation by 10%), as estimated from an ML fit of the reference PDFs to the systematic-affected 100pb^{-1} dataset (see text), occupy the next four columns. The total systematic uncertainty, including a 10% uncertainty on the integrated luminosity can be found in the last column.



(a)



(b)

Figure 8.6: Number of jets of $E_{\mathrm{T}} > 40\ GeV$ after the application of the tight muon cuts. The one-jet bin is not included and the five-jet bin also includes the higher jet multiplicity events. The signal and different sources of background are shown stacked above and the signal and total background distributions (each normalized to unit area) are shown below.

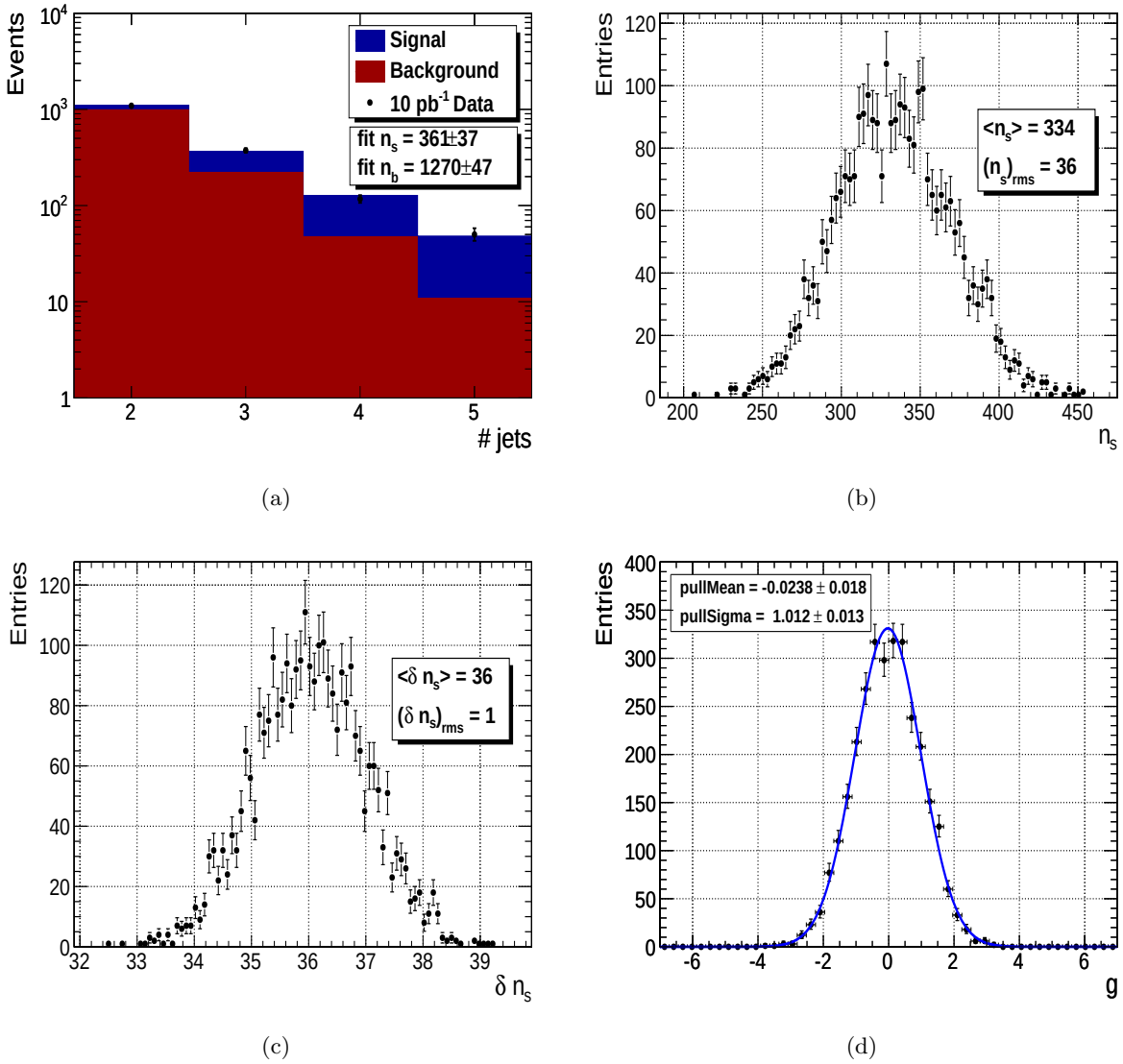


Figure 8.7: ML fit of the number of jets of $E_{\mathrm{T}} > 40\ GeV$ after the application of the tight muon cuts (see text). An example fit on toy MC data (equivalent to $10\mathrm{pb}^{-1}$) is shown in the top-left figure. The distributions of the estimated number of signal events (top-right), the fit error (bottom-left) and the pull (bottom-right) obtained from three thousand toy MC experiments are also shown.

change under different systematic effects. The systematic uncertainties on the fit results have been estimated as for the M_3 fit and added to Tab. 8.1. The fits to the 100pb^{-1} systematic-affected datasets for the tight selection strategy are shown in Fig. 8.9.

Predictably, the misalignment and miscalibration have a very small effect on the number of reconstructed jets (Fig. 8.8(a)), as they did in the case of the M_3 distribution. They also have a minimal effect on the distribution shapes (Fig. 8.8(b)). As a result, the uncertainty due to this kind of effects is very small. The extra tracks and calorimeter deposited energy in the pileup events is a much more important source of uncertainty. As more jets pass the 40 GeV threshold, there is a number of new events entering the selection and some already selected events moving into higher jet multiplicity bins. On the other hand, the added "noise" interferes with the lepton isolation as explained above. As a result, the number of selected background events increases (since more of them now fulfill the jet requirements) but the number of selected signal events decreases (since signal is already rich in jets and only suffers from the poorer muon isolation). As for the M_3 fit, this effect is stronger when the selection places higher demands on the muon isolation, leading to an important reduction in signal events (Fig. 8.8(c)). The change in the distribution shapes (Fig. 8.8(d)), which would normally cause the fit to overestimate the signal is countered by this reduction. The final result is a much smaller bias for the tight selection ($\approx 3\%$) than for the loose ($\approx 12\%$). Finally, a jet energy scale increased by 10% has a similar effect to the jets as the pileup events. It causes more jets to pass the 40 GeV threshold, increasing the jet multiplicity of many events. The already jet-rich signal does not increase as much as the background because of this effect (Fig. 8.8(e)). Nevertheless, the result of the fit is strongly affected, due to the change in the shape of the jet multiplicity distributions (Fig. 8.8(f)), which lead to a considerably overestimated signal (at approximately 30% of its reference value).

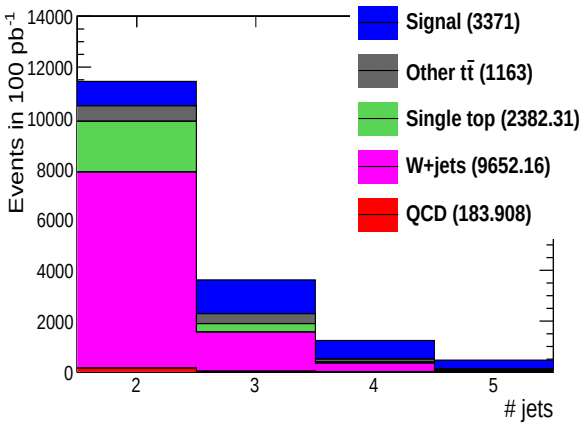
8.4 Conclusions

In this chapter we investigated an alternative method of measuring the $t\bar{t}$ cross-section using an extended ML fit. Compared to the basic event-counting method presented in Ch. 7, this method has two important advantages:

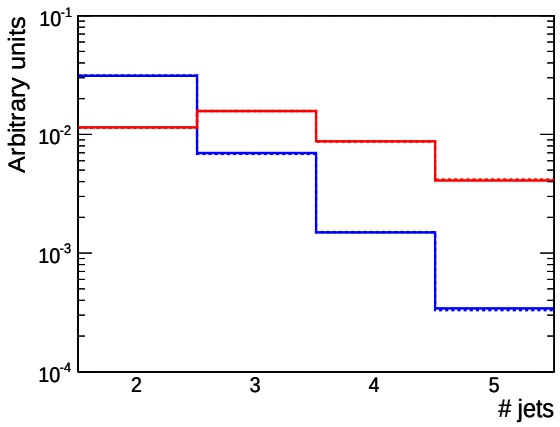
- It does not depend on the MC to predict the correct cross-section for the background. Instead it only uses the MC to obtain the distribution of the discriminating variable.
- It can provide us with a cross-section measurement which is less prone to systematic uncertainty than the result we obtain from the event-counting method.

Two discriminating variables were examined. The first was the invariant mass of the three jets assigned to the hadronic side of the $t\bar{t}$ decay. A Monte Carlo study showed that the ML fit of this variable is stable and can be relied on to measure the cross section using 10pb^{-1} of data. The main disadvantage of using this variable is that the statistical uncertainty of the M_3 fit is significantly higher than that of the event-counting method (see the first column of Tab. 7.2 and Tab. 8.1). However, the systematic uncertainty of the M_3 fit is much smaller than that of the event-counting method (see Sec. 8.2.4. It is interesting to note that the systematic uncertainty for the M_3 fit, estimated to be less than 20%, is below the level of the (statistical) uncertainty of the fit, which is higher than 30% for most selection/jet-parton assignment strategies. The opposite is true for the event-counting method (the best result was $\pm 14\%(\text{stat.})^{+38\%}_{-29\%}(\text{syst.})$). This indicates that as the amount of available data increases beyond 10 pb^{-1} the advantage of the M_3 fit over the event-counting method will grow.

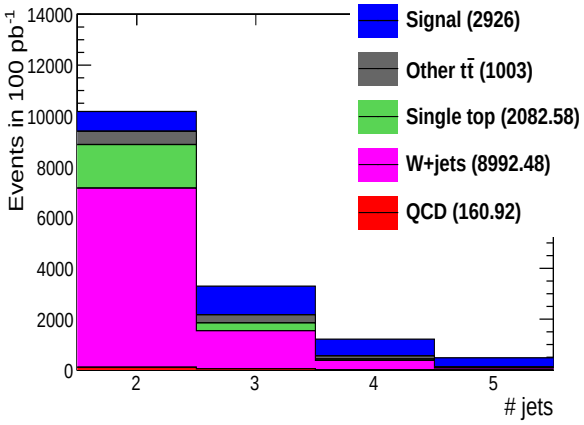
The second discriminating variable examined was the jet multiplicity. The advantage the jet multiplicity has over the invariant mass fit is that, by allowing the two- and three-jet bins into the selection, it increases the available statistics while maintaining a clear difference between the signal and background distribution. Even more importantly, the inclusion of the background-rich



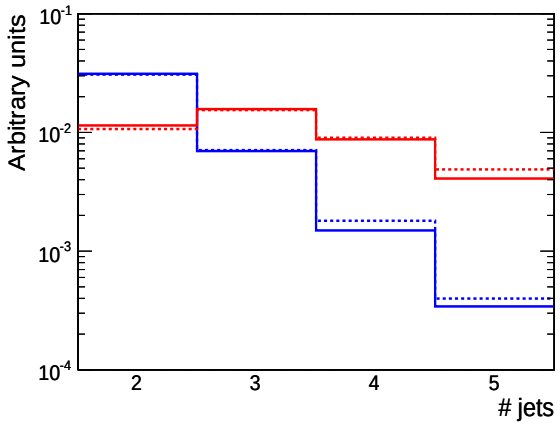
(a) Ideal, Tight.



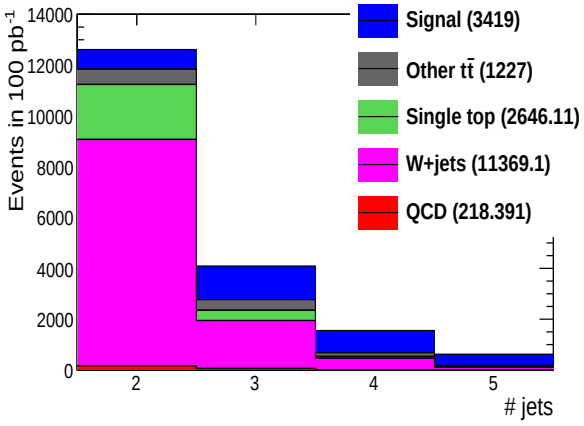
(b) Startup and Ideal, Tight.



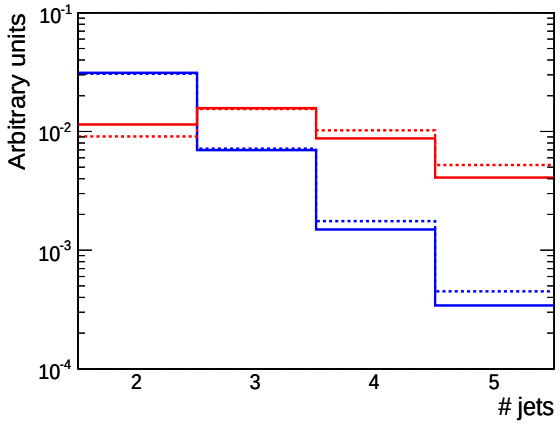
(c) Pileup, Tight.



(d) Startup and with pileup, Tight.

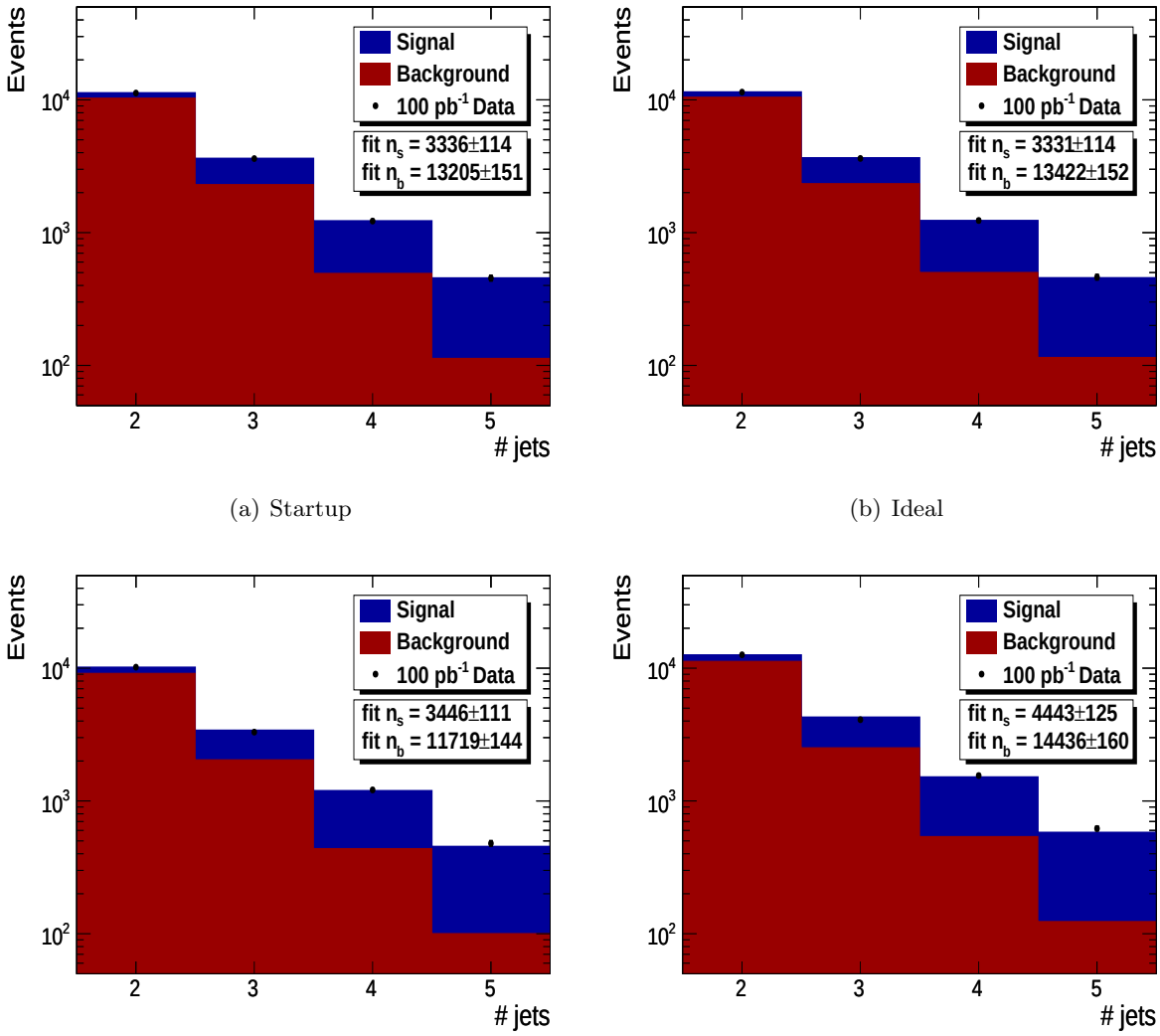


(e) E_{jet} + 10%, Tight.



(f) Startup and with E_{jet} + 10%, Tight.

Figure 8.8: Effect of misalignment & miscalibration (top), pileup (middle) and a 10% jet energy scale bias (below) on the jet multiplicity distribution. The selection requires a "tight" muon. The contributions of the various sources of background and the signal under each systematic effect are drawn stacked on the left. On the right, the signal (background) PDFs are drawn in blue (red). The solid (dashed) lines correspond to the reference (systematic-affected) PDFs. All PDFs are normalized to the same area.



(a) Startup

(b) Ideal

(c) Pileup

(d) E_{jet} $+10\%$

Figure 8.9: Systematic biases of the jet multiplicity ML fit. Each of the figures shows the fit of the reference PDF to the full 100pb^{-1} dataset produced under the systematic effect. The selection requires a "tight" muon. The systematic bias is calculated by subtracting the result of the fit to the systematic-affected dataset from the result of the fit to the reference dataset (top left).

two-jet bin ensures that the background is properly normalized to the data, elegantly removing the dependence on the MC prediction of the cross-section. The result is a much smaller statistical uncertainty ($\pm 11\%$ for the tight selection). On the other hand, jet multiplicity distribution is very sensitive to the jet energy scale, resulting in a larger systematic uncertainty ($^{+33\%}_{-25\%}$ for the tight selection).

Chapter 9

Conclusions and outlook

The high LHC production cross-section of $t\bar{t}$ pairs allows an early "rediscovery" of the top quark at the CMS experiment. A significant signal can be established using $O(10\ \text{pb}^{-1})$ of integrated luminosity by searching for single leptonic $t\bar{t}$ decays in the muon channel.B-tagging and missing transverse energy reconstruction, which depend on a good knowledge of the detector, are not required by the selection. The excess observed over the background expectations can be identified as $t\bar{t}$ signal by plotting the invariant mass of the three jets coming from the hadronically decaying top quark (M_3). The jets corresponding to the correct partons can be efficiently selected using specific jet-parton assignment strategies.

A first measurement of the cross-section can be made with the same amount of integrated luminosity. A simple, event-counting method relying on the Monte Carlo prediction on the background would have a statistical uncertainty of approximately 10%, but would be very dependent on the jet energy scale, leading to a systematic uncertainty of at least 40%.

Using a maximum likelihood (ML) fit to measure the cross-section is a good alternative to the event-counting method, as it does not depend on the Monte Carlo prediction on the background cross-section, but only on the shape of the distribution of the observable quantity. Using the invariant mass of the three jets assigned to the hadronically decaying top quark as the observable, we can obtain a measurement with a statistical uncertainty of approximately 30% using 10 pb^{-1} of data. This is more than the statistical uncertainty of the event counting method, it is however compensated for by the reduced systematic uncertainty, which is estimated to be at the level of 20% (a more accurate estimation would require more MC data than the amount available for this analysis). Furthermore, at this early stage of the experiment, it will probably be easier to obtain more statistics than to improve the jet energy calibration or the jet resolution.

The prospect of using the jet multiplicity as the observable for the ML fit is even more promising. Using a selection that places tight requirements on the muon transverse momentum and isolation we can achieve an uncertainty of $\pm 11\%(\text{stat.})^{+33\%}_{-25\%}(\text{syst.})$. The jet energy scale remains the most important source of systematic uncertainty for this measurement method.

9.1 Possible improvements

The purpose of this final section is to give a brief account of a number of possible changes or additions to this work that could enhance the possibility of an early observation of single-leptonic $t\bar{t}$ decays and the precision of the cross-section measurement.

9.1.1 MC samples

As was explained in Sec. B.3, the Monte Carlo samples used for this analysis, particularly for the $pp \to \mu X$ background, are limited in size. Given sufficient CPU power and storage space,

it would be feasible to repeat the analysis using a larger amount of simulated data. This would lead to a much better understanding of the systematic uncertainties and would significantly add to the early $t\bar{t}$ rediscovery potential and to the reliability and precision of the cross-section measurement.

On a related note, the systematics-affected MC samples produced for this analysis have required an educated guess on the size of the systematic effects at startup. Assuming that the average number of pileup events per collision will be five and that the energy scale will be overestimated by 10% might be correct. However, it would be useful to create more samples with different alignment and calibration, pileup and jet energy scale, to obtain an even better idea on the importance of these effects.

9.1.2 Estimating the background from data

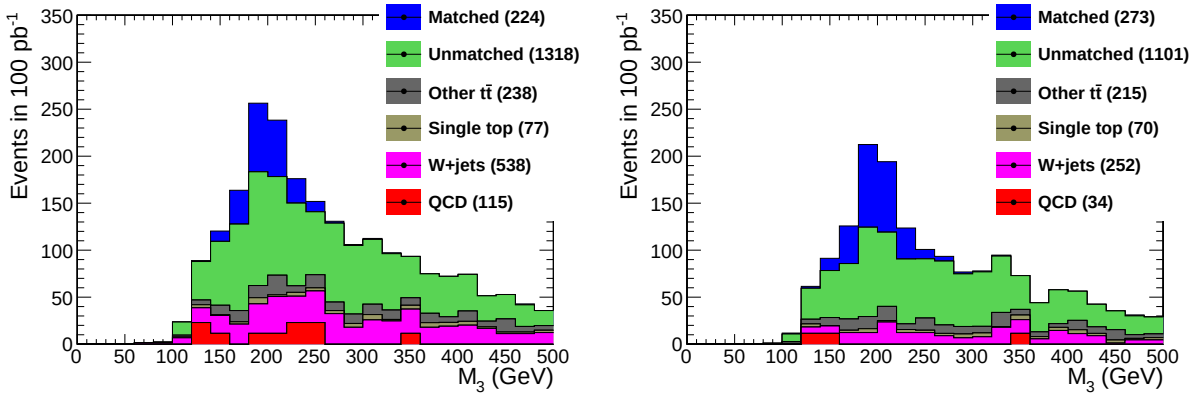
The best way to address the need for accurate background PDFs is to find a method of estimating them from data instead of simulation. The CMS top physics group has started investigating different possible strategies but has not reached a definitive conclusion on what the optimal one is. The main idea behind some of the most promising options is to invert one of the selection cuts that makes a very large difference in the purity of the sample, while leaving the rest of the cuts unchanged. This will eliminate the signal from the selection, without altering the background distribution used for the maximum likelihood fit.

For example, if we require that the muon is not isolated, we will collect a sample rich in QCD events. The amount of QCD will be so large that the contribution of other physical processes will become negligible. By leaving the rest of the cuts unchanged, we can hope that the distribution of a variable which is not strongly correlated with the muon, such us the jet transverse energy, will remain almost unchanged. We will thus have obtained a PDF which is constructed from a very large amount of data and is also free of systematic uncertainties: The Monte Carlo uncertainties are eliminated and all systematic effects present in the signal-rich selection are also present in the PDF.

9.1.3 b-tagging

As mentioned in Sec. 5.5.1, b-tagging is a technique that allows to identify jets originating from b-quarks. In the context of a $t\bar{t}$ analysis, this can prove particularly useful both in reducing the background and in improving the jet-parton assignment. There are two reasons why the use of b-tagging techniques has not been thouroughly investigated in this work. The first is that all b-tagging algorithms rely on well performing track and vertex reconstruction. The misalignment of the tracking systems expected at startup makes it doubtful that the necessary performance will be achieved in the first period of data taking. The second reason is that b-tagging has been found to slightly over-perform in the version of Fast Simulation used for this analysis.

Figure 9.1, which shows the possible improvement that b-tagging can bring to the event selection, should therefore not be taken as an accurate evaluation of the usefulness of b-tagging, but as a clear indication that it is indeed useful. The b-tagging algorithm used here is called the *track-counting* algorithm. The track counting algorithm relies on the computation of the impact parameters of the tracks corresponding to the jet in question. The first step in applying it is therefore to determine which tracks belong to the jet. This is achieved by cutting on the $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ distance between each track and the jet. A set of quality cuts is applied on the tracks considered. Each track needs to have a minimum number of hits in total and in the pixel detector, a high enough transverse momentum and a good χ^2 value. It also has to be located close to the primary vertex and to the jet axis, so as to reject tracks from sources other than the decay of the b-meson (e.g. from photon conversions). Next, the significance of the impact parameter is calculated for each of the tracks associated with the jet from the estimated uncertainties of the track parameters. The impact parameter is the three dimensional distance



(a) No b-tag required

(b) At least one b-tag required

Figure 9.1: Distribution of the invariant mass of the three hadronic-side jets with and without (left) the requirement for a single b-tag. Left: Normal stack plot of the signal and various background components for the loose selection, with jets assigned to partons using the c_1 criterion. Right: Result of the same selection strategy, with the additional requirement for at least one b-tagged jet. The signal to background ratio for the latter is considerably improved. Of the possible solutions only the ones assigning the b-tagged jet to a b-quark are considered. This leads the number of "matched" signal events to a significant increase.

of the track from the primary vertex of the event. It is defined as positive if the track originates downstream from the primary vertex with respect to the jet direction and negative otherwise. After ordering the tracks by impact parameter significance, we take the Nth track and use its impact parameter significance as the discriminator. The b-jets will tend to have more tracks of higher significance than other kinds of jets because of the large lifetime of B-hadrons, therefore a cut can be applied on that discriminator. The jets passing the cut are tagged as b-jets.

Recent versions of the CMS Fast Simulation software guarantee the accurate simulation of the b-tagging performance. Furthermore, extensive studies of the tagging algorithms have investigated their sensitivity to misalignment. As detailed in [49] there are taggers available at CMS which offer a promising compromise between efficiency and robustness, even under startup detector conditions. It would definitely be wortwhile to update the analysis to make it compatible with the latest version of the CMS software and to investigate what benefit can be gained from the use of these algorithms.

9.1.4 Particle flow jets

As explained in Sec. 5.2 the standard jet reconstruction relies solely on calorimeter information and a single calibration factor is used for the whole jet. This is suboptimal for a number of reasons, most notably because:

- Different kinds of particles that might form part of a jet require different calibration factors. Photons in particular, which carry a significant portion of the energy of the jet, would require a calibration different than the one required for hadrons.
- Charged particles are bent by the magnetic field and, depending on their initial direction and their momentum, might exit the jet algorithm cone before they reach the calorimeter, leading to energy loss.
- Magnetic bending changes the final position of charged particles within the jet cone compromising the jet position calculation.

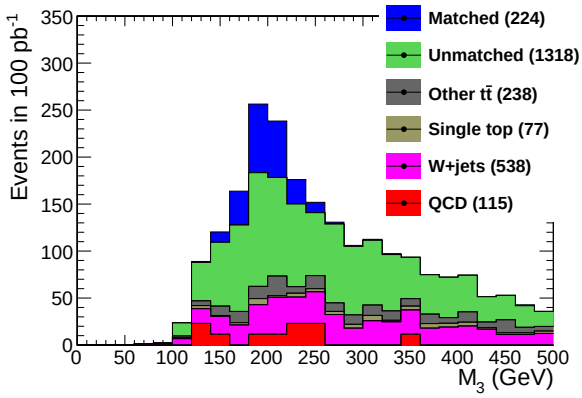
Particle flow is a newly developed reconstruction technique with the potential to produce reconstructed objects of more accurately measured properties and with higher efficiency. This is achieved by reconstructing all stable particles constituting the jet separately, combining information from all subdetectors. This added knowledge reduces or completely avoids the shortcomings of the standard, calorimetric jet. Detailed studies [50] have clearly shown that both the energy and the position resolution can be significantly improved, in particular for the relatively low-energy jets often produced in $t\bar{t}$ decays.

Given the importance of jet reconstruction to this study, it is expected that replacing the standard jets with particle flow jets should lead to a considerable improvement.

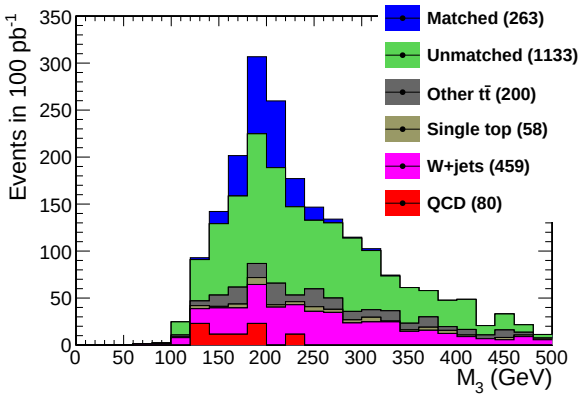
Appendix A

Invariant mass distributions and signal significance for all selection strategies

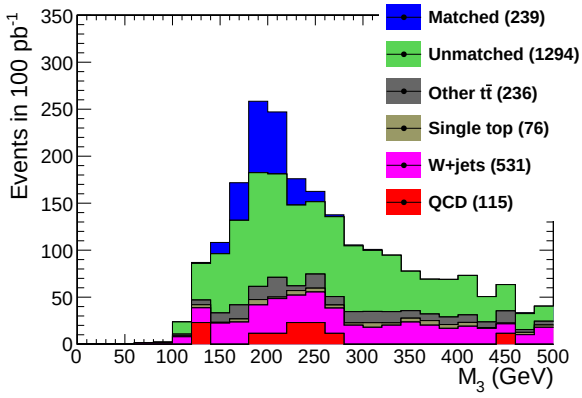
We present here the distributions of the invariant mass of the three jets assigned to the hadronic side of the $t\bar{t}$ decay corresponding to all selection and jet-parton assignment strategies presented in Ch. 6 (Fig. A.1-A.4). The corresponding plots of the significance as a function of integrated luminosity are presented on Fig. A.5-A.7. The method used to calculate the significance for any given integrated luminosity is detailed in Sec. 7.1.



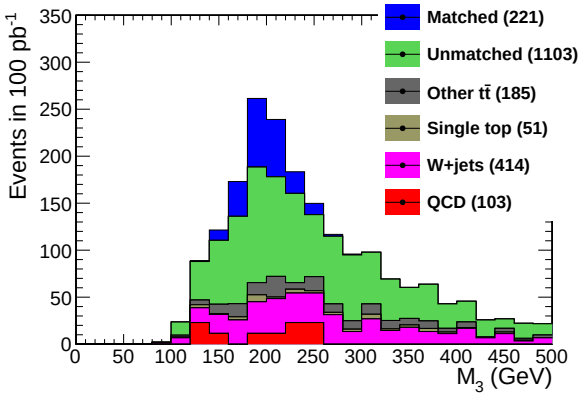
(a) No pruning



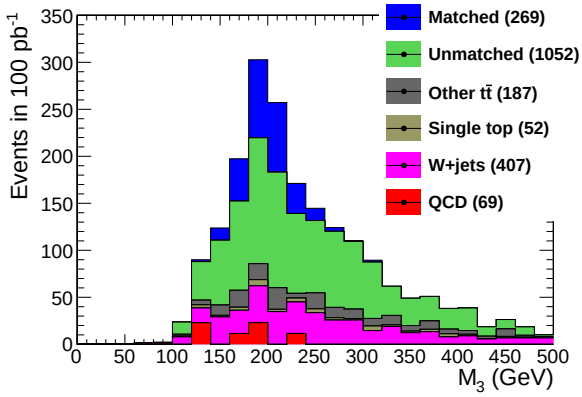
(b) PTM pruning



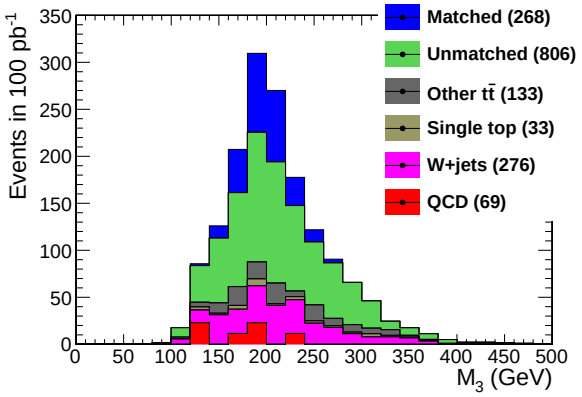
(c) Angle pruning



(d) m_W pruning

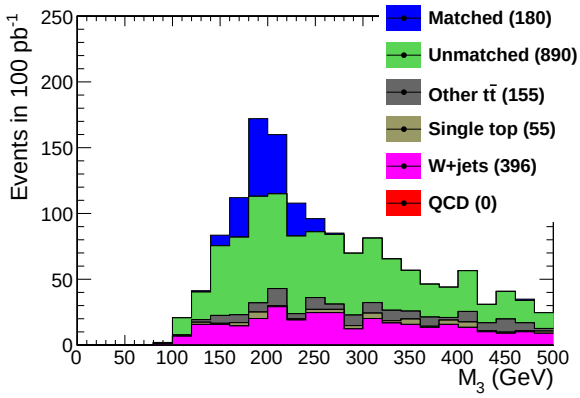


(e) PTM and angle pruning

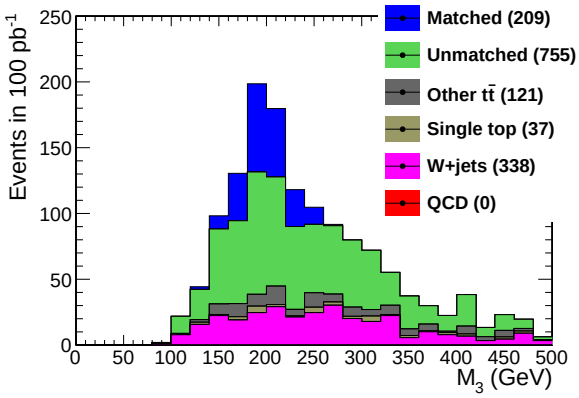


(f) All three pruning methods applied.

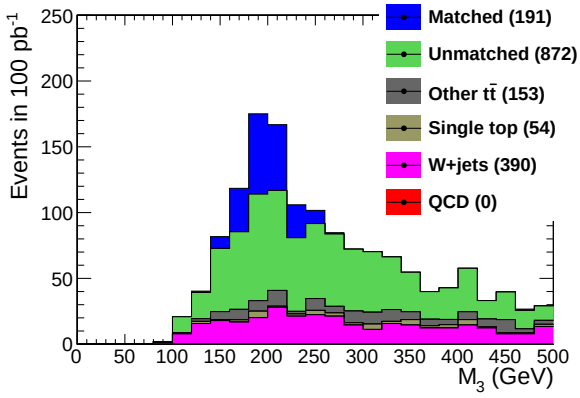
Figure A.1: Invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, after pruning. The plots correspond to the loose selection and the c_1 criterion has been used to select the hadronic-side jets.



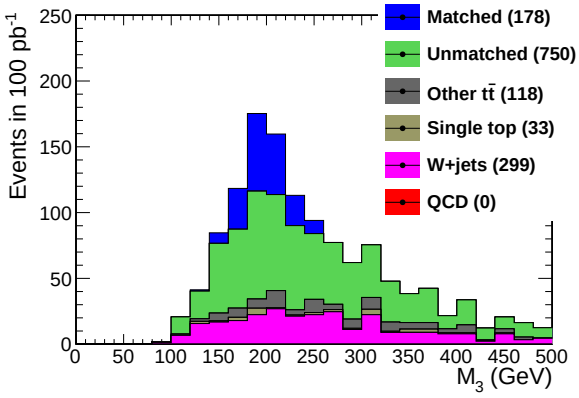
(a) No pruning



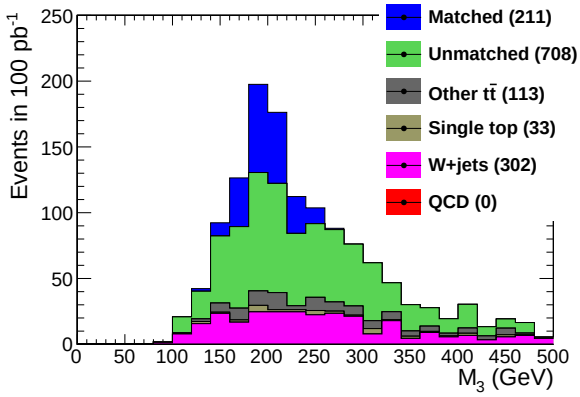
(b) PTM pruning



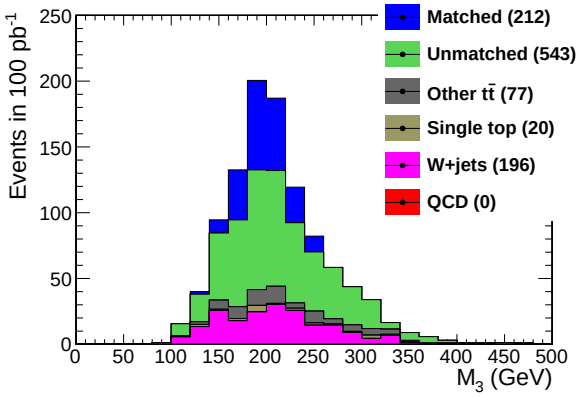
(c) Angle pruning



(d) m_W pruning

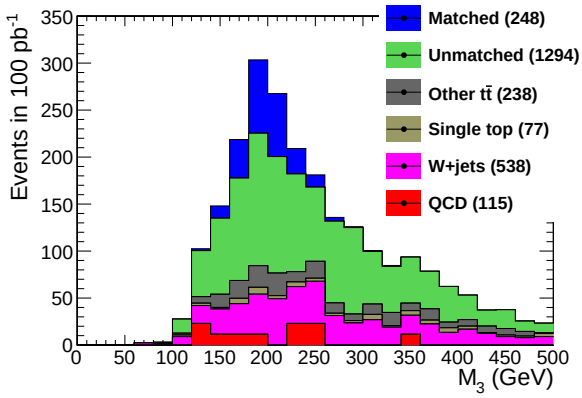


(e) PTM and angle pruning

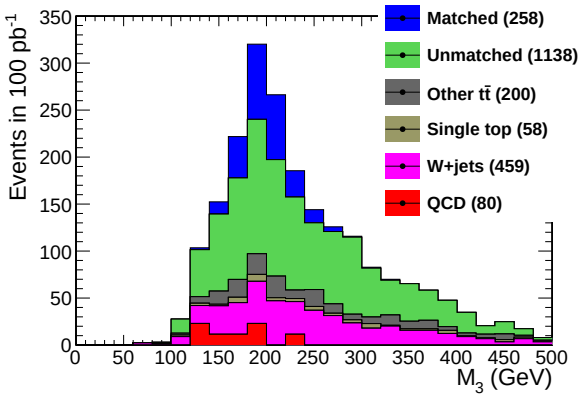


(f) All three pruning methods applied

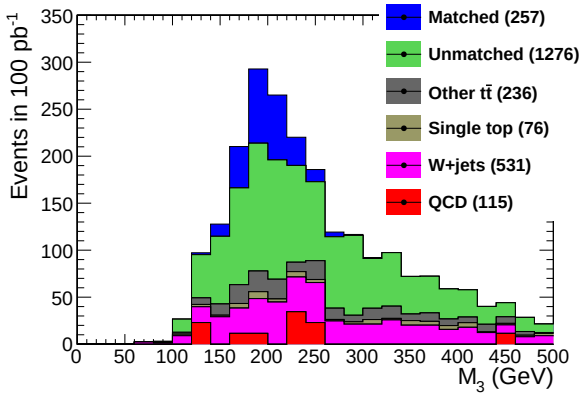
Figure A.2: Invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, after pruning. The plots correspond to the tight selection and the c_1 criterion has been used to select the hadronic-side jets.



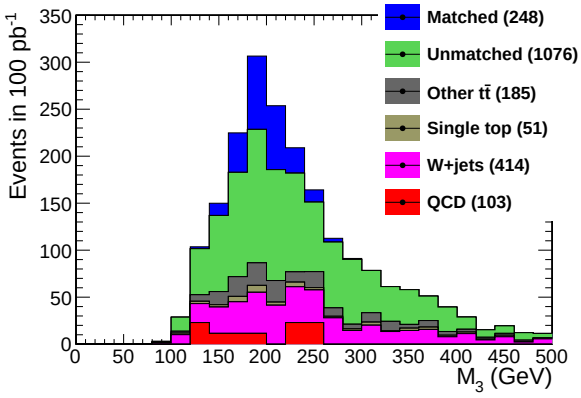
(a) No pruning



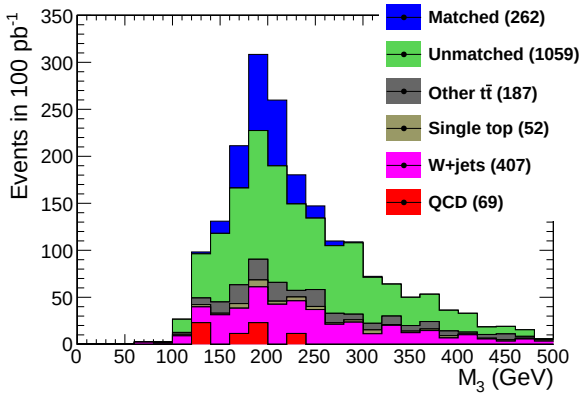
(b) PTM pruning



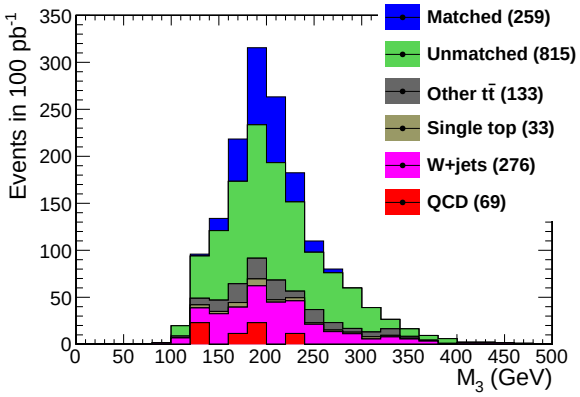
(c) Angle pruning



(d) m_W pruning

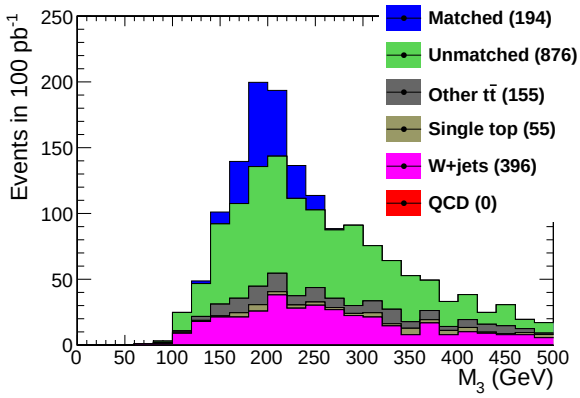


(e) PTM and angle pruning

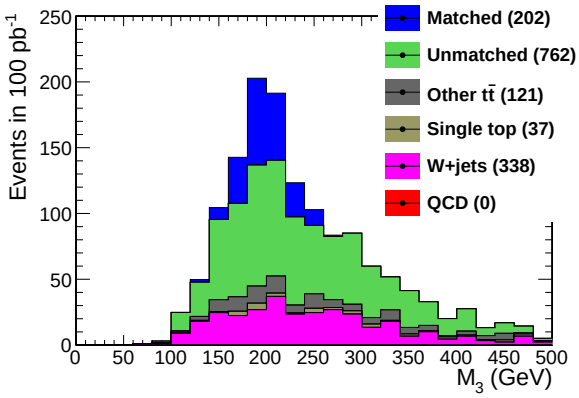


(f) All three pruning methods applied.

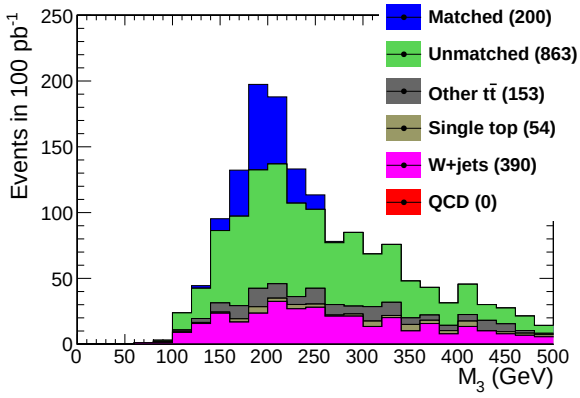
Figure A.3: Invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, after pruning. The plots correspond to the loose selection and the c_2 criterion has been used to select the hadronic-side jets.



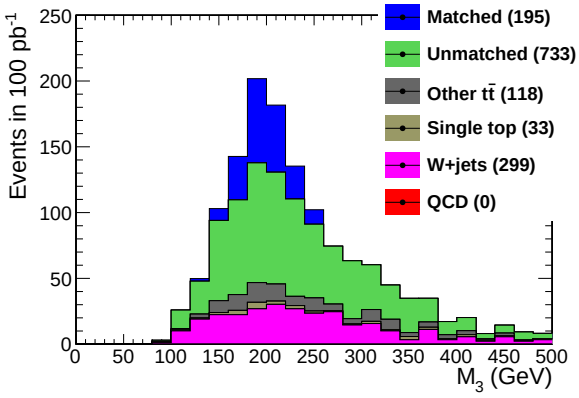
(a) No pruning



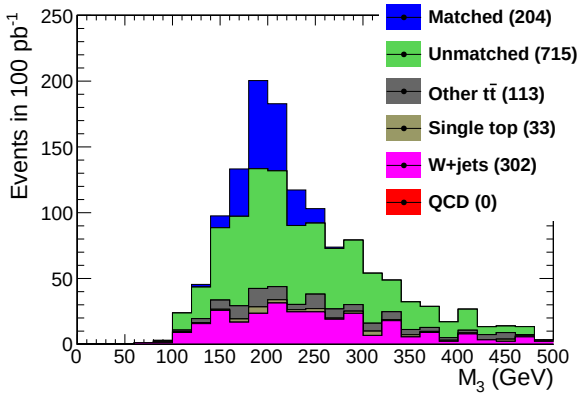
(b) PTM pruning



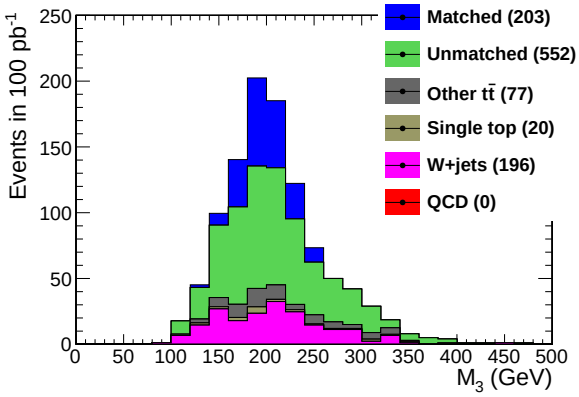
(c) Angle pruning



(d) m_W pruning

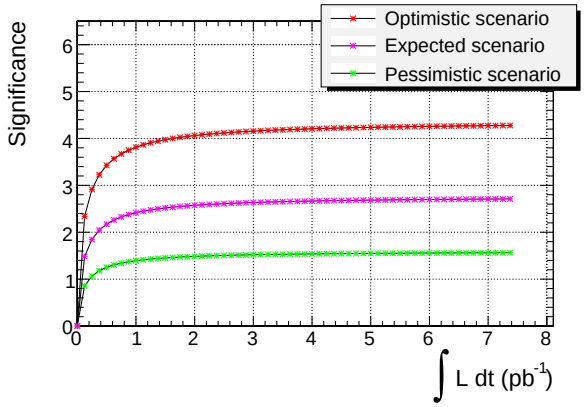


(e) PTM and angle pruning

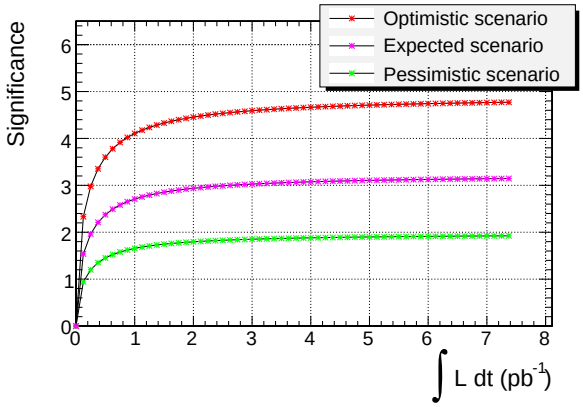


(f) All three pruning methods applied

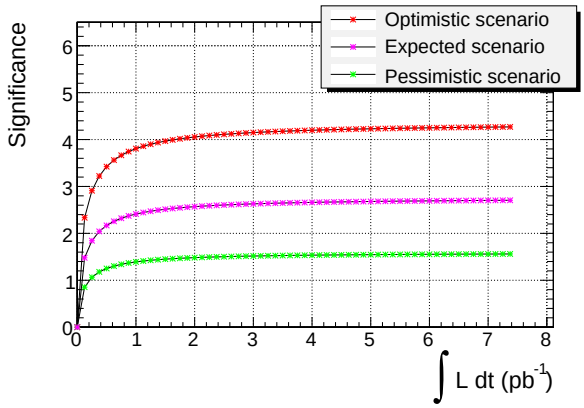
Figure A.4: Invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, after pruning. The plots correspond to the tight selection and the c_2 criterion has been used to select the hadronic-side jets.



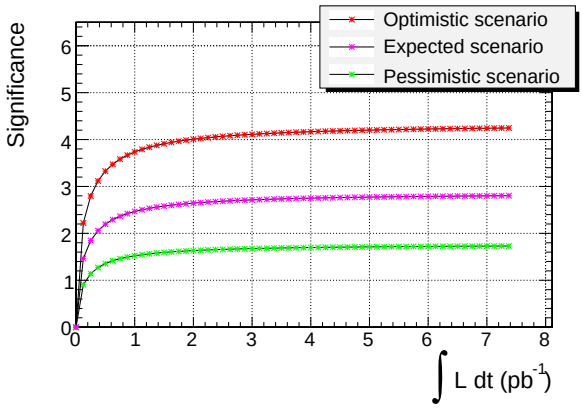
(a) No pruning



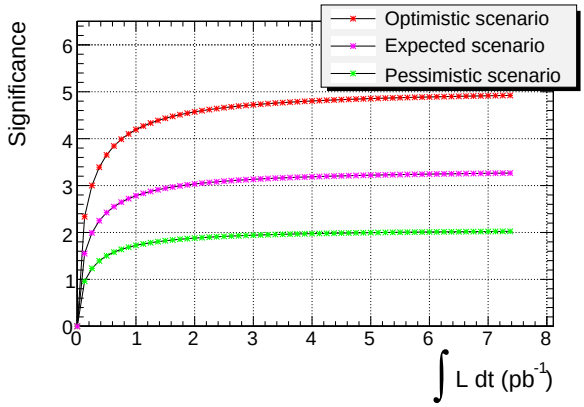
(b) PTM pruning



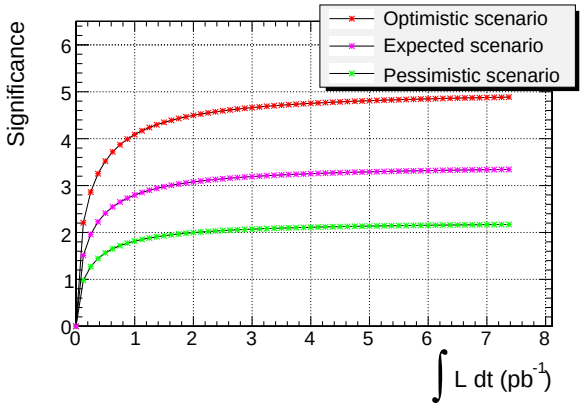
(c) Angle pruning



(d) m_W pruning

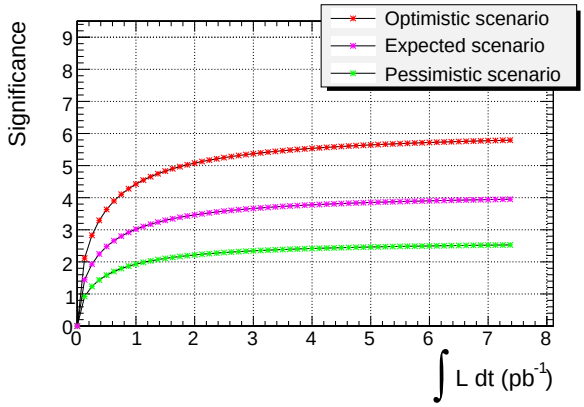


(e) PTM and angle pruning

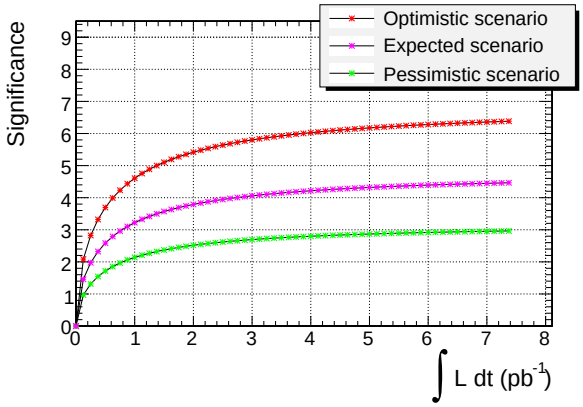


(f) All three pruning methods applied.

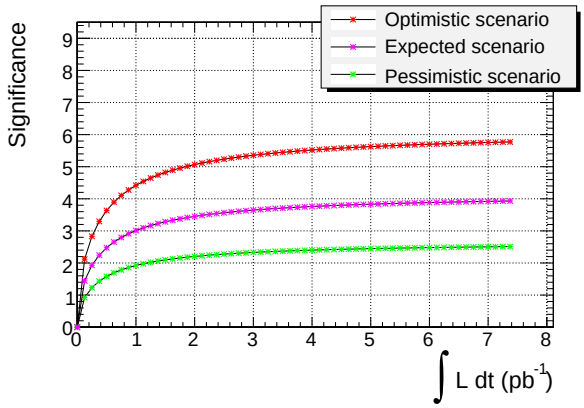
Figure A.5: Significance vs Integrated Luminosity. The selection requires a loose muon. Different combinations of solution pruning methods are displayed.



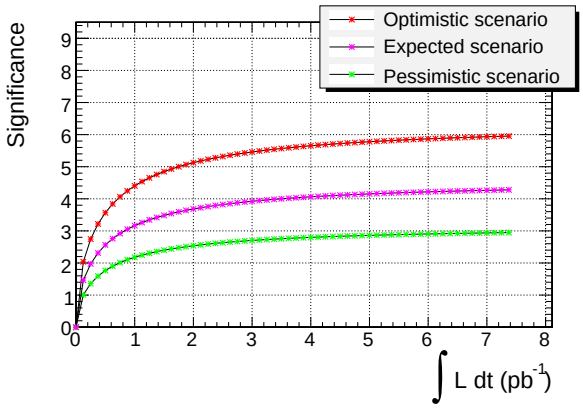
(a) No pruning



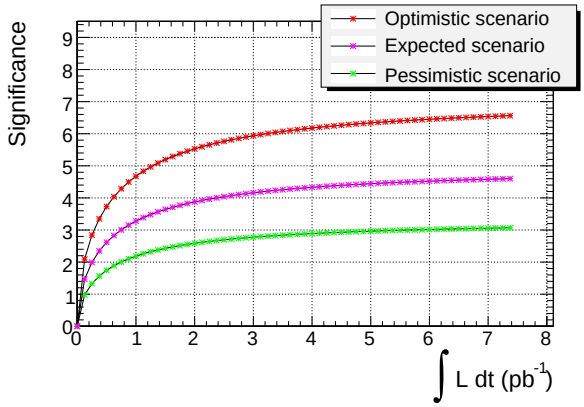
(b) PTM pruning



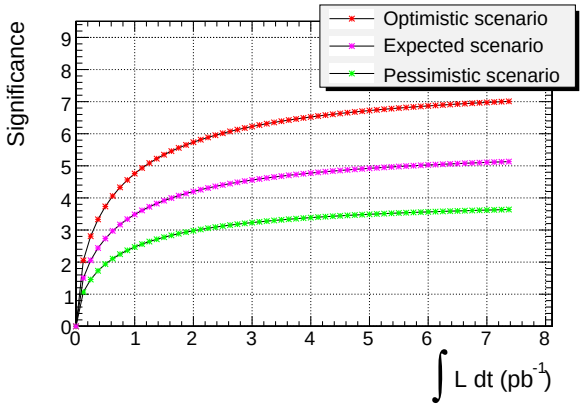
(c) Angle pruning



(d) m_W pruning

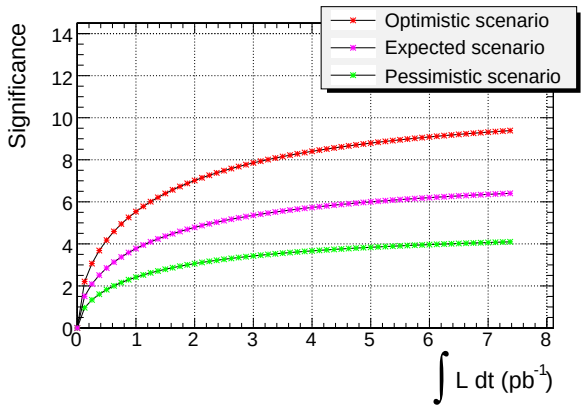


(e) PTM and angle pruning

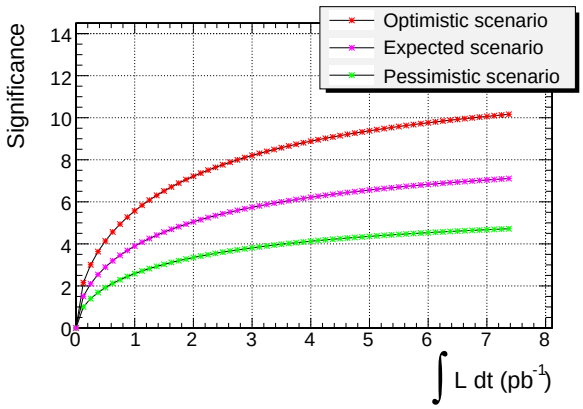


(f) All three pruning methods applied.

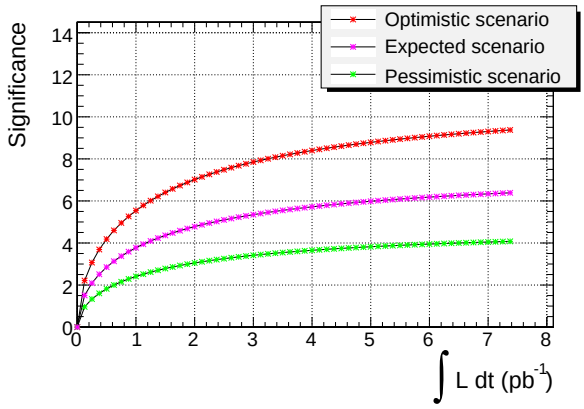
Figure A.6: Significance vs Integrated Luminosity. The selection requires a tight muon. Different combinations of solution pruning methods are displayed.



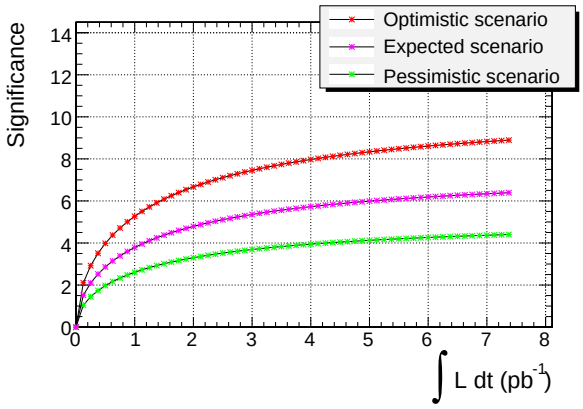
(a) No pruning



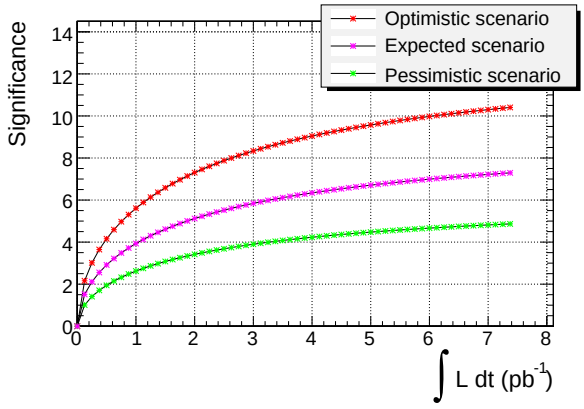
(b) PTM pruning



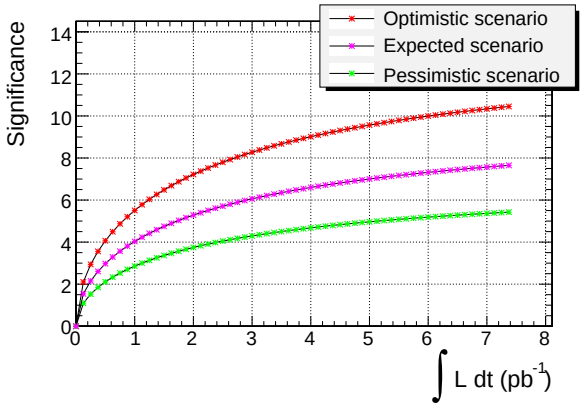
(c) Angle pruning



(d) m_W



(e) PTM and angle pruning



(f) All three pruning methods applied.

Figure A.7: Significance vs Integrated Luminosity. The jet energy scale uncertainty considered is 5%. The selection requires a tight muon. Different combinations of solution pruning methods are displayed.

Appendix B

Notes on the maximum likelihood fit

This appendix contains a few notes on the maximum likelihood fit and its application on the cross-section measurement. It begins with a brief description of the method of maximum likelihood. It continues with a note on the Monte Carlo studies used to confirm the stability of the fit. Finally, it deals with the topic of selecting the bin size for the fit and the possibility of replacing it for an unbinned likelihood fit.

B.1 The method of maximum likelihood

Suppose there is a random variable x, described by a known probability density function (PDF) $f(x;\vec{\theta})$, where $\vec{\theta}$ is a vector of m undetermined parameters that we wish to estimate. If we make n independent measurements of the variable, we define the *likelihood* of getting a result $(x_1, ..., x_n)$ as the product of the PDF values corresponding to these results:

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i;\vec{\theta}) \tag{B.1}$$

In the case where the number of measurements of x, n, is itself a Poisson random variable with mean ν, we need to multiply with the Poisson probability factor:

$$L(\nu,\vec{\theta}) = \frac{\nu^n}{n!}e^{-\nu}\prod_{i=1}^{n} f(x_i;\vec{\theta}) = \frac{e^{-\nu}}{n!}\prod_{i=1}^{n}\nu f(x_i;\vec{\theta}) \tag{B.2}$$

This is the *extended likelihood* function.

The general idea of the method of maximum likelihood is to estimate the parameters $\vec{\theta}$ by calculating the values $\hat{\vec{\theta}}$ which maximize the likelihood function for the available sample, $(x_1, ..., x_n)$. The goal is therefore to solve the set of m equations:

$$\frac{\partial L(\vec{\theta})}{\partial\theta_i} = 0, \quad i = 1, ..., m \tag{B.3}$$

It is often convenient to determine $\hat{\theta}$ by maximizing the logarithm of the likelihood function. This is allowed because the maximum of any function and that of its logarithm is the same.

$$\log L(\nu,\vec{\theta}) = -\nu(\vec{\theta}) - \log n! + \sum_{i=1}^{n}\log(\nu f(x_i;\vec{\theta})) \tag{B.4}$$

This is the *log-likelihood* function. The term $\log n!$ does not depend on $\vec{\theta}$, it is therefore usually

left out.

$$\log L(\nu, \vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^{n} \log(\nu f(x_i; \vec{\theta})) \tag{B.5}$$

B.1.1 Application to the cross-section measurement

In order to use the method of maximum likelihood for the purpose of the cross-section measurement, we formulate the problem as follows. Each of the n events passing the selection cuts is treated as an independent measurement of a variable x, which can be any measurable quantity (e.g. the jet multiplicity). We can obtain the signal and background PDFs, $f_s(x)$ and $f_b(x)$, from the Monte Carlo simulation but, in order to get the total PDF, we need to sum the two in the correct proportion. We thus write:

$$f(x) = \frac{\nu_s}{\nu_s + \nu_b} f_s(x) + \frac{\nu_b}{\nu_s + \nu_b} f_b(x) \tag{B.6}$$

Where the number of signal and background events are both Poisson variables and ν_s, ν_b are their mean values that we need to estimate. The extended log-likelihood function that we have to maximize now becomes

$$\log L(\nu_s, \nu_b) = -\nu_s - \nu_b + \sum_{i=1}^{n} \log(\nu_s f_s(x_i) + \nu_b f_b(x_i)) \tag{B.7}$$

since $\nu = \nu_s + \nu_b$. In the case of binned data, the principle is the same, but the sum is over the N bins, instead of the n individual measurements:

$$\log L(\nu_s, \nu_b) = -\nu_s - \nu_b + \sum_{i=1}^{N} \log(\nu_s \nu_{s,i} + \nu_b \nu_{b,i}) \tag{B.8}$$

Where $\nu_{s,i}, \nu_{b,i}$ are the expected signal and background entries in bin i:

$$\nu_{s/b,i} = \nu_{s/b} \int_{x_i^{min}}^{x_i^{max}} f_{s/b}(x) dx. \tag{B.9}$$

The advantage of the maximum likelihood approach is that we do not have to rely on the Monte Carlo prediction for the background cross-section to obtain the number of signal events, as we did for the event counting method. We only have to know the shape of the background and signal distributions to make an estimate. The accuracy of this estimate strongly depends on the choice of the variable x - the larger the difference between the shapes of the signal and background distributions, the easier it is to determine the signal and background proportions from a finite data sample.

B.2 Statistical properties of a maximum likelihood fit

Suppose we find a suitable variable x, make a histogram of that variable from our sample and perform the maximization of the $\log L$ function (maximum likelihood fit). The result is an estimate $\hat{n}_s$ of the mean number of signal events. It is useful to study the statistical properties of the fit, to make sure that it is stable and that we have a reliable estimate of the uncertainty of the result. It is thus useful to know how n_s is distributed. The usual way to achieve this is using Monte Carlo. When working with real data, we tune our Monte Carlo so as to fix the mean to the result of the maximum likelihood fit. We then use it to produce a large number of fake data samples and perform the maximum likelihood fit on each of them. If our estimated

values of $\hat{n}_s$ are not biased, then we will find that they are distributed around the expected value of signal events in an approximately Gaussian shape. This property of the maximum likelihood estimators is referred to as *asymptotic normality*. The standard deviation of our results can be taken as the statistical uncertainty of our maximum likelihood fit.

B.2.1 The pull distribution

A Monte Carlo study such as the one described above allows us not only to estimate the uncertainty of our result, but also to check that our fit is stable and unbiased. In order to do that, we plot the *pull* distribution. In the case of an extended ML fit, where the "true" value of the signal events is allowed to follow a Poisson distribution around the value ν_s, the pull is defined as the difference of a fit result, n_s, from ν_s divided by the error of the fit as estimated by MINUIT.

$$g \equiv \frac{n_s - \nu_s}{(\delta n_s)_{fit}} \tag{B.10}$$

An unbiased fit is expected to result in a Gaussian pull distribution, with a mean of zero and a standard deviation of one. The pull distribution can provide useful information on the fit properties, as it can be sensitive to different effects. For example, a pull distribution that does not have a mean of zero indicates a systematic bias affecting the fit and causing a tendency to overestimate or underestimate the parameter. A pull distribution that has $\sigma > 1$ indicates that the fit error is underestimated.

The distributions appearing on Fig. B.2(d),B.3(d) and B.4(d) are indicative of a good fit behaviour.

B.3 Choosing the bin width

An important consideration when extracting the signal and background PDFs from the histograms is the number of bins. It can be seen on Fig. B.1, that the narrow mass peak is more clearly distinguishable when the number of bins is larger. This positive effect is countered by the limited size of the Monte Carlo sample, which leads to significant statistical fluctuations. This is fine for data, but limited statistics in the Monte Carlo templates can be a problem. The effect on the background distribution is stronger, as the simulated samples used are smaller. To increase the available statistics per bin, it is necessary to decrease the number of bins (Fig. B.1(c), B.1(d) and B.1(e),B.1(f)) at the cost of reducing our sensitivity to the peak and thus the power of the maximum likelihood fit.

We sample the distributions obtained from the whole Monte Carlo in order to create a "data" sample corresponding to 10pb^{-1} and perform an extended ML fit of the 100pb^{-1} shapes to this "data". The true number of signal and background events in the "data" are allowed to vary following a Poisson distribution around the expected value as estimated from the Monte Carlo. As an interface to the MINUIT [46] minimization library we use the RooFit toolkit [47] within the ROOT data analysis framework [48], which helps automate the necessary tasks. Example fits with different binning can be found on Fig. B.2, B.3 and B.4. Each example fit is accompanied by the distributions (obtained by one thousand similar MC experiments) of:

- The number of signal events estimated from the fit, n_s.
- The fit error (as calculated by MINUIT), δn_s.
- The pull, g.

Comparing the two distributions for different bin widths clearly shows the impact of an increased bin width on the precision of the measurement. The fit error increases and the distribution of

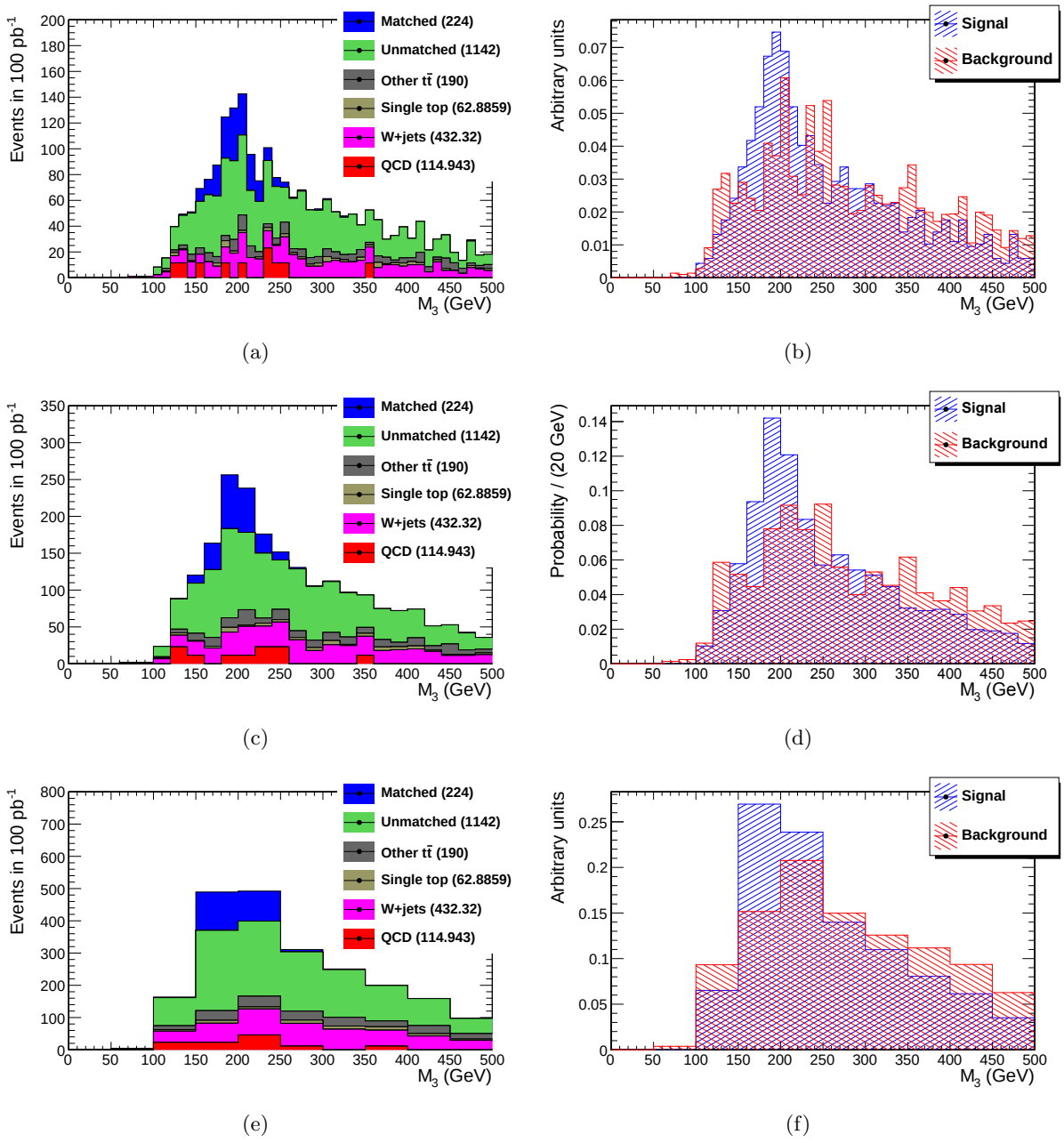


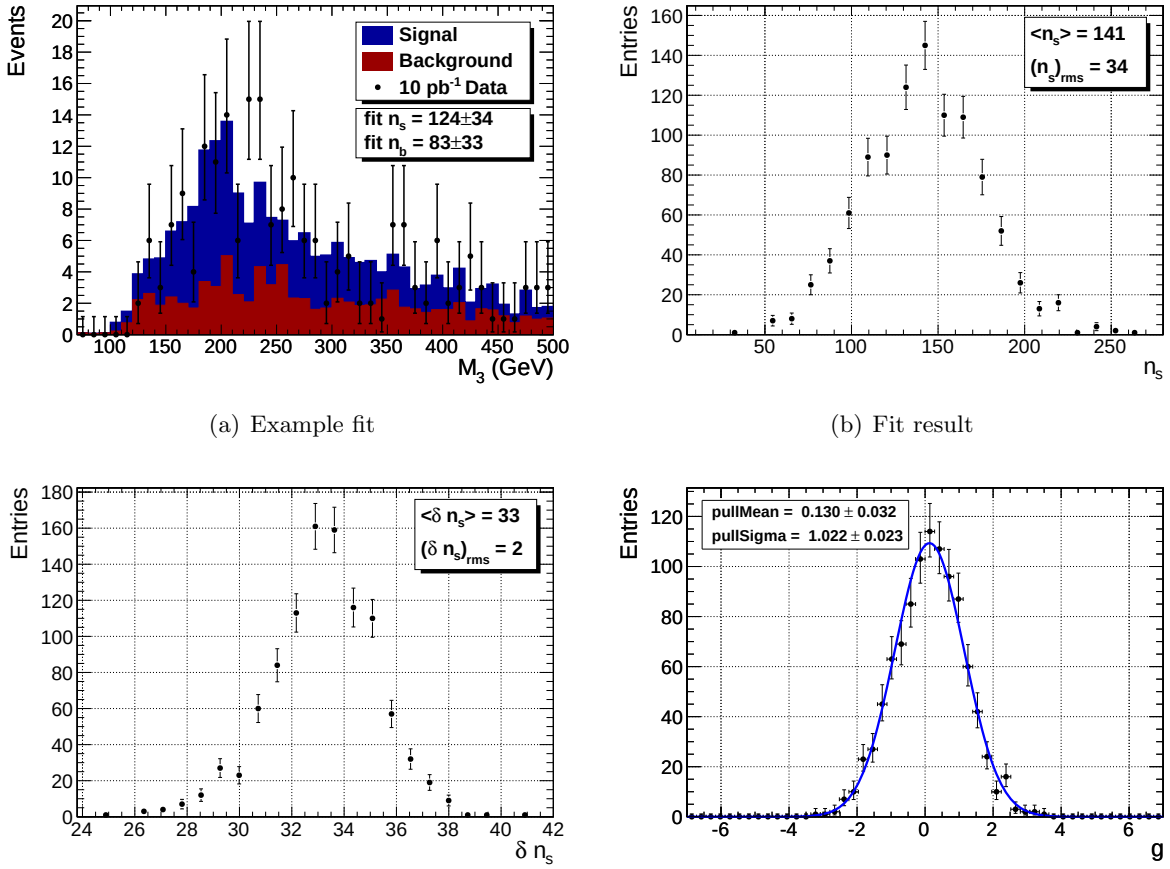
Figure B.1: The figure shows the effect a different binning has on the distribution of the invariant mass of the three hadronic-side jets obtained from the MC. The selection strategy here corresponds to a "loose" muon and a jet-parton assignment by the c_1 criterion with no solution pruning (see text). The stack plots of the different types of signal and background events are displayed on the left and the normalized total signal and background shapes are displayed on the right. The plots with $N_{bins} = 50$ (above) and $N_{bins} = 25$ (middle) show a much clearer top mass peak but are strongly affected by statistical fluctuations, whereas those with $N_{bins} = 10$ (below) are almost free of statistical fluctuations but include the whole peak in only two bins.

the estimated signal events widens as the bin width increases. However, as explained above, this does not imply that we can use an arbitrarily small bin width and increase the fit precision. The reason is that the shapes from which the MC experiment "data" is sampled are the same shapes used to fit this "data". This means that the statistical fluctuations affecting the PDFs used also affect the "data" and thus do not affect the fit. In reality, we would fit our MC shapes to data produced by the "natural PDFs" and we cannot afford large statistical fluctuations in them. In order to perform the real measurement with the amount of MC data available to this analysis, a large bin width would have to be chosen, or the fit could become less precise and even unstable. The error shown on Fig. B.4 is therefore a more realistic estimate of the uncertainty that the real measurement would have in the absence of larger MC samples.

B.3.1 Using continuous PDFs

One way to address the binning-related problem described above is by trying to remove the statistical fluctuations by fitting a suitable PDF to the MC distribution. This has the disadvantage that some arbitrariness is introduced through the choice of fitting functions to discribe the signal and background shapes. We can however use a goodness-of-fit test to ensure that whatever PDF we choose is compatible with the shapes of the Monte Carlo. The PDF fits for the loose selection, with c_1 jet-parton assignment and no solution pruning and the signal and background shapes extracted can be seen on Fig B.5.

After the signal and background PDFs have been determined in this way, we can sample them to make a Monte Carlo study as described above. The results are shown on Fig. B.6.



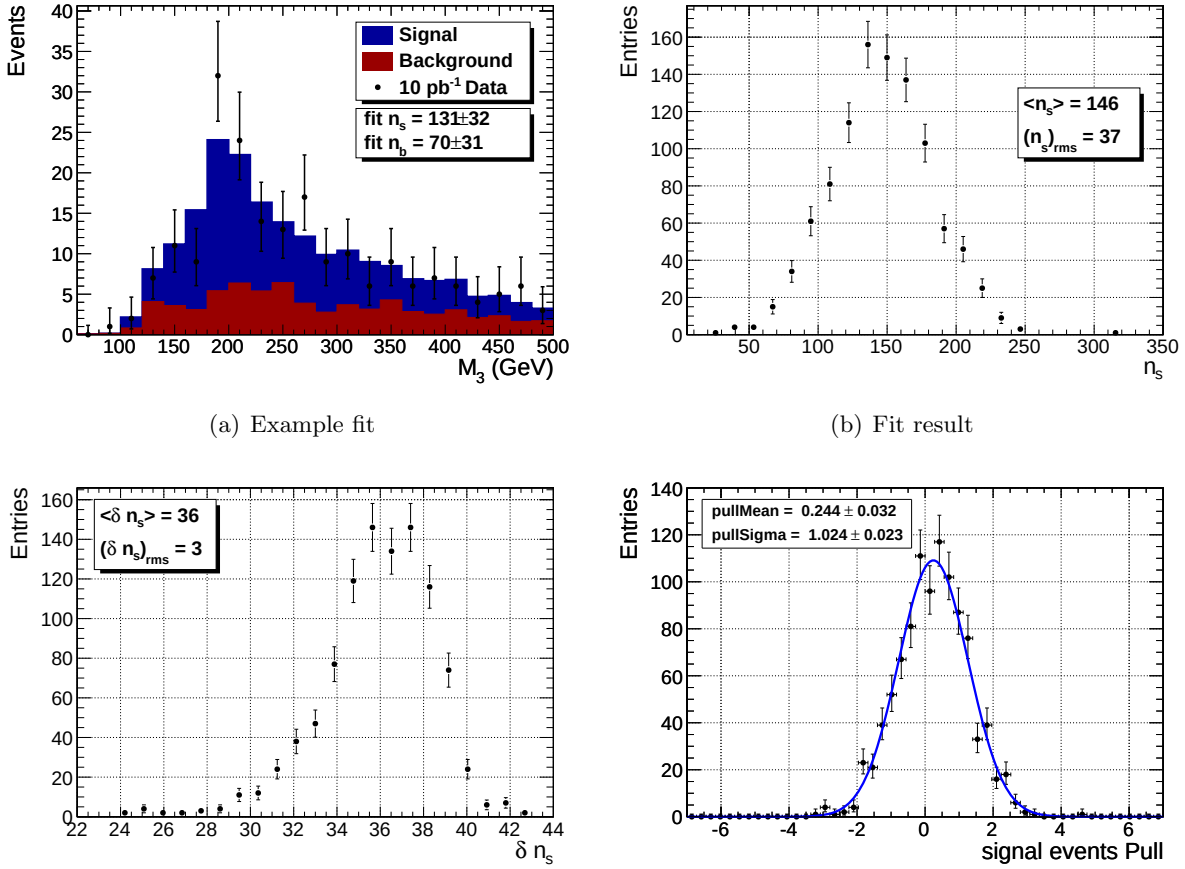
(a) Example fit

(b) Fit result

(c) Error

(d) Pull

Figure B.2: ML fit of the invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, M_3. The plots correspond to the loose selection, the c_1 criterion has been used for the jet-parton assignment and no solution pruning has been applied (see text). An example fit on toy MC data (equivalent to 10pb^{-1}) is shown in the top-left figure. The distributions of the estimated number of signal events (top-left), the fit error (bottom-left) and the pull (bottom-right) obtained from one thousand toy MC experiments are also shown. The bin width is 10GeV.



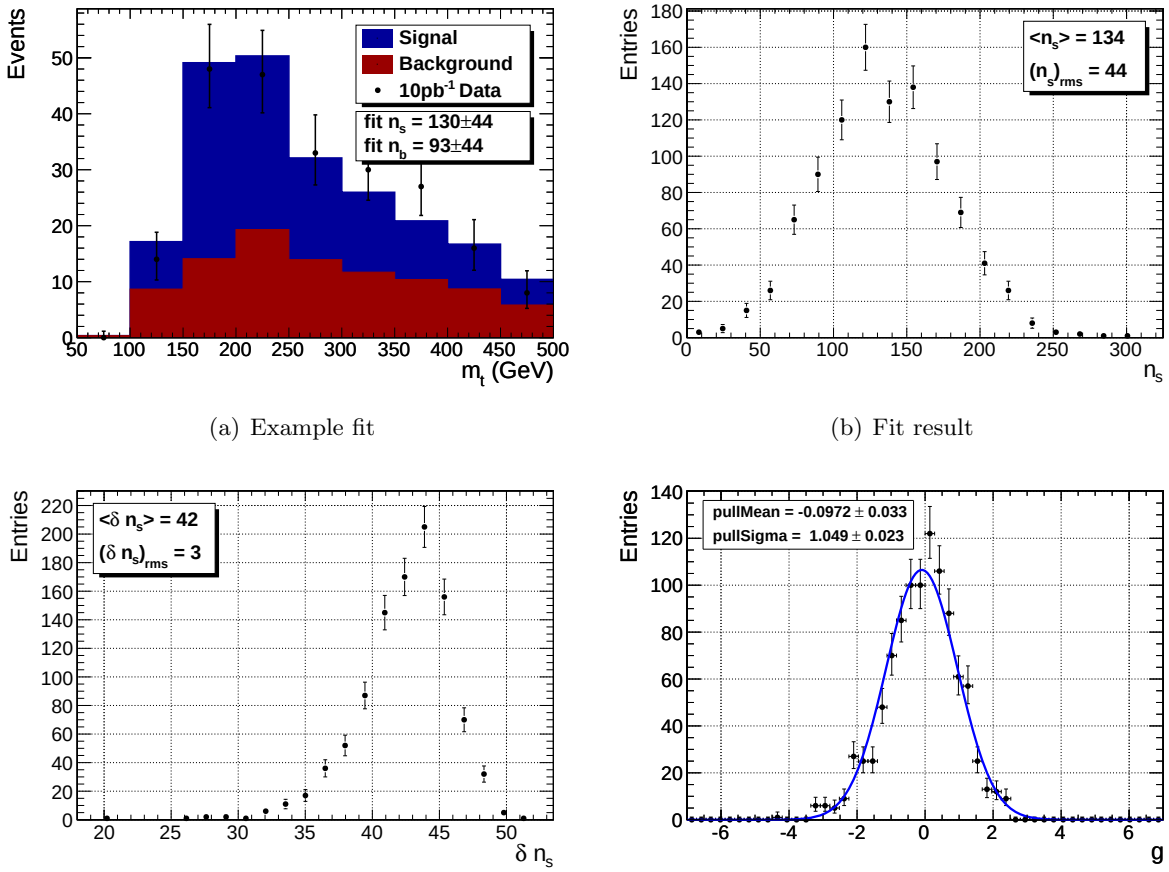
(a) Example fit

(b) Fit result

(c) Error

(d) Pull

Figure B.3: ML fit of the invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, M_3. The plots correspond to the loose selection, the c_1 criterion has been used for the jet-parton assignment and no solution pruning has been applied (see text). An example fit on toy MC data (equivalent to 10pb^{-1}) is shown in the top-left figure. The distributions of the estimated number of signal events (top-left), the fit error (bottom-left) and the pull (bottom-right) obtained from one thousand toy MC experiments are also shown. The bin width is 20GeV.



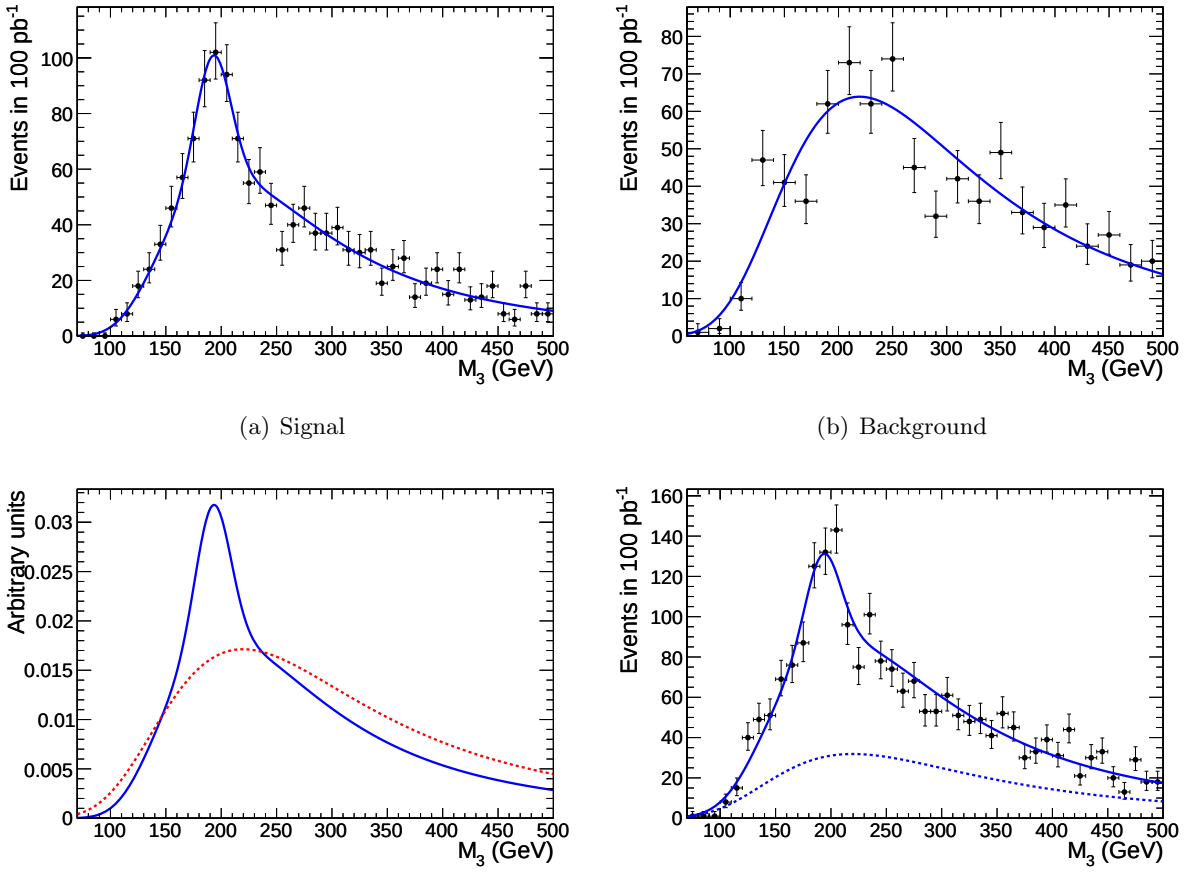
(a) Example fit

(b) Fit result

(c) Error

(d) Pull

Figure B.4: ML fit of the invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, M_3. The plots correspond to the loose selection, the c_1 criterion has been used for the jet-parton assignment and no solution pruning has been applied (see text). An example fit on toy MC data (equivalent to 10pb^{-1}) is shown in the top-left figure. The distributions of the estimated number of signal events (top-left), the fit error (bottom-left) and the pull (bottom-right) obtained from one thousand toy MC experiments are also shown. The bin width is 50GeV.



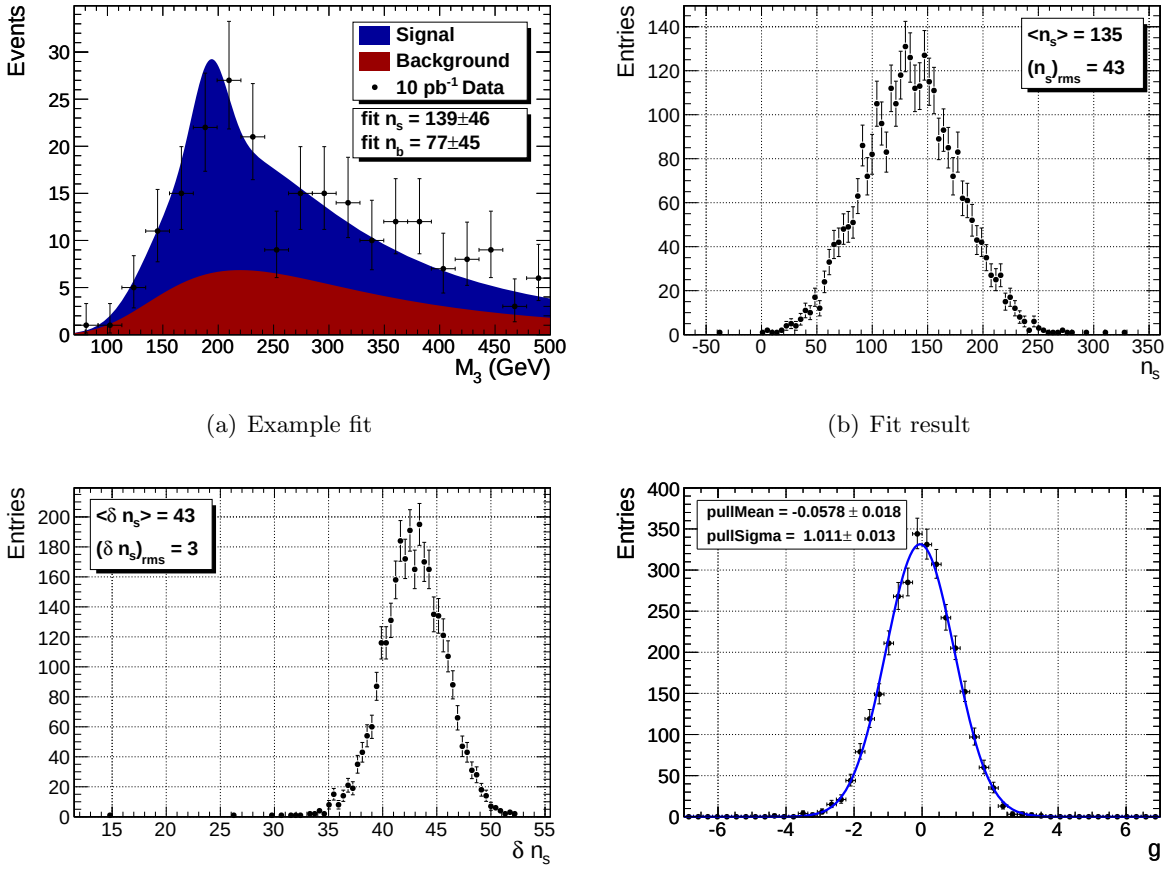
(a) Signal

(b) Background

(c) Normalized shapes

(d) Total

Figure B.5: ML fits to determine the signal and background PDFs. The plots correspond to the loose selection, the c_1 criterion has been used for the jet-parton assignment and no solution pruning have been applied (see text). The PDF used for the signal (left) is the sum of a Landau and a gaussian distribution and the shape used for the background (right) is a Landau distribution. The shapes extracted from the fits are shown normalized (below left) and stacked in the correct proportion with the whole MC sample as "data" points (below right).



(a) Example fit

(b) Fit result

(c) Error

(d) Pull

Figure B.6: Unbinned ML fit of the invariant mass of the three jets assigned to the partons of the hadronically decaying side of the $t\bar{t}$ system, M_3. The plots correspond to the loose selection, the c_1 criterion has been used for the jet-parton assignment and no solution pruning have been applied (see text). An example fit on toy MC data (equivalent to 10pb^{-1}) is shown in the top-left figure. The distributions of the estimated number of signal events (top-left), the fit error (bottom-left) and the pull (bottom-right) obtained from three thousand toy MC experiments are also shown.

Bibliography

[1] S. W. Herb et al. Observation of a dimuon resonance at 9.5-GeV in 400-GeV proton - nucleus collisions. *Phys. Rev. Lett.*, 39:252–255, 1977.

[2] B. Adeva et al. NEW PARTICLE SEARCHES. *Phys. Lett.*, B152:439, 1985.

[3] F. Abe et al. A search for the top quark in the reaction $\bar{p}p \to e + \text{jets}$ at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 64:142, 1990.

[4] D. Decamp et al. A SEARCH FOR NEW QUARKS AND LEPTONS FROM Z0 DECAY. *Phys. Lett.*, B236:511, 1990.

[5] C. Albajar et al. SEARCH FOR NEW HEAVY QUARKS IN PROTON - ANTI-PROTON COLLISIONS AT s**(1/2) = 0.63-TeV. *Z. Phys.*, C48:1–12, 1990.

[6] T. Akesson et al. SEARCH FOR TOP QUARK PRODUCTION AT THE CERN anti-p p COLLIDER. *Z. Phys.*, C46:179, 1990.

[7] F. Abe et al. A Lower limit on the top quark mass from events with two leptons in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 68:447–451, 1992.

[8] S. Abachi et al. Search for the top quark in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 72:2138–2142, 1994.

[9] F. Abe et al. Evidence for top quark production in $\bar{p}p$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 73:225–231, 1994.

[10] B. Jacobsen. Top mass from electroweak measurements. Talk given at 29th Rencontres de Moriond: QCD and High Energy Hadronic Interactions, Meribel les Allues, France, 19-26 Mar 1994.

[11] F. Abe et al. Observation of top quark production in $\bar{p}p$ collisions. *Phys. Rev. Lett.*, 74:2626–2631, 1995.

[12] S. Abachi et al. Observation of the top quark. *Phys. Rev. Lett.*, 74:2632–2637, 1995.

[13] Scott Willenbrock. The standard model and the top quark. ((U)). 2002.

[14] Matteo Cacciari, Stefano Frixione, Michelangelo M. Mangano, Paolo Nason, and Giovanni Ridolfi. Updated predictions for the total production cross sections of top and of heavier quark pairs at the Tevatron and at the LHC. 2008.

[15] C. Amsler et al. Review of particle physics. *Phys. Lett.*, B667:1, 2008.

[16] J. A. Aguilar-Saavedra. Top flavour-changing neutral interactions: Theoretical expectations and experimental detection. *Acta Phys. Polon.*, B35:2695–2710, 2004.

[17] The LHC webpage. `http://lhc.web.cern.ch/lhc`.

[18] Tao Han. Collider phenomenology: Basic knowledge and techniques. 2005.

[19] CMS, the Compact Muon Solenoid: Technical proposal. CERN-LHCC-94-38.

[20] R. Adolphi et al. The CMS experiment at the CERN LHC. *JINST*, 0803:S08004, 2008.

[21] W. W. Armstrong et al. ATLAS: Technical proposal for a general-purpose p p experiment at the Large Hadron Collider at CERN. CERN-LHCC-94-43.

[22] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.

[23] ALICE: Technical proposal for a large ion collider experiment at the CERN LHC. CERN-LHCC-95-71.

[24] K. Aamodt et al. The ALICE experiment at the CERN LHC. *JINST*, 0803:S08002, 2008.

[25] S. Amato et al. LHCb technical proposal. CERN-LHCC-98-04.

[26] A. Augusto Alves et al. The LHCb Detector at the LHC. *JINST*, 3:S08005, 2008.

[27] TOTEM: Total cross section, elastic scattering and diffraction dissociation at the LHC: Technical Proposal. CERN-LHCC-99-07.

[28] G. Anelli et al. The TOTEM experiment at the CERN Large Hadron Collider. *JINST*, 3:S08007, 2008.

[29] O. Adriani et al. Technnical proposal for the CERN LHCf experiment: Measurement of photons and neutral pions in the very forward region of LHC. CERN-LHCC-2005-032.

[30] O. Adriani et al. The LHCf detector at the CERN Large Hadron Collider. *JINST*, 3:S08006, 2008.

[31] The CMS website. `http://cms.web.cern.ch/cms`.

[32] D. Acosta et al. CMS technical design report, volume I: Detector perfomance and software. CERN-LHCC-2006-001.

[33] (ed.) Sphicas, P. CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger. CERN-LHCC-2002-026.

[34] D. Acosta et al. CMS technical design report, volume II: Physics performance. *J. Phys.*, G34:995–1579, 2007.

[35] Michelangelo L. Mangano, Mauro Moretti, Fulvio Piccinini, Roberto Pittau, and Antonio D. Polosa. ALPGEN, a generator for hard multiparton processes in hadronic collisions. *JHEP*, 07:001, 2003.

[36] Torbjorn Sjostrand, Stephen Mrenna, and Peter Skands. PYTHIA 6.4 physics and manual. *JHEP*, 05:026, 2006.

[37] Fabio Maltoni and Tim Stelzer. MadEvent: Automatic event generation with MadGraph. *JHEP*, 02:027, 2003.

[38] V. N. Ivanchenko. Geant4 toolkit for simulation of HEP experiments. *Nucl. Instrum. Meth.*, A502:666–668, 2003.

[39] N. Amapane et al. Volume-based representation of the magnetic field. Prepared for Computing in High-Energy Physics (CHEP '04), Interlaken, Switzerland, 27 Sep - 1 Oct 2004.

[40] Sunanda Banerjee. Readiness of CMS Simulation Towards LHC Startup. Presented at International Conference on Computing in High Energy and Nuclear Physics (CHEP 07), Victoria, BC, Canada, 2-7 Sep 2007.

[41] D. Acosta et al. *CMS technical design report, volume I, section 2.6: Fast simulation p55.* CERN-LHCC-2006-001.

[42] Andrea Giammanco and Andrea Perrotta. Fast simulations of the ATLAS and CMS experiments at LHC. CERN-CMS-CR-2007-010.

[43] Guenter Grindhammer and S. Peters. The parameterized simulation of electromagnetic showers in homogeneous and sampling calorimeters. 1993.

[44] Lars Sonnenschein. The t anti-t production in p p collisions at S**2 = 14- TeV. PITHA-01-04.

[45] Suggested by D. Hits in private correspondence.

[46] F. James and M. Roos. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput. Phys. Commun.*, 10:343–367, 1975.

[47] W. Verkerke and D Kirkby. Roofit users manual. `http://root.cern.ch/root/doc/RooFit_Users_Manual_2.91-33.pdf`.

[48] R. Brun et al. The root system home page. `http://root.cern.ch`.

[49] Alexander Schmidt. Beauty production and identification at cms. *Nuclear Physics B (Proceedings Supplements)*, 187:216–223, 2009.

[50] The Particle Flow Group. Particle flow reconstruction of jets, met and taus in cms. *CMS Note in preparation (PFT-09-001)*, 2009.

List of Figures

List of Tables