ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Measurement of the differential $t\bar{t}$ production cross section at 13 TeV with CMS in the all-hadronic boosted regime

Relatore:
Prof. Andrea Castro

Presentata da:
Federico Celli

Anno Accademico 2016/2017

A Chiara e Davide

Sommario

È qui presentata una misura inclusiva e differenziale della sezione d'urto di produzione $t\bar{t}$ nel canale completamente adronico, usando dati provenienti da collisioni pp a 13 TeV, raccolti dal rivelatore CMS ad LHC nel 2016, corrispondenti a una luminosità integrata di 37 fb^{-1}. L'analisi è stata svolta nel regime "boosted", caratterizzato da un alto momento trasverso per il quark top, i cui prodotti di decadimento sono ricostruiti, a causa dell'elevato boost di Lorentz, in jet di ampia larghezza. La sezione d'urto inclusiva di produzione $t\bar{t}$ è stata misurata essere pari a $\sigma_{t\bar{t}} = 572 \pm 14(\mathrm{stat}) \pm 118(\mathrm{syst}) \pm 14(\mathrm{lumi})$ pb, un valore sensibilmente inferiore rispetto alla predizione teorica. Una sovrastima dell'efficienza di selezione del campione Monte Carlo generato con POWHEG + PYTHIA 8 potrebbe spiegare questo effetto.

La sezione d'urto differenziale al detector-level è poi misurata in funzione di alcune variabili di interesse ed è confrontata alle predizioni teoriche.

Una procedura di "unfolding" è infine svolta al fine di rimuovere gli effetti del rivelatore ed estrapolare la sezione d'urto differenziale all'intero spazio delle fasi partonico.

Abstract

An inclusive and differential measurement of the $t\bar{t}$ production cross section in the all-hadronic channel is presented here, using data from 13 TeV pp collisions, collected by the CMS detector at the LHC in 2016, corresponding to an integrated luminosity of 37 fb^{-1}. The analysis has been performed in the boosted regime, characterized by high-p_T top quarks whose decay products are reconstructed, due to the large Lorentz boost, into two wide jets. The inclusive $t\bar{t}$ production cross section has been measured to be $\sigma_{t\bar{t}} = 572 \pm 14(\text{stat}) \pm 118(\text{syst}) \pm 14(\text{lumi})$ pb, a value quite lower than the theoretical prediction. An overestimation of the POWHEG + PYTHIA 8 Monte Carlo selection efficiency could explain this effect.

The detector-level differential cross section is then measured as a function of some variables of interest and it is compared to the theoretical predictions.

Finally, an unfolding procedure is performed in order to remove detector effects and to extrapolate the differential cross section measurement to the full parton-level phase space.

Desidero esprimere tutta la mia gratitudine al Professor Andrea Castro, per la costante disponibilità, l'impegno e la pazienza con cui mi ha guidato in questo percorso.
Grazie alla mia famiglia e a tutti gli amici che mi hanno consigliato e sostenuto in questi anni.

Contents

Chapter 1

Introduction

The top quark is the latest discovered quark and nowadays it is one of the fundamental fermions of the Standard Model, belonging to the third generation of quarks. Being a fermion, it has an half-integer spin ($\frac{1}{2}\hbar$) and it is also subject to the electromagnetic interaction (having an electric charge, $+\frac{2}{3}e$), to the strong interaction (having a color charge) and to the weak interaction (being part, together with the bottom quark, of a weak isospin doublet). Furthermore, the top quark is the most massive fundamental particle of the Standard Model, having a mass of $173.21 \pm 0.51(\text{stat.}) \pm 0.71(\text{syst.})$ GeV [1], and many of its peculiar properties derive from this aspect.

The discovery of the top quark dates back to the 1995, when the two experiments CDF [2] and DØ [3] operating at the Tevatron announced independently the observation of a particle compatible with a new quark. The process that allowed the discovery was the $t\bar{t}$ pair production, which has a larger cross section than the processes that lead to a single top quark production. The $t\bar{t}$ production can occur at hadron accelerators like the Tevatron and the LHC, through different mechanisms, the dominant ones being shown in Fig 1.1. While the dominant production mechanism at the Tevatron was $q\bar{q}$ annihilation, at the LHC the $t\bar{t}$ pairs are mainly produced through gluon fusion processes.

The $t\bar{t}$ production inclusive cross section was measured at the Tevatron in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV (see [4, 5] and references therein) and it was found to be consistent with the theoretical estimate of $\sigma_{t\bar{t}}^{theor} = 7.16^{+0.20}_{-0.23}$ pb, obtained by perturbative QCD calculations. Being the $t\bar{t}$ cross section strongly dependent on the center-of-mass energy of the collision, the theoretical estimate for LHC collisions at 13 TeV is $\sigma_{t\bar{t}}^{theor} = 832^{+20}_{-29}(\text{scale}) \pm 35(\text{PDF} + \alpha_s)$ pb. The first uncertainty is related to different choices for the factorization and renormalization scales. The second uncertainty is associated to the choice of parton distribution functions and of the strong coupling.

The inclusive $t\bar{t}$ cross section measurement is important to check the QCD prediction and it helps to constrain some of its parameters. In fact the pair production cross section can be written as

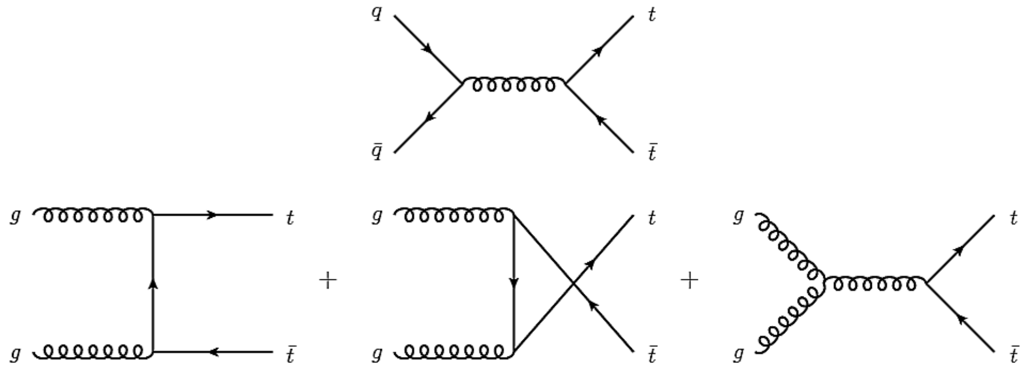


Figure 1.1: Feynamn diagrams for the dominant $t\bar{t}$ pair production mechanisms at the Tevatron (upper row) and at the LHC (lower row).

$$\sigma(pp \to t\bar{t}) = \sum_{i,j} \int dx_i f_i(x_i, \mu^2) \int dx_j f_j(x_j, \mu^2) \hat{\sigma}_{ij}(\hat{s}, \mu^2, m_t), \tag{1.1}$$

where indices i, j indicates partons, f_k are the parton distribution functions (PDFs) of gluons and light quarks, x_i are the momentum fractions of the partons, $\hat{s} = x_i x_j s$ is the center-of-mass energy of the partons, $\hat{\sigma}_{ij}$ is the cross section of the partonic process and μ is the factorization scale, related to the perturbative order of the calculations. Moreover, $t\bar{t}$ processes could be sensitive to physics beyond the Standard Model: the production of new exotic particles could manifest as an excess in the measured cross section with respect to the theory predictions.

The top quark is not a stable particle: its behavior is greatly influenced by its huge mass and its lifetime is so short that it decays well before hadronization can occur. The top quark lifetime can be estimated by using

$$\tau_t = \frac{\hbar}{\Gamma_t}, \tag{1.2}$$

where $\hbar$ is the reduced Plank constant and Γ_t is the top quark decay width, that at leading order can be computed as

$$\Gamma_t = \frac{G_F}{8\pi\sqrt{2}} m_t^3 \left(1 - \frac{m_W^2}{m_t^2}\right)^2 \left(1 + 2\frac{m_W^2}{m_t^2}\right) \approx 1.5 \text{ GeV}, \tag{1.3}$$

yielding $\tau_t \approx 5 \cdot 10^{-25}$ s. Since the typical time scale for the hadronization process is about $3 \cdot 10^{-23}$ s, mesonic or barionic states composed of top quarks have never been observed.

A t ($\bar{t}$) quark decays almost exclusively into a W^+b ($W^-\bar{b}$) pair. As a matter of facts, the top decay fraction into Wb can be computed as

$$R_{Wb} = \frac{|V_{tb}|^2}{|V_{tb}|^2 + |V_{ts}| + |V_{td}|^2} \approx 0.93, \tag{1.4}$$

where V_{tb}, V_{ts} and V_{td} are elements of the CKM matrix, which contains informations on the strength of flavor-changing weak decays. For this reason, $t\bar{t}$ events are then classifiable depending on the decay products of the W bosons and analyses can be carried independently on events belonging to different decay channels.

- Dilepton channel: both the W bosons decay into a lepton-neutrino pair ($(l, \bar{\nu}_l)$ and $(\bar{l}', \nu_{l'})$). This process has a branching ratio of 5% if only electrons and muons are considered (tau leptons are usually considered separately).
- Sigle-lepton channel: one of the W bosons decays into a lepton-neutrino pair, while the other decays hadronically into a pair of quarks. This process has a branching ratio of about 30% when considering only electrons and muons.
- All-hadronic channel: both the W bosons decay hadronically into a pair of quarks, which will then hadronize by turning into jets of particles. This process is the most probable of the three channels, having a branching ratio of about 46%.

Fig. 1.2 summarizes the the top quark decay modes and their branching ratios.

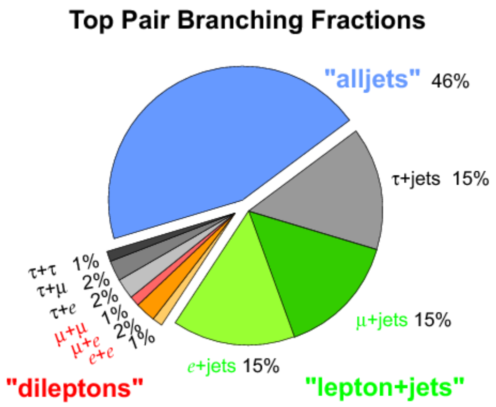


Figure 1.2: Summary of the $t\bar{t}$ decay modes and their relative branching ratio.

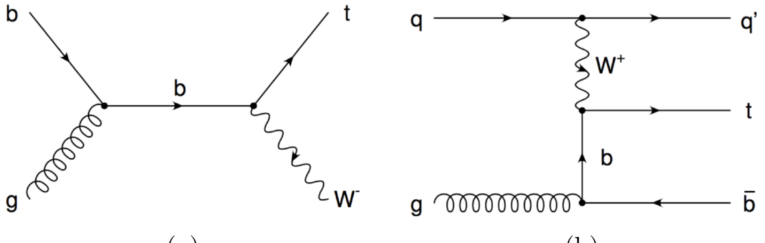
In the analysis presented here, $t\bar{t}$ events belonging to the all-hadronic channel have been selected. The use of this event topology has the important advantage of allowing a full reconstruction of the $t\bar{t}$ decay products, in contrast to the leptonic channels, where the presence of one or two neutrinos lead to an ambiguity in the event interpretation. Despite this advantage, the presence of jets in the final state leads to larger uncertainties (e.g. jet energy scale and resolution) on the measured quantities with respect to leptonic channels. Furthermore, this channel also has a very high background, mainly coming from QCD multijet events.

At the LHC, the *pp* collisions at 13 TeV can produce top quarks with very high kinetic energies. In the laboratory frame, the decay products of such quarks are collimated in the particle flight direction. If a top quark with a Lorentz boost $\gamma = E/m$ is considered, then the angular distance $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ in the pseudorapidity-azimuth space of its decay products will be approximately $\Delta R = 2/\gamma$. If the boost is high enough, a jet clustering algorithm could merge into a single *wide jet* the three jets of particles coming from the hadronic decay of the top quark. The study of the substructures of these wide jets is then useful to find the signature of a top quark decay. Events like these are called *boosted topology events*, and despite the fact that their study poses new experimental challenges, they play an important role in particle physics studies at high energy.

In this analysis, the measurements are performed by selecting only the boosted topology all-hadronic $t\bar{t}$ events. These events are thus characterized by the presence of two wide jets whose substructures suggest that they each come from a top quark decay. When defining the selection of the signal it is also important to know the possible sources of background.

The main source of background are QCD multijet events, namely QCD events with large cross section in which multiple quarks and gluons are produced, leading to the presence, in the final state, of several jets of particles. These events can easily mimic the boosted topology all-hadronic $t\bar{t}$ decay and thus an effective method for extracting the signal from this background needs to be introduced.

Along with QCD multijet events, many other subdominant background processes may disturb the boosted topology $t\bar{t}$ event selection.



(a) (b)

Figure 1.3: Feynman diagrams for the single top quark *tW* (a) and t-channel (b) processes.

In this analysis only the most important ones were considered:

- Single top quark with associated W production (tW and $\bar{t}W$);
- Single top quark (and single antiquark) t-channel process;
- W production with hadronic decay (W + jets);

- Drell-Yan process with hadronic final state (Z + jets).

The contribution to the total background coming from these processes is found to be very small when compared to the QCD multijet background. Some Feynman diagrams for the single top sub-dominant background processes are shown in Fig. 1.3.

Chapter 2

Physics at the LHC

2.1 The Large Hadron Collider

2.1.1 Generalities

The LHC is a superconducting accelerator [6, 7] and collider installed in a 27 km long circular tunnel buried about 100 m underground in the border of France and Switzerland, near the city of Geneva. The LHC was turned on September 10, 2008, and remains the latest addition to the CERN accelerator complex, the European Organization for Nuclear Research.

Nowadays LHC is the world largest and most powerful particle accelerator producing collisions between protons with a record center-of-mass energy of $\sqrt{s} = 13$ TeV, or lead ions.

Inside the accelerator, two high-energy particle beams travel at close to the speed of light before they are made to collide. The beams travel in opposite directions in separate beam pipes, two tubes kept at ultrahigh vacuum, and they are guided around the accelerator ring by a strong magnetic field maintained by superconducting electromagnets. The electromagnets are built from coils of special electric cable that operate in a superconducting state, efficiently conducting electricity without resistance or loss of energy. This requires chilling the magnets to -271.3 °C, a temperature colder than outer space. For this reason, much of the accelerator is connected to a distribution system of liquid helium, which cools the magnets.

All the controls for the accelerator, its services and technical infrastructure are housed under one roof at the CERN Control Center. From there, the beams inside the LHC are made to collide at four locations around the accelerator ring, corresponding to the positions of four particle detectors: ATLAS, CMS, ALICE and LHCb.

2.1.2 The CERN complex

LHC is not the only accelerator at CERN: the CERN complex is composed of various machines and accelerating rings which have different power. Each machine injects the particle beam into the next one, which takes over to bring the beam to a higher energy. The particles at the end of this process enter into LHC where they are accelerated to the maximum energy.
Most of the accelerators in the CERN complex have their own experimental halls, where the beams are used for experiments at different energies.
A simple schematic of the CERN complex is shown in Fig. 2.1.

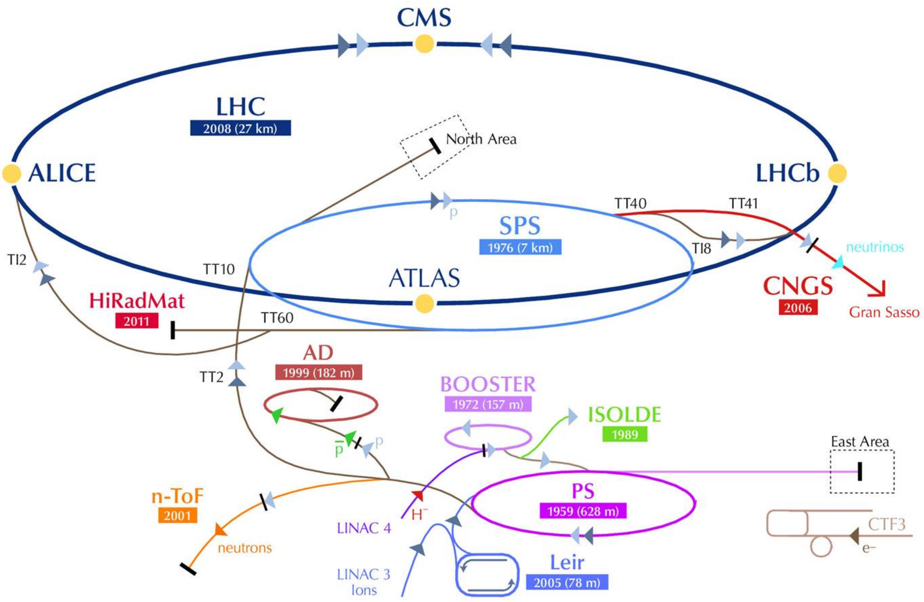


Figure 2.1: CERN accelerator complex.

The process of accelerating protons at CERN starts from a bottle of hydrogen. Protons are extracted from hydrogen atoms by stripping orbiting electrons thanks to a strong electric field. Protons are then injected into the PS Booster (PSB) at an energy of 50 MeV from Linac2. The booster accelerates them to 1.4 GeV. The beam is then fed to the Proton Synchrotron (PS) where it is accelerated to 25 GeV. Protons are then sent to the Super Proton Synchrotron (SPS) where they are accelerated to 450 GeV. They are finally transferred to the LHC (both in a clockwise and an anticlockwise direction) where they reach the maximum energy.
The complex can accelerate not exclusively protons but also lead ions, which are produced from a highly purified lead sample heated to a temperature of about 500 °C. The lead vapour is ionized by an electron current. Many different charge states are produced

with a maximum around Pb^{29+}. These ions are selected and accelerated to 4.2 MeV/u (energy per nucleon) before passing through a carbon foil, which strips most of them to Pb^{54+}.

The Pb^{54+} beam is accumulated, then accelerated to 72 MeV/u in the Low Energy Ion Ring (LEIR), which transfers it to the PS. The PS accelerates the beam to 5.9 GeV/u and sends it to the SPS after first passing it through a second foil where it is fully stripped to Pb^{82+}. The SPS accelerates it to 177 GeV/u and finally sends it to the LHC.

2.1.3 The LHC machine

Structure The LHC ring is made of eight arcs and eight "insertions". The arcs contain the dipole bending magnets, with 154 magnets in each arc. Their aim is to bend the beams using a strong magnetic field so that the particles can fly in the almost circular orbit of the LHC ring.

An insertion consists of a long straight section plus two transition regions (one at each end), the so-called "dispersion suppressors". The exact layout of the straight section depends on the specific use of the insertion: physics (beam collisions within an experiment), injection, beam dumping, beam cleaning.

A sector is defined as the part of the machine between two insertion points. The eight sectors are the working units of the LHC: the magnet installation happens sector by sector, the hardware is commissioned sector by sector and all the dipoles of a sector are connected in series and are in the same continuous cryostat. Powering of each sector is essentially independent.

An octant starts from the middle of an arc and ends in the middle of the following arc and thus spans a full insertion. Therefore, this description is more practical when we look at the use of the magnets to guide the beams into collisions or through the injection, dumping, and cleaning sections.

A simple schematic of the LHC structure is shown in Fig. 2.2.

Vacuum LHC has three vacuum systems made up to handle three different tasks: insulation vacuum for cryomagnets, insulation vacuum for the helium distribution line, beam vacuum.

Since their only aim is insulation, the first two systems do not provide a vacuum as high as the last system. The beam vacuum instead has to be very high, 10^{-13} atm (ultrahigh vacuum), because we want to avoid collisions between the beam particles and the gas in the beam pipes.

Magnets There is a large variety of magnets in the LHC, including dipoles, quadrupoles, sextupoles, octupoles, decapoles, etc. giving a total of about 9600 magnets. Each type of magnet contributes to optimizing the beam trajectory: the dipoles bend the beam in the correct direction along the LHC ring, the other multipoles focus the beam reducing

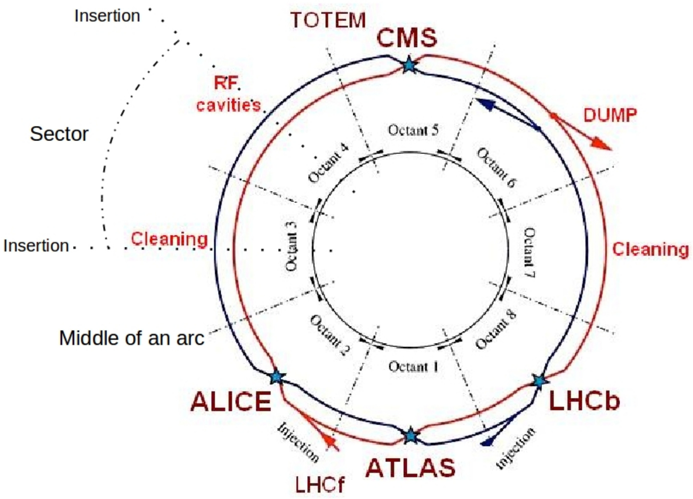


Figure 2.2: LHC layout.

its transverse section to increase the interaction probability during the collisions.
The dipoles of the LHC represented the most important technological challenge for the LHC design. Each dipole is 15 m long and weighs around 35 t. In a proton accelerator like the LHC, the maximum energy that can be achieved is directly proportional to the strength of the dipole field, given a specific acceleration circumference. At the LHC the dipole magnets are superconducting electromagnets and able to provide the very high field of 8.33 T over their length. No practical solution could have been designed using "warm" magnets instead of superconducting ones.
The LHC dipoles use coils made of niobium-titanium (NbTi) cables, which become superconducting below a temperature of 10 K (−263.2 °C), that is, they conduct electricity without resistance. In fact, the LHC will operate at 1.9 K (−271.3 °C), which is even lower than the temperature of outer space (2.7 K or −270.5 °C). A current of 11850 A flows in the dipoles, to create the high magnetic field of 8.33 T required to bend the beam.
The temperature of 1.9 K (−271.3 °C) is reached by pumping superfluid helium into the magnet systems.

Cavities The main role of the LHC cavities is to keep the proton bunches, which constitute the beam, tightly bunched to ensure high luminosity at the collision points and hence, maximize the number of collisions. Each bunch contains about 10^{11} protons and measures a few centimetres in length and a millimetre in width when far from the collision points. However, as they approach the collision points, they are squeezed to about 16 μm in width to allow for a greater chance of proton-proton collisions. Increasing the number of bunches is one of the ways to increase the instantaneous luminosity $\mathcal{L}$ in a machine, namely the number which, multiplied by the total cross section, gives the total number of collisions per unit time. At full luminosity the LHC uses a bunch spacing of 25 ns (or about 7 m) which corresponds to a frequency of 40 MHz. However, for practical reasons there are several bigger gaps in the pattern of bunches which lead to a frequency of 31.6 MHz.

The cavities also deliver radiofrequency (RF) power to the beam during acceleration to the top energy. Protons can only be accelerated when the RF field has the correct orientation when particles pass through an accelerating cavity, which happens at well specified moments during an RF cycle. The LHC will use eight cavities per beam, each delivering 2 MV (an accelerating field of 5 MV/m) at 400 MHz. The cavities will operate at 4.5 K (−268.7 °C).

2.2 The Compact Muon Solenoid

2.2.1 Generalities

The Compact Muon Solenoid (CMS) [8] is one of the six detectors installed at the LHC. The other five detectors are: A Large Ion Collider Experiment (ALICE), A Toroidal LHC ApparatuS (ATLAS), the Large Hadron Collider beauty (LHCb), the Large Hadron Collider forward (LHCf) and the TOTal Elastic and diffractive cross section Measurement (TOTEM). ALICE, ATLAS, CMS and LHCb are installed in four huge underground caverns built around the four collision points of the LHC beams.

These detectors have different research purposes in nuclear physics.

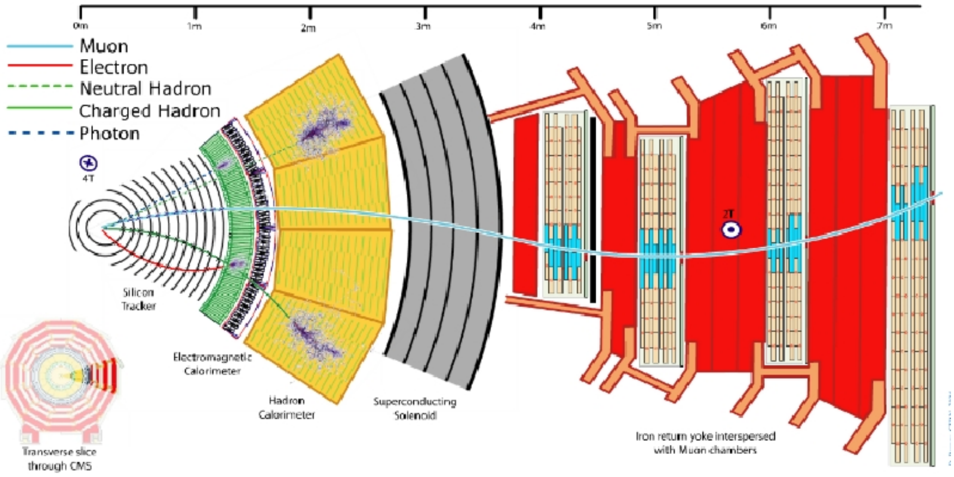
CMS is a general-purpose detector built around a huge superconducting solenoid which takes the form of a cylindrical coil of superconducting cable. It can generate a magnetic field of 4 T.

CMS was mainly designed to look for the Higgs boson, to measure its properties, and also to scrutinize various currently unproven models as supersimmetry, the existence extra dimensions, or the origin of the dark matter. CMS possesses the necessary versatility to uncover unexpected phenomena at LHC energies.

2.2.2 Structure

To work as a precise detector and achieve its goals, CMS must be able to reconstruct the collision events in the best possible way. For this reason it is composed of several sub-detectors of different type to reveal most of the particles produced in the collisions, measuring their energy and momentum.

Since a magnetic field can be used to measure the momentum of a particle through the bending of the track left in the detector, the central feature of the CMS apparatus is a



©CERN

Figure 2.3: CMS transverse section.

superconducting solenoid, of 6 m internal diameter, which provides a magnetic field of 3.8 T. Such a strong magnetic field is needed to produce a large bending power to measure precisely the momentum of high-energy charged particles like muons. This forces a choice of superconducting technology for the magnets.
Within the field volume are the silicon tracker, the crystal electromagnetic calorimeter, and the brass/scintillator hadron calorimeter. Muons are measured in gas-ionization detectors located externally to the other elements of the detector.
A simple schematic of the CMS transverse section is shown in Fig. 2.3.

The tracker The inner tracking system of CMS is designed to provide a precise and efficient measurement of the trajectories of charged particles emerging from the LHC collisions (like electrons, protons, muons), as well as a precise reconstruction of secondary vertices. It surrounds the interaction point and has a length of 5.8 m and a diameter of 2.5 m.
Being the nearest part of the detector to the collision point, the tracker will experience a huge particle flux. Therefore, a detector technology featuring high granularity and fast response is required, such that the trajectories can be identified reliably and attributed to the correct bunch crossing. However, these features imply a high power density of the on-detector electronics which in turn requires efficient cooling. This is in direct conflict with the aim of keeping to the minimum the amount of material in order to limit multiple scattering, bremsstrahlung, photon conversion and nuclear interactions. A compromise had to be found in this respect. The intense particle flux will also cause severe radiation

damage to the tracking system. The main challenge in the design of the tracking system was to develop detector components able to operate in this harsh environment for an expected lifetime of 10 years. These requirements on granularity, speed and radiation hardness lead to a tracker design entirely based on silicon detector technology.

The silicon tracker is composed of a pixel detector and strip detectors.

The pixel detector covers an area of about 1 m^2 and has 66 million pixels. When a charged particle passes through this detector it releases enough energy for electrons to be ejected from the silicon atoms of the pixels, creating electron-hole pairs. Each pixel uses an electric field to collect these charges on the surface as a small electric signal which is then amplified. In this way the pixel detector provides three high precision space points for each charged particle and thus gives the possibility to reconstruct the track.

After the pixels and on their way out of the tracker, particles pass through ten layers of silicon strip detectors, reaching out to a radius of 130 centimetres. This part of the tracker contains 15200 highly sensitive modules with a total of 10 million detector strips read by 80000 microelectronic chips. Each module consists of three elements: a set of sensors, its mechanical support structure and readout electronics.

The electromagnetic calorimeter (ECAL) The aim of the electromagnetic calorimeter is to measure the energy of photons and electrons. It is made of lead tungstate ($PbWO_4$) crystals whose usage guarantees high speed, fine granularity and radiation resistance, all important characteristics in the LHC environment.

These crystals have the important property to scintillate when electrons or photons pass through them, namely they produce light in proportion to the particles energy.

The ECAL has a cylindrical shape with two endcaps. The central part is called "the barrel". A total of about 61200 crystals are located in the barrel and 7324 in the endcaps. Avalanche photodiodes are used as photodetectors in the barrel and vacuum phototriodes in the endcaps. Each photodetector produces an electric signal whose intensity is proportional to the energy of the photons coming from the crystal. As a result it is possible to measure the energy of the particles (electrons or photons) produced during the collisions.

The hadron calorimeter (HCAL) The CMS detector is designed to study a wide range of high-energy processes involving different signatures of final states. For this reason the hadron calorimeter (HCAL) is particularly important for the measurement of hadron jets and neutrinos or exotic particles resulting in apparent missing transverse energy.

The hadron calorimeter is radially restricted between the outer extent of the electromagnetic calorimeter (R = 1.77 m) and the inner extent of the magnet coil (R = 2.95 m). As the electromagnetic calorimeter, the HCAL has a cylindrical shape composed of a barrel and two endcaps.

The HCAL is a sampling calorimeter meaning that it finds a particle position, energy and arrival time using alternating layers of “absorber” and fluorescent “scintillator” materials that produce a rapid light pulse when the particle passes through. Special optic fibres collect up this light and feed it into readout boxes where photodetectors amplify the signal. When the amount of light in a given region is summed up over many layers of tiles in depth, called a “tower”, this total amount of light is a measure of a particle’s energy.
Measuring hadrons is important as they can tell us if new particles such as the Higgs boson or supersymmetric particles have been formed.

The muon detectors As the name suggest, one of the CMS main aims is to measure muons, fundamental particles similar to the electron but with a mass about 200 times heavier. A precise measure of these particles is important because we expect them to be produced in the decay of a number of potential new particles; for instance, one of the clearest “signatures” of the Higgs Boson is its decay into four muons.
Unlike most of the particles produced in the LHC collisions, muons can penetrate several layers of matter without interacting. For this reason, while most of the other particles are stopped in the internal calorimeters (ECAL, HCAL), muons are revealed in special detectors located externally to them.
The track of a muon is measured by fitting a curve to hits among the four muon stations, which sit outside the magnet coil and are interleaved with iron plates.
By tracking their position through the multiple layers of each station, combined with tracker measurements, the detectors precisely trace the muons paths. Furthermore, thanks to the strong magnetic field, the muons momenta can be measured by observing the bending of their tracks.

Chapter 3

Data analysis

3.1 Samples

In this analysis both data and Monte Carlo (MC) simulatated samples have been used.

The data sample was collected during the 2016 LHC run from pp collisions at 13 TeV, corresponding to an integrated luminosity of 37.0 fb^{-1}.

As far as MC samples are concerned, different simulation programs have been used to describe different processes. For the signal $t\bar{t}$ sample, POWHEG [9, 10] was used as the generator, computing the QCD matrix element to the next-to-leading order (NLO), while PYTHIA 8 [11, 12] was used to simulate the parton shower development using the tune CUETP8M2T4 [13]. As a standard in MC samples, $t\bar{t}$ events were generated by assuming a top quark mass of 172.5 GeV.

The dominant background process is the QCD multijet production, which was simulated using MADGRAPH [14] for different intervals of H_T, thus obtaining six independent samples, H_T being the scalar sum of the transverse momenta of the partons produces in the hard scatter. For the subdominant backgrounds, single top quark production processes were simulated by using POWHEG, while W production and Drell-Yan events were simulated with MADGRAPH. In all the cases, the hard process was interfaced with the PYTHIA 8 simulation of the parton shower development.

The detector effects were simulated by using GEANT 4 [15].
Table 3.1 summarizes the MC samples used in this analysis, and for each of them shows the corresponding cross section and the total number of generated events.

3.2 Reconstruction and event selection

The raw data that are collected by the CMS detector are nothing but the output, event by event, of the response of each subdetector. Despite carrying the whole information, this format is not suitable for the analysis. The events are thus reconstructed from raw data by applying a series of algorithms which are able to identify each object (particles,

Sample	**Program**	$\boldsymbol{\sigma}$**(pb)**	$\mathbf{N_{gen}}$
$t\bar{t}$ pair production	POWHEG + PYTHIA 8	832	77081150
QCD (300 < H_T < 500 GeV)	MADGRAPH + PYTHIA 8	$3.67 \cdot 10^5$	54537900
QCD (500 < H_T < 700 GeV)	MADGRAPH + PYTHIA 8	$2.94 \cdot 10^4$	62271340
QCD (700 < H_T < 1000 GeV)	MADGRAPH + PYTHIA 8	$6.52 \cdot 10^3$	45412780
QCD (1000 < H_T < 1500 GeV)	MADGRAPH + PYTHIA 8	$1.06 \cdot 10^3$	15127290
QCD (1500 < H_T < 2000 GeV)	MADGRAPH + PYTHIA 8	121.5	11826700
QCD (H_T > 2000 GeV)	MADGRAPH + PYTHIA 8	25.4	6039005
Single t (tW)	POWHEG + PYTHIA 8	35.60	6952830
Single $\bar{t}$ ($\bar{t}W$)	POWHEG + PYTHIA 8	35.60	6933094
Single t (t-channel)	POWHEG + PYTHIA 8	136.02	67240810
Single $\bar{t}$ (t-channel)	POWHEG + PYTHIA 8	80.95	38811020
W + jets	MADGRAPH + PYTHIA 8	3539	22402470
Drell-Yan + jets	MADGRAPH + PYTHIA 8	1460	12055100

Table 3.1: List of all the MC samples used in this analysis along with the corresponding MC generator, cross section σ, and total number of generated events N_{gen}.

jets), compute their defining variables (momentum, energy) and reconstruct some of the properties of the whole event (interaction vertex).

3.2.1 Jet reconstruction

A jet can be defined as a collimated spray of stable particles arising from the fragmentation and hadronisation of a parton (quark or gluon) after a collision. Jet reconstruction algorithms are used to combine the calorimetry and tracking information to define jets. Some aspects of an algorithm that need to be considered are the jet size and whether the algorithm is infra-red and collinear safe. The jet size and area determine the susceptibility of a jet to soft radiation. A larger jet radius is important as it allows the jet to capture enough of the hadronised particles for the accurate calculation of the jet mass and energy. However, a smaller jet radius is useful in reducing the amount of the underlying event and pile-up captured by the jet, preventing the overestimation of the jets mass and energy. The splitting of a hard particle, while using a collinear unsafe algorithm, will result in the altering of the number and contents of the jets. A similar problem arises when a soft gluon is added to the system while an infra-red unsafe algorithm is in use.

Sequential clustering algorithms assume that particles within jets will have small differences in transverse momenta and thus group particles based on momentum space, resulting in jets that have fluctuating areas in the η-ϕ space. Sequential clustering algorithms are also infra-red and collinear safe. In such an algorithm, first the following quantities are defined

$$d_{ij} = min(p_{T,i}^{2a}, p_{T,j}^{2a})\frac{\Delta R_{ij}^2}{R^2}, \qquad \Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2, \tag{3.1a}$$

$$d_{iB} = p_{T,i}^{2a}, \tag{3.1b}$$

where d_{ij} is the "distance" between particle i and particle j, d_{iB} is the distance between particle i and the beam, R is the so-called distance parameter, η_i and ϕ_i are the pseudorapidity $\eta = -ln(\tan\theta/2)$ and the azimuthal angle of particle i and a is an integer which can take values -1, 0, 1. Then the procedure starts:

1. Evaluate all the distances d_{ij} and d_{iB} from the list of all the final state particles;
2. Find the minimum distance;
3. If it is a d_{ij}, recombine together particle i and j, then come back to step 1;
4. Otherwise declare particle i to be a jet, remove it from the final list and come back to step 1;
5. Algorithm stops when no particles remain.

In this analysis jet reconstruction is implemented using the *anti-k_T* algorithm [16],which corresponds to $a = -1$, with distance parameter $R = 0.8$ (AK8 algorithm). The particles which are used as inputs to the jet reconstruction are identified using the CMS *particle flow* (PF) algorithm [17],which aims at identifying and reconstructing all the particles from the collision by combining optimally the information of the different CMS subdetectors. It relies on a efficient and pure track reconstruction, on a clustering algorithm able to disentangle overlapping showers, and on an efficient link procedure to connect together the deposits of each particle in the sub-detector. Once all the deposits of a particle are associated, its nature can be assessed, and the information of the sub-detectors combined to determine optimally its four-momentum. In order to mitigate the effect of pileup, charged PF candidates that are unambiguously associated to pileup vertices are removed prior to the jet clustering.
The use of the PF algorithm allows to enhance the energy resolution of jets, if compared to methods which exploit information from calorimeters only; resolution of 15% at 10 GeV, 8% at 100 GeV and 4% at 1 TeV can be reached. Finally, using simulation, corrections to the jets energy is applied, defined as functions of variables such as η and p_T of the jets.

As previously stated, in this analysis boosted topology $t\bar{t}$ events are studied, namely events in which the top quarks have enough boost that their decay products are strongly collimated. For this reason, the anti-k_T algorithm reconstruct them as a single top quark jet with large distance parameter R (AK8 jets). However, a study of these wide jet substructures is useful to reject background events (QCD multijet events or subdominant background processes). An AK8 jet coming from a top quark decay should in fact contain three localized energy deposits, each one coming from the hadronization of a quark. These energy substructures are called *subjets* and, in order to identify them, the *n-subjettiness* [18] variable τ_i can be defined

$$\tau_i = \frac{1}{\sum_k p_{T,k} R} \sum_k p_{T,k} \min(\Delta R_{1k}, \Delta R_{2k}, ...\Delta R_{ik}), \tag{3.2}$$

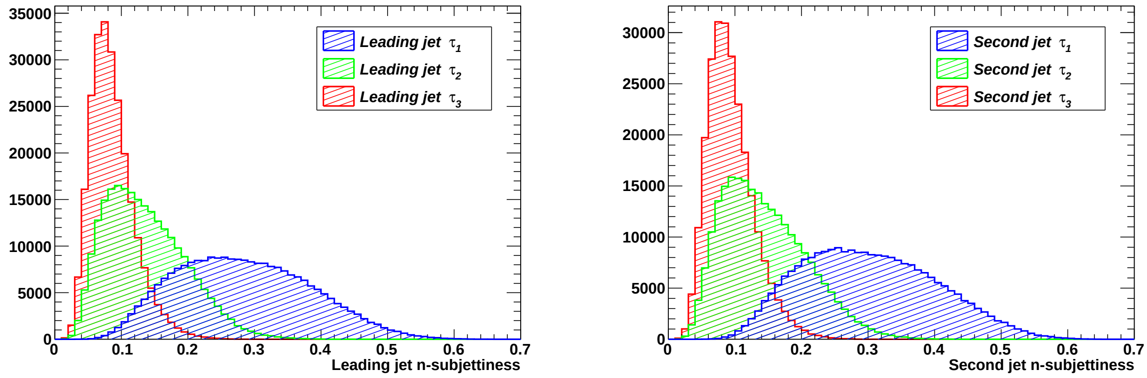
where k enumerates the constituent particles in a given jet, $p_{T,k}$ are their transverse momenta, and $\Delta R_{i,k} = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ is the distance in the $\eta - \phi$ plane between a candidate subjet i and a constituent particle k. The subjettiness variable τ_i thus measures the compatibility of a jet with the hypothesis that it is composed of i subjets. Jets with $\tau_i \approx 0$ have all their energy deposits aligned with the candidate subjet directions and therefore are composed of i (or fewer) subjets. On the other hand, jets with $\tau_i \gg 0$ have a large fraction of their energy distributed away from the candidate subjet directions and therefore have at least $i + 1$ subjets. Since for all-hadronic top quark decays a topology with three subjects is expected, the τ_1, τ_2 and τ_3 values can be effectively used to help discriminating between boosted top quark jets and QCD jets that tend to contain fewer subjets.

Fig. 3.1 shows the n-subjettiness distributions for the leading and second jets, obtained from the $t\bar{t}$ simulated sample. In order to fill the histograms a pre-selection was applied, asking for at least two AK8 jets in the event, each one with $p_T > 400$ GeV, $|\eta| < 2.4$ and jet mass > 50 GeV. As it is possible to notice, since in the simulation only $t\bar{t}$ events are present, jets are less likely to be constitued by one or two subjets, so that τ_1 and τ_2 are distributed farther from zero than τ_3.
In order to prove that the n-subjettiness can be used to help discriminating the signal from background events, in Fig. 3.2 the signal and background τ_3/τ_1 and τ_3/τ_2 normalized distributions are compared, both for the leading and second jet, as obtained from the $t\bar{t}$ and QCD simulated samples.

For the calculation of a jet invariant mass with suppressed radiation, underlying event, and pileup contamination, the SoftDrop method [19] is used, which removes from the jet clustering soft or collinear particles. In this method, the clustering of the jet j with distance parameter R is reverted step by step, breaking j into j_1 and j_2 at each iteration. Then, the SoftDrop condition is examined:

$$\frac{\min(p_{T,1}, p_{T,2})}{p_{T,1} + p_{T,2}} > z_{cut} \cdot \left(\frac{\Delta R_{12}}{R}\right)^{\beta}. \tag{3.3}$$

If the condition holds then j is considered the final jet and the procedure stops. Otherwise, the leading subjet is relabelled as j and the softer one is discarded. The two



(a) (b)

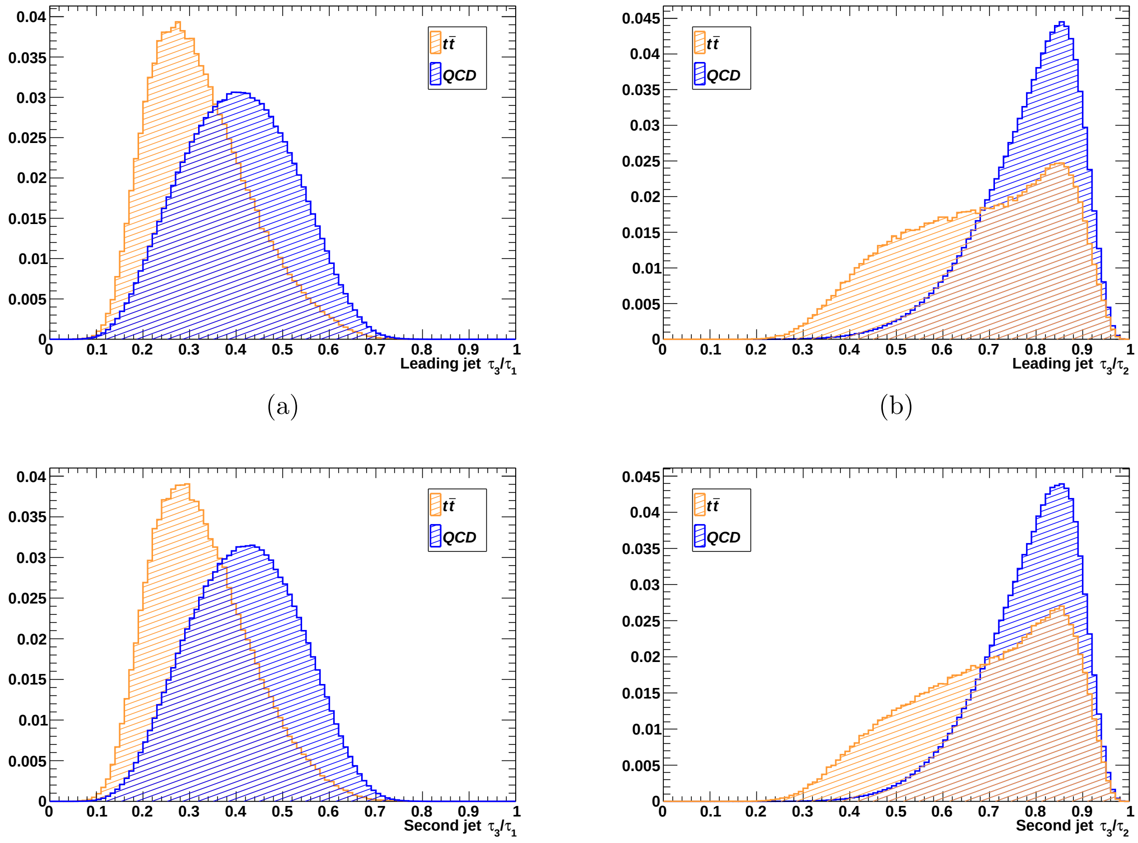
Figure 3.1: n-subjettiness distributions for the leading (a) and second (b) jets obtained from the $t\bar{t}$ simulated sample.

parameters of the algorithm, z_{cut} and β, control the strength of the fractional p_T selection and the suppression of the collinear radiation, respectively. For the present study the values $z_{cut} = 0.1$ and $\beta = 0$ are used, which have been found to be optimal for CMS analyses targeting boosted top quarks [20].

The identification of jets that likely originate from the hadronization of b quarks is done with the *Combined Secondary Vertex version 2* b-tagger [21], which combines in an optimal way the information from track impact parameters and identified secondary vertices within a given jet, and provides a continuous discriminator output. The *medium* working point, is here used, which corresponds, on average, to about 60% b jet efficiency and about 1% misidentification probability for non-b jets. In this analysis the b-tagging algorithm is applied to AK8 jets to identify those containing subjets coming from b quark hadronization.

3.2.2 Lepton reconstruction

Lepton (electrons and muons) candidates are reconstructed using the PF algorithm, and are required to have $p_T > 10$ GeV and $|\eta| < 2.4$. In addition, an isolation criterion $I_{rel} < 0.15$ (0.07) is applied for muon (electron) candidates, where I_{rel} is defined as the sum of the p_T of all reconstructed particle candidates inside a cone around the lepton in η-ϕ space of radius ΔR, excluding the lepton itself, divided by the p_T of the lepton. For muons (electrons) the radius of the isolation cone is 0.4 (0.3). Finally, a correction is applied to the isolation definition in order to take into account the dependence on pileup. In this study, lepton reconstruction is important to select only events with no isolated leptons, since the all-hadronic $t\bar{t}$ decay is studied.



(a)

(b)

(c)

(d)

Figure 3.2: τ_3/τ_1 (left) and τ_3/τ_2 (right) normalized distributions for the leading (upper) and second (lower) jets, obtained from the $t\bar{t}$ and QCD simulated samples.

3.2.3 Trigger selection

The collision rate at the LHC is heavily dominated by QCD processes with large cross section, which are not interesting for the physics program of the CMS experiment. Since it is not possible to register all the events, it becomes mandatory to use a trigger system in order to select events according to physics-driven choices. The CMS experiment features a two-level trigger architecture. The first level (L1), hardware, operates a first selection of the events to be kept, using subdetector informations. The second level, called High Level Trigger (HLT), is implemented in software and aims to further reduce the event rate to about 800 Hz on average. The HLT contains many trigger paths, each corresponding to a dedicated trigger.

There are two main types of trigger paths. The *signal trigger paths* aim to collect interesting signal events and they are always unprescaled. This means that every event passing the trigger is recorded and for this reason these trigger paths consume most of the available bandwidth. Besides the signal triggers, there are *control trigger paths*, which are mainly used for the study of the background. In order not to consume too much bandwidth in collecting background events, a prescale is usually applied, meaning that only a fraction of the events that pass the trigger is collected.

In this analysis the *HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV* signal trigger path has been used. Starting from a L1 trigger which requires a single jet with $p_T > 180$ GeV, the HLT requires at least two AK8 jets with $p_T > 280$ GeV and 200 GeV, where at least one is b-tagged (online b-tagging).

The good quality of a trigger path is evincible from an efficiency study. To compute the efficiency a *reference trigger* is needed, namely another trigger path which applies looser cuts to select the events. In this analysis the reference trigger path *HLT_IsoMu27* has been used, which requires an isolated muon with $p_T > 27$ GeV. Then, the efficiency can be computed as the ratio between the number of events which pass both trigger paths and the number of events only passing the reference trigger path. Some offline cuts are also applied:

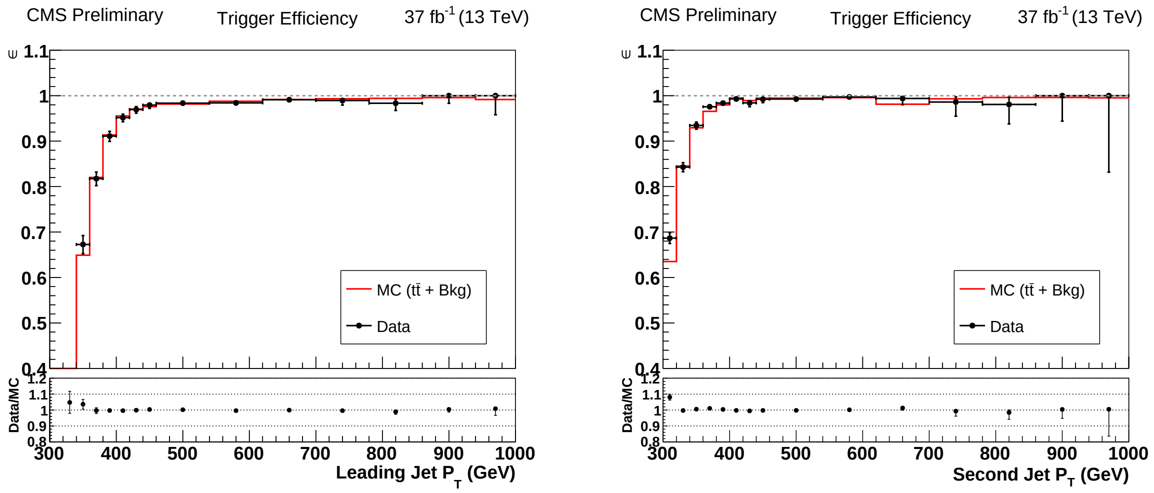
- at least two jets per event, since two wide jets coming from the $t\bar{t}$ decay should be present in signal events;
- second jet $p_T > 300$ GeV. Since in this analysis only boosted top quarks are considered, the trigger efficiency must be studied in the high-p_T region;
- at least one of the two leading jets should contain at least one b-tagged subjet.

This offline requirement is introduced in order to study the trigger efficiency in a phase space close to the signal phase space, which will be defined in the next section. By including the offline selection, the trigger efficiency is then given by

$$\epsilon = \frac{\text{events passing signal trigger \& offline cuts \& reference trigger}}{\text{events passing offline cuts \& reference trigger}}. \tag{3.4}$$

Because the set associated to the numerator of eq. 3.4 is a sub-set of the one associated to its denominator, and they are thus correlated, the computation of the efficiency error bars is not trivial. For this reason the ROOT *TEfficiency* class [22] has been used, which allows to compute the error bars with an exact coverage of 68.3%, using the frequentist Clopper-Pearson method.

In Fig. 3.3 the trigger efficiency is shown, as measured in data, as a function of the leading and second jet p_T, along with the corresponding efficiency in MC. Since the ratio between data and MC, shown in the bottom panels of Fig. 3.3a, 3.3b, is essentially equal to 1 for $p_T > 400$ GeV, it is possible to state that no scale factor is needed in this analysis, to make the MC efficiency agree with the value expected on data.



(a)

(b)

Figure 3.3: Trigger efficiency, measured in data, as a function of the first (a) and second (b) jet p_T, along with the corresponding efficiency in MC (upper panel); data/MC scale factor (bottom panel).

3.2.4 Event selection

In order to extract the $t\bar{t}$ events candidates amongst the wide variety of processes with boosted topology that can occurr at the LHC in pp collisions, a selection needs to be applied. The selected phase space needs to be defined according to a balance between efficiency and purity. The higher the efficiency of the selection, the higher the number of the collected events, but the lower the purity, since a larger number of background events will be selected.

In this analysis, two different selections are defined: an *extended*, much looser, selection, used for the inclusive $t\bar{t}$ cross section measurent, and a *reduced*, tighter, selection, used for the differential measurement.

The extended selection starts by requiring the trigger path introduced in the previous section: HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV. Then events are selected with at least two AK8 jets, each one with $p_T > 400$ GeV, $|\eta| < 2.4$, SoftDrop mass $m_{SD} > 50$ GeV. Because in this analysis only hadronic $t\bar{t}$ decays are studied, no isolated lepton has to present in the selected events (lepton veto). Furthermore, as previously stated, jets coming from $t\bar{t}$ decays should contain one subjet originated from the hadronization of a b quark. Events are thus divided into three different categories: 2-btag category events in which both jets contain at least one b-tagged subjet; 1-btag category events in which only one of the two jets contains at least one b-tagged subjet; events containing jets with no b-tagged subjets (0-btag category). For this reason, in order to improve the signal selection, the condition category = 2-btag is also asked. Finally, a neural network is constructed from the n-subjettiness τ_1, τ_2 and τ_3 of the two leading jets. The samples used for the training are a simulated $t\bar{t}$ sample for the signal and a simulated QCD multijet sample for the background. Fig. 3.4 shows the neural network architecture and peformance.

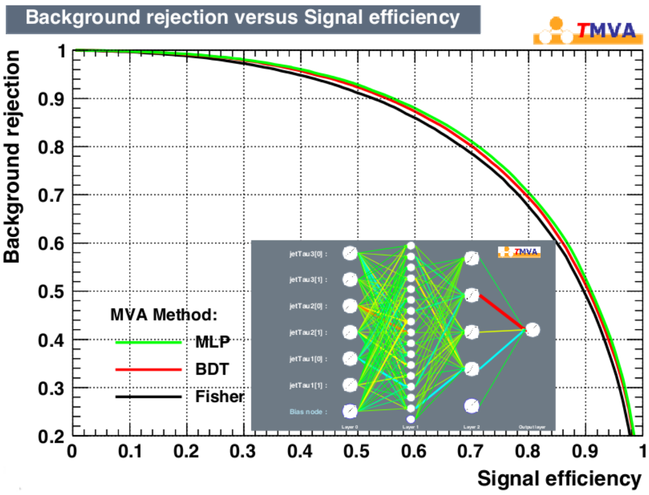


Figure 3.4: Representation of the neural network architecture together with a comparison of the ROC curves of different multivariate methods: neural network (green), boosted decision tree (red), Fisher discriminant (black).

After the training, it turns out that an appropriate selection requires the condition $mva > 0.9$ on the neural network output. Table 3.2 summarizes the conditions applied in the extended selection.

Extended Selection
Signal Trigger
$N_{\text{AK8 jets}} \geq 2$ with $p_T > 400$ GeV, $\lvert\eta\rvert < 2.4$, $m_{SD} > 50$ GeV
50 GeV $< m_{SD}^{(1)} <$ 300 GeV
lepton veto
2-btag category
$mva > 0.9$

Table 3.2: Definition of the extended selection.

The reduced selection used for the differential measurement, is simply obtained from the extended one by applying an additional cut on the SoftDrop mass of both jets:

$$140 \text{ GeV} < m_{SD}^{(1)} < 200 \text{ GeV}, \qquad 140 \text{ GeV} < m_{SD}^{(2)} < 200 \text{ GeV}.$$

An important parameter, which is necessary for computing the inclusive cross section, is the selection efficiency ϵ, defined as

$$\epsilon = \left[\frac{N_{sel}}{N_{gen}}\right]_{MC}, \tag{3.5}$$

where N_{sel} is the number of selected events in the MC sample and N_{gen} is the total number of generated events in the sample. For the extended and reduced selections, the efficiencies computed on the $t\bar{t}$ MC sample are

$$\epsilon_{ext} = 2.07 \times 10^{-4}, \tag{3.6a}$$
$$\epsilon_{red} = 1.11 \times 10^{-4}. \tag{3.6b}$$

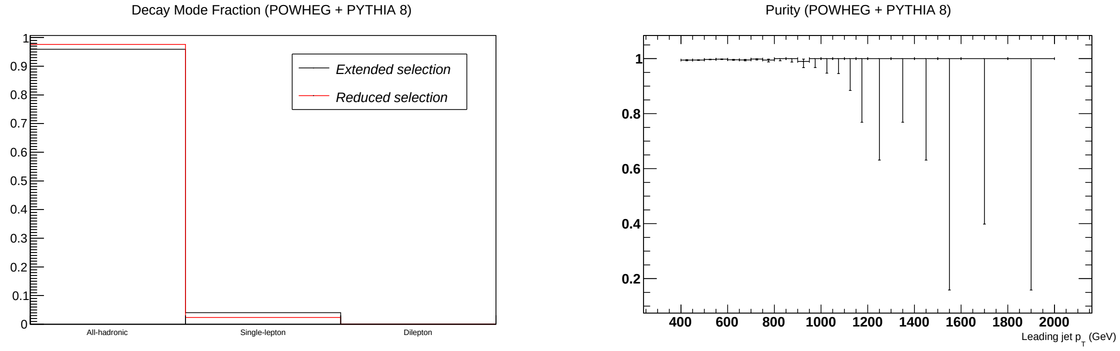
In order to evaluate the quality of the selections, some quantities have been studied, using the $t\bar{t}$ simulated sample. First of all, the fraction of events as a function of lepton number has been studied. As expected, in Fig.3.5a it is shown that 96% of the events retained by the extended selection are all-hadronic. Furthermore, if the reduced selection is asked, a higher fraction of leptonic and semileptonic events is discarded, resulting in a more pure hadronic sample (98%).

Then, the probability is studied that a selected jet really comes from a generated top quark, since there could be sources of jets other than top quarks (e.g. underlying event, pileup) that may enter the signal phase space. For each event, the distance ΔR in the $\eta - \phi$ plane between jets and generated top quarks is computed. If this distance is less than 0.2, the jet and its corresponding top quark are said to be matching. In Fig. 3.5b the sample purity is shown, namely the probability, as a function of the leading jet p_T, that a selected jet matches the corresponding generated top quark. It is possible to notice that, by using the extended selection, a purity of about 100% is achieved along the whole spectrum.

Finally, the b-tagging selection is studied. Fig. 3.6 compares the leading jet SoftDrop mass distributions obtained before and after the b-tagging requirement. It is possible to notice that the b-tagging condition removes the W boson mass peak from the spectrum, thus helping to select a more pure $t\bar{t}$ sample.

3.3 Data and simulation comparison

The aim of this section is to prove that the simulated samples used in this analysis correctly reproduce the distributions observed in data. As a matter of fact, the input parameters of the MC programs have uncertainties and even the choice of the tuning may lead to a mismodeling of the distributions.



(a) Decay mode fraction for simulated $t\bar{t}$ events, selected by requiring the extended selection (with no lepton veto).

(b) Purity distribution as a function of the leading jet p_T obtained from the $t\bar{t}$ simulated sample by requiring the extended selection.

Figure 3.5

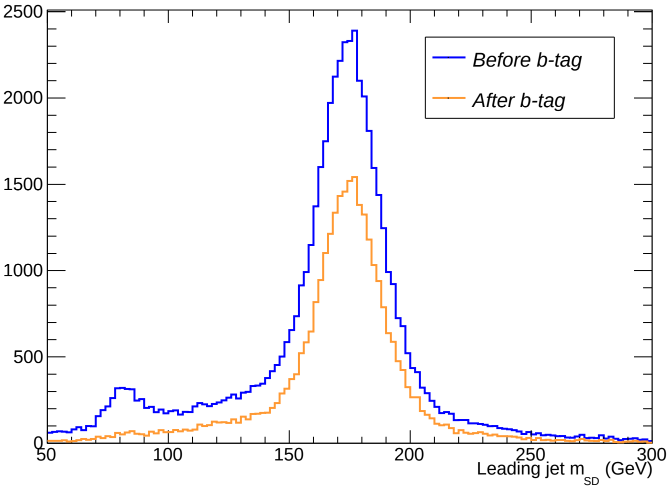


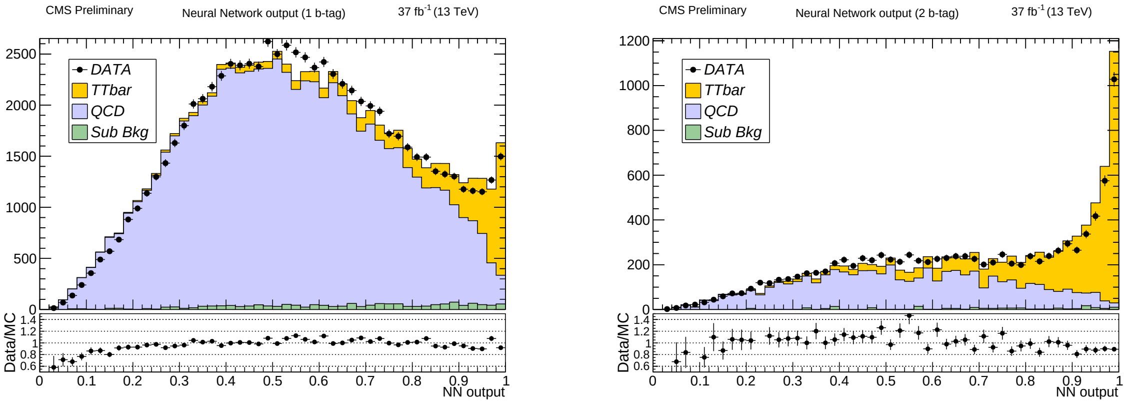
Figure 3.6: Leading jet SoftDrop mass distribution before and after the b-tagging requirement, obtained from the $t\bar{t}$ simulated sample.

Fig. 3.7, 3.8, 3.9 show some plots in which the distributions observed in the data are compared to the simulated yields coming from different samples: $t\bar{t}$, QCD and subdominant backgrounds. In order to generate these plots, first each contribution coming from the different MC samples has been normalized to the integrated luminosity, using the theoretical cross sections reported in Table 3.1. Then, a global normalization factor has been applied to all the contributions so that the total simulated yield is equal to the number of events in data. The error bars in the ratio plots shown in the bottom panel of each figure, are computed as the sum in quadrature of the data and MC statistical uncertainties.

Fig. 3.7 shows the multivariant discriminant output distributions for both the 1-btag and 2-btag categories after the extended selection. These plots, other than showing a good agreement between data and simulations, prove that the cut $mva > 0.9$ on the multivariate discriminant selects a sample highly dominated by signal.

Fig. 3.8 shows the leading jet SoftDrop mass distributions for both the 1-btag and 2-btag categories, obtained after asking the extended selection. These datasets are important in this analysis as they are used to perform the inclusive cross section measurement after fitting the distribution with appropriate signal and background templates, as it will be explained later.

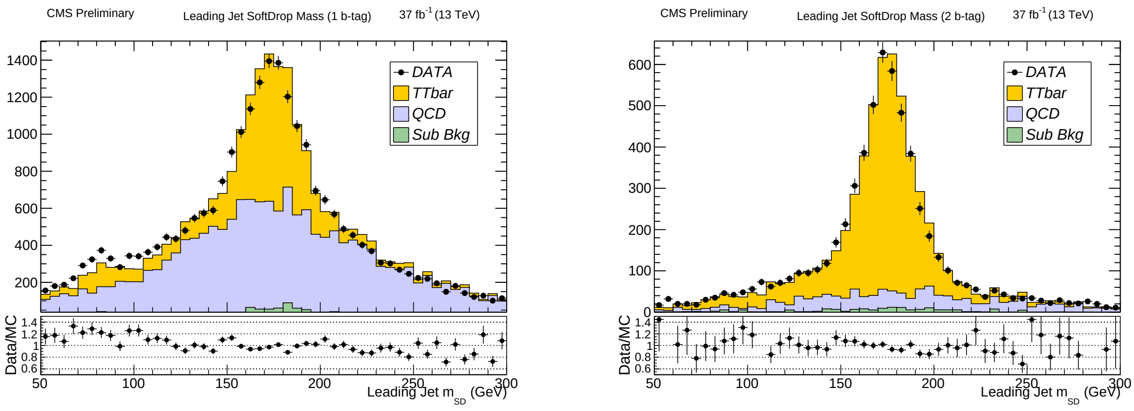
Finally, Fig. 3.9 shows the data-simulation comparisons for other distributions obtained, for both categories, by asking the reduced selection. A general good agreement is observed for each plot. Causes of the small discrepancies could be a mismodeling of the boosted top quark jets in the $t\bar{t}$ sample, or more likely a mismodeling of the QCD jets that mimic the properties of top quark jets. It must be pointed out however, that the QCD simulation is not directly used in this measurement, since the background prediction will be made starting from a control sample in data, as it will be clarified later.



(a)

(b)

Figure 3.7: Comparison between the neural network output distributions obtained from data and simulations for the 1-btag (a) and 2-btag (b) categories by asking the extended selection.



(a) (b)

Figure 3.8: Comparison between the leading jet SoftDrop mass distributions obtained from data and simulations for the 1-btag (a) and 2-btag (b) categories by asking the extended selection.

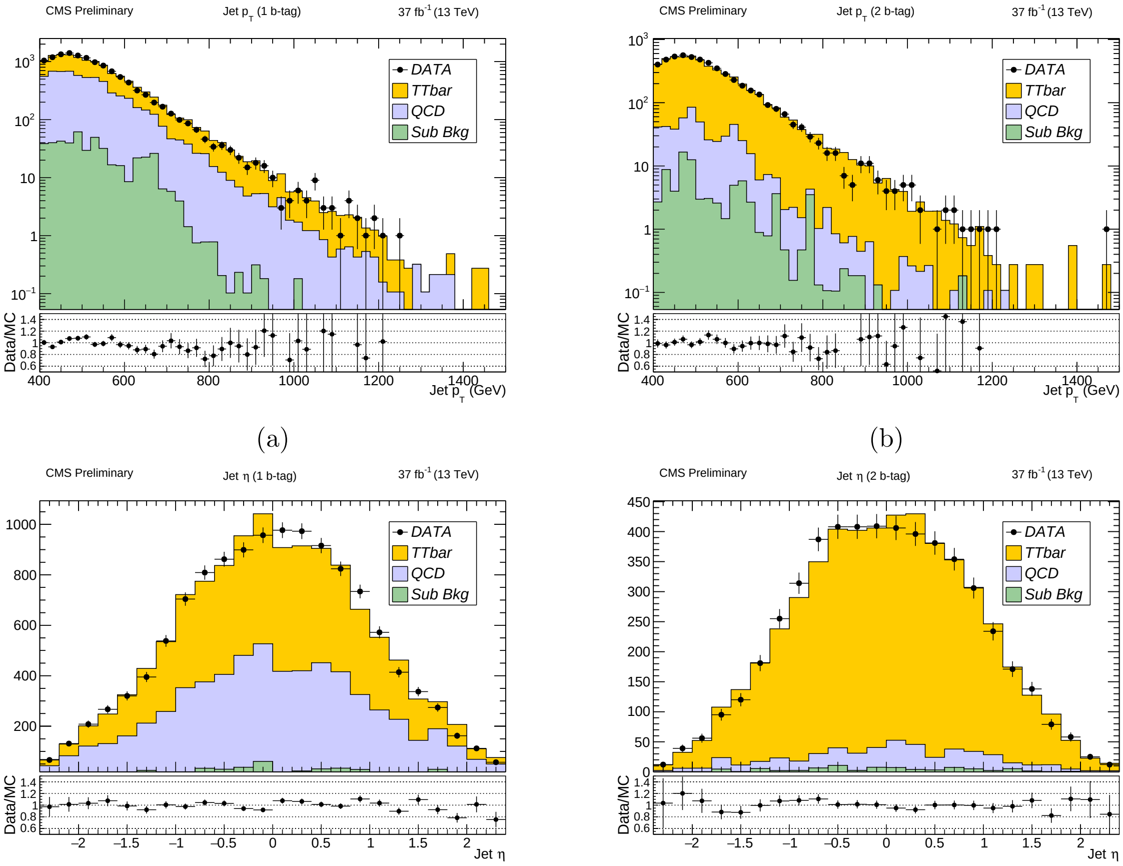
3.4 Inclusive cross section

3.4.1 Background prediction

Due to their large cross section in pp collisions, QCD multijet events are the dominant background for the $t\bar{t}$ events in the all-hadronic final state. Unfortunately MC simulations of the QCD background do not describe the data accurately enough, mainly for what concerns the yield. Furthermore, despite the large number of simulated events, very few pass the selection, resulting in a very small sample of available events. For these reasons it is advisable to define a *control sample*, which can be used for a data-based estimate of the QCD distributions of any variable of interest.

The control sample can be extracted from data by asking the same cuts used to select signal phase space, but reverting the conditions that involve b-tagging. A *control trigger path* is chosen, which applies the same kinematic cuts as the signal trigger, but it does not feature the HLT online b-tagging. Furthermore, only the events with exactly no b-tagged jets can enter the control sample. In this way it is possible to obtain an almost pure QCD sample showing a behaviour similar to that of QCD multijet events in the signal region. A summary of the selection for the control sample is shown in Table 3.3.

As previously stated, besides QCD multijet events, many other sub-dominant processes contribute to the background in the signal region. Unlike QCD, these processes are safely described by MC methods, so there is no need to define other control samples. The same cuts used to define the signal region can be applied to these MC samples to estimate the sub-dominant background shape.



(a)

(b)

(c)

(d)

Figure 3.9: Comparison between the two leading jets p_T (upper) and η (lower) distributions obtained from data and simulations for the 1-btag (left) and 2-btag (right) categories by asking the reduced selection.

3.4.2 Signal extraction

In order to perform both inclusive and differential $t\bar{t}$ cross section measurements, it is necessary to extract the signal yield from the events selected from data. This is done by fitting the data with appropriate parametrized functions (templates). The variable of interest used to extract the signal yield is the SoftDrop mass m_{SD} of the leading jet. It is possible to extract from data two m_{SD} distributions: the first asking for the event selection cuts (2-btag category), the second by requiring the same cuts as the first, but asking for just one b-tagged subjet per event (1-btag category). In this analysis the information coming from both distributions has been used to extract the signal yield. The signal and background templates have been fitted on both distributions at the same time, a procedure that is called *simultaneous fitting*. The fit procedure has been carried

Control Sample Selection
Control Trigger
$N_{\text{AK8 jets}} \geq 2$ with $p_T > 400$ GeV, $\lvert\eta\rvert < 2.4$, $m_{SD} > 50$ GeV
50 GeV $< m_{SD}^{(1)} <$ 300 GeV
lepton veto
b-tag veto
$mva > 0.9$

Table 3.3: Selection applied to data in order to define the QCD-enriched control sample.

on by using the ROOT package RooFit [23].

Signal templates As in the data, two distributions have been obtained from the $t\bar{t}$ MC for the 1-btag and 2-btag categories. These distributions are independently fitted with the model $S(m_{SD})$, which features the following components:

- a Gaussian $G_t(m_{SD}, k_{scale}\mu_t, k_{res}\sigma_t)$ describing the top quark mass peak. The quantities k_{scale} and k_{res} are parameters which are left free in this fit and they are determined only at the end of the simultaneous fit procedure.
- a Gaussian $G_W(m_{SD}, k_{scale}\mu_W, k_{res}\sigma_W)$ describing the W mass peak. The quantities k_{scale} and k_{res} are the same parameters left free in the previous function.
- a Bernstein polinomial with 8 parameters $P_8(m_{SD})$, describing the combinatorial background.

Figs. 3.10a and 3.10b show the two signal templates $S_{1b}(m_{SD})$ and $S_{2b}(m_{SD})$, containing two free parameters: k_{scale} and k_{res}.

Sub-dominant background The two distributions obtained from the sub-dominant background MC samples (1 b-tag, 2 b-tag categories), are independently fitted with the model $B^{sub}(m_{SD})$, which features the following components:

- a Gaussian $G_t(m_{SD}, k_{scale}\mu_t, k_{res}\sigma_t)$ describing the background top quark mass peak (e.g. single top quark processes). The parameters k_{scale} and k_{res} are left free in this fit and they are determined only at the end of the simultaneous fit procedure.
- a Gaussian $G_W(m_{SD}, k_{scale}\mu_W, k_{res}\sigma_W)$ describing the background W mass peak. The quantities k_{scale} and k_{res} are the same parameters left free in the previous function.

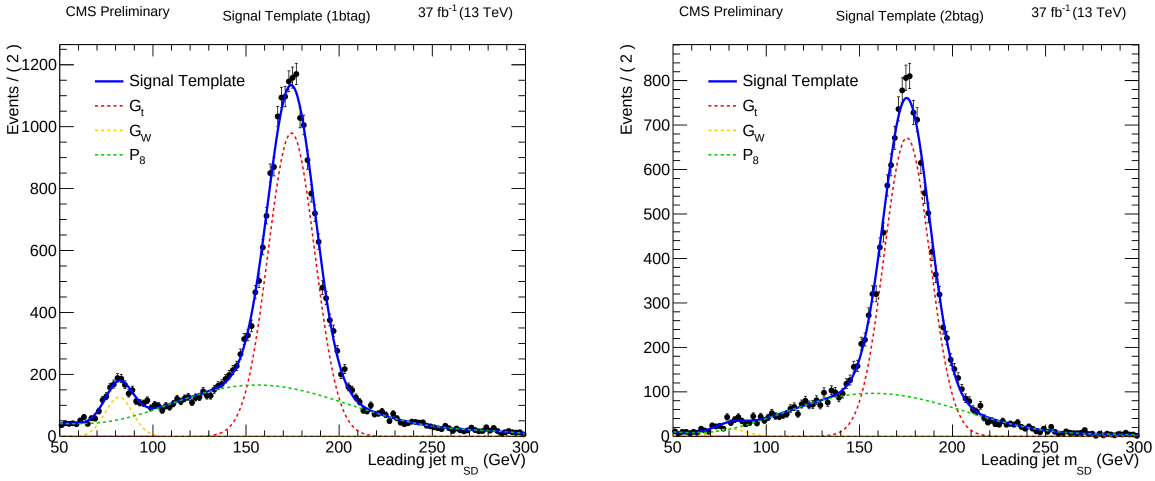
- a Bernstein polinomial with 3 parameters $P_3(m_{SD})$, describing the combinatorial background.

Figs. 3.11a and 3.11b show the two sub-dominant background templates $B_{1b}^{sub}(m_{SD})$ and $B_{2b}^{sub}(m_{SD})$, containing the two free parameters k_{scale} and k_{res}.

QCD template The m_{SD} distribution obtained from the QCD-enriched control sample is fitted with the model $B^{QCD}(m_{SD})$, which features the following components:

- a Gaussian $G_{QCD}(m_{SD}, \mu_{QCD}, \sigma_{QCD})$.
- a Bernstein polinomial with 4 parameters $P_4(m_{SD})$.

Fig. 3.12 shows the QCD background template $B^{QCD}(m_{SD})$, containing no free parameters.

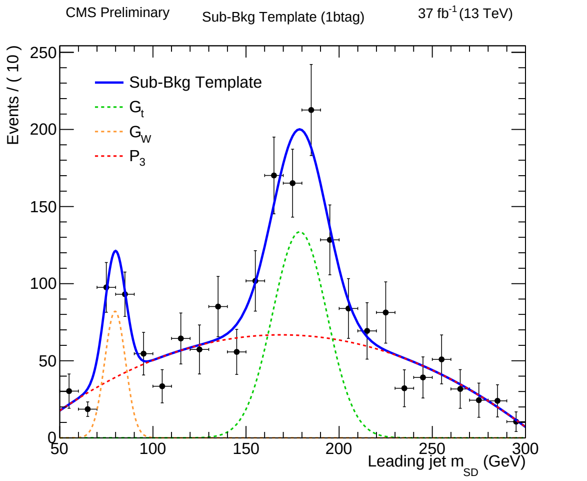


(a) Signal template (blue) fitted on the $t\bar{t}$ MC sample distribution (1 b-tag category). The dashed lines represent the components of the complete model.

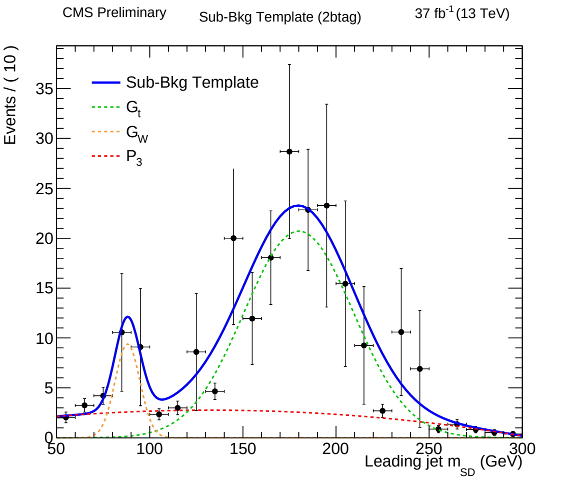
(b) Signal template (blue) fitted on the $t\bar{t}$ MC sample distribution (2 b-tag category). The dashed lines represent the components of the complete model.

Figure 3.10

The QCD template is extracted from a control sample which is outside the signal phase space. Thus, in order to safely use this template to fit the data, a correction should be applied. The corrected QCD templates $B_{1b}^{QCDcor}(m_{SD})$ and $B_{2b}^{QCDcor}(m_{SD})$ are obtained by multiplying the variable m_{SD} of the original QCD template $B^{QCD}(m_{SD})$, by the correction factors $(1 + k_{1b}^{QCD} m_{SD})$ and $(1 + k_{2b}^{QCD} m_{SD})$ respectively. The quantities k_{1b}^{QCD} and k_{2b}^{QCD} are free parameters whose value will be determined at the end of



(a) Sub-dominant background template (blue) fitted on the sub-dominant background MC samples distribution (1 b-tag category). The dashed lines represent the components of the complete model.



(b) Sub-dominant background template (blue) fitted on the sub-dominant background MC samples distribution (2 b-tag category). The dashed lines represent the components of the complete model.

Figure 3.11

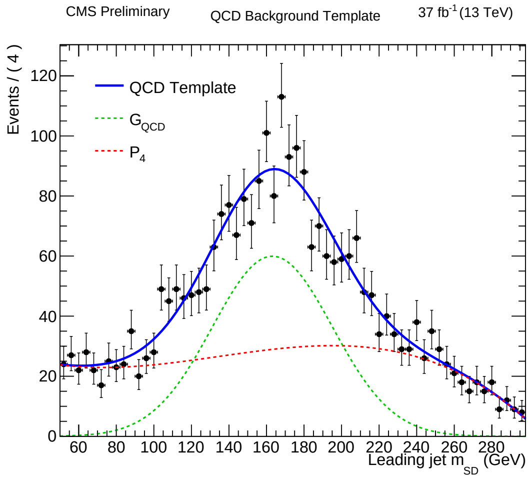


Figure 3.12: QCD background template (blue) fitted on the control sample distribution. The dashed lines represent the components of the complete model.

the simultaneous fit procedure. In the end, the corrected QCD background templates B_{1b}^{QCDcor} and B_{2b}^{QCDcor} depend on the free parameters k_{1b}^{QCD} and k_{2b}^{QCD}.

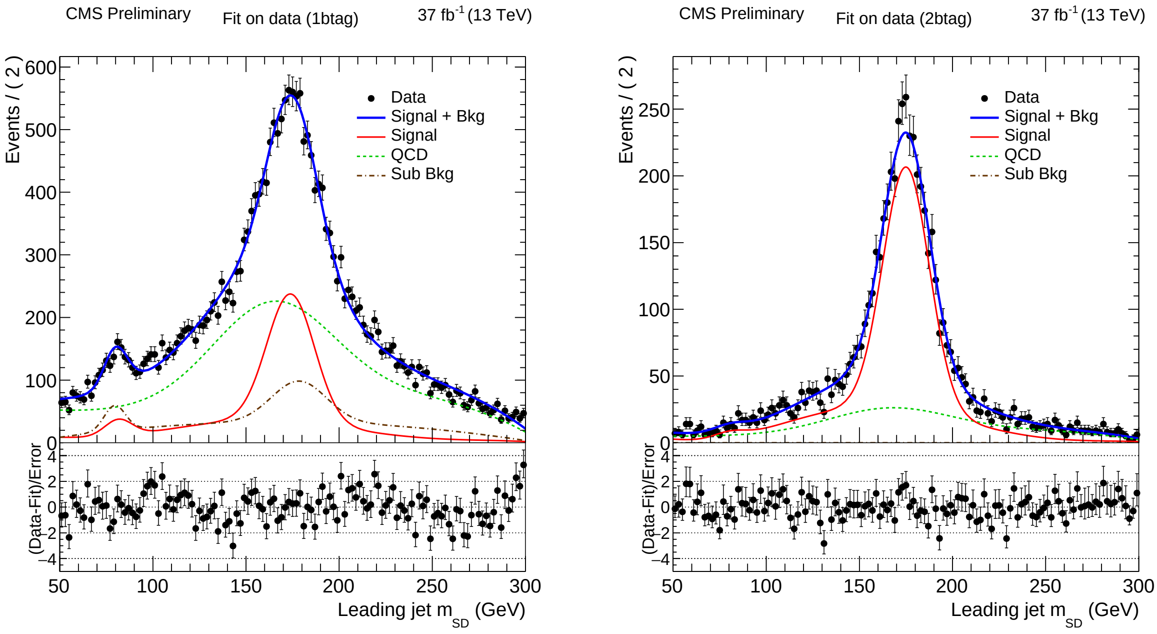
Finally, for each data sample (1 b-tag, 2 b-tag categories), a complete model is defined by adding the respective components: the signal template, the corrected QCD template and the sub-dominant background template. The final models $D_{1b}(m_{SD})$ and $D_{2b}(m_{SD})$

are simultaneously fitted on the two data samples and the values of the 10 free parameters are determined.

The signal yields N_{1b} and N_{2b}, respectively extracted from the 1 b-tag and from the 2 b-tag distributions, can be defined in terms of the *b-tagging efficiency* ϵ_b and the total number of selected $t\bar{t}$ events $N_{t\bar{t}}$ as

$$N_{1b} = 2\epsilon_b(1-\epsilon_b)N_{t\bar{t}} = 5742 \pm 283, \tag{3.7a}$$
$$N_{2b} = \epsilon_b^2 N_{t\bar{t}} = 4382 \pm 107. \tag{3.7b}$$

Figs. 3.13a and 3.13b show the two m_{SD} data distributions (1 b-tag, 2 b-tag) simultaneously fitted with the complete models.



(a) 1 b-tag data sample fitted with the complete model (blue). The signal, QCD and sub-dominant background templates are also shown.

(b) 2 b-tag data sample fitted with the complete model (blue). The signal, QCD and sub-dominant background templates are also shown.

Figure 3.13

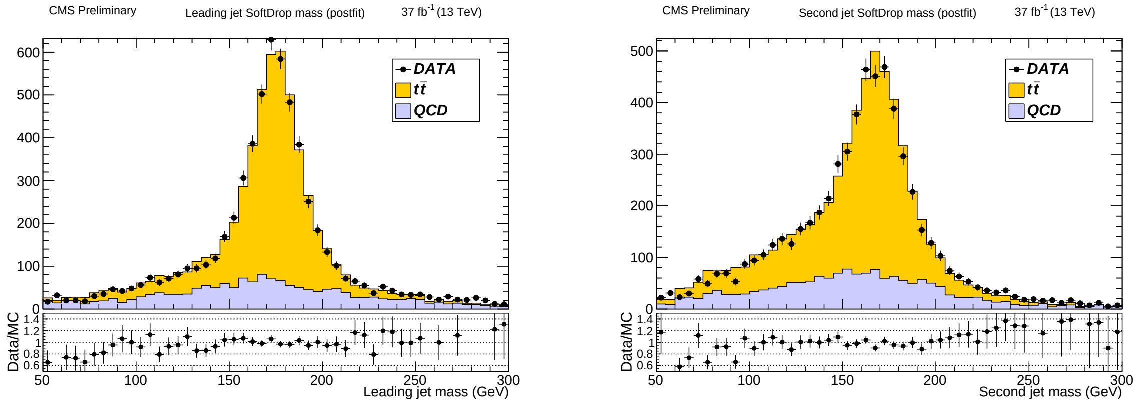
Since the total number of events extracted from data by asking the extended selection (2-btag category) is 6002, the signal and background yields are then given by

$$N_{sig} = 4382 \pm 107, \qquad N_{bkg} = 1620 \pm 93. \tag{3.8}$$

After the fit, the contribution to N_{bkg} of the subdominant processes is found to be completely negligible and thus the background in the signal region is only determined by QCD events as inferred from the control sample.

3.4.3 Postfit comparisons

In order to support the validity of the fit procedure described in the previous section, some postfit comparisons are here presented. The signal distributions are taken from the $t\bar{t}$ simulated sample, while the QCD background distributions are taken from the control sample defined on data (the sub-dominant background contribution is here neglected). In Fig. 3.14 the leading and second jet mass distributions are shown. The $t\bar{t}$ and QCD contributions are normalized respectively to the signal and background fitted yields (eq. 3.8). The plots shown in Fig. 3.15 are instead obtained by requiring that the mass of both the leading and second jet belong to the interval 140 - 200 GeV (reduced selection), and the distributions are normalized to the signal and background yields in this window. The error bars in the ratio plots shown in the bottom panel of each figure, are computed as the sum in quadrature of the data and MC statistical uncertainties.
Each plot show a good agreement between data and the normalized templates, thus increasing the confidence on the validity of the fit procedure.



(a)

(b)

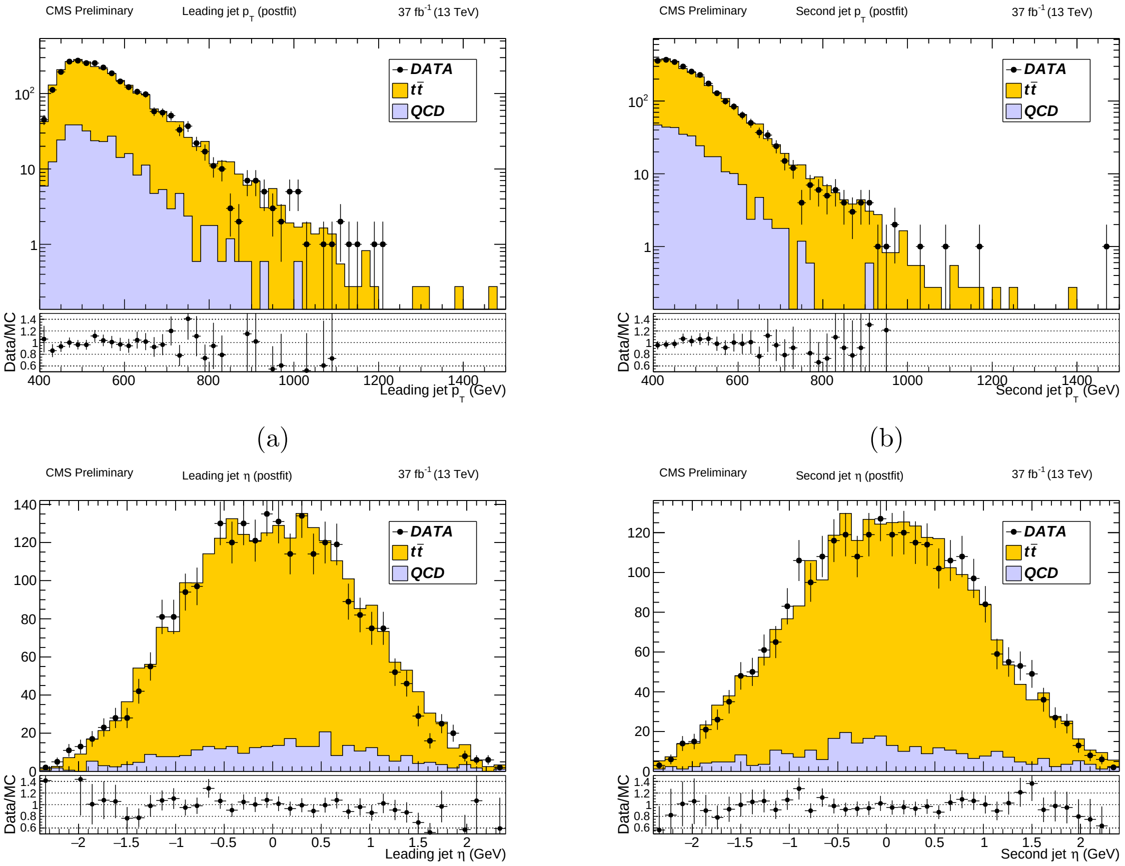
Figure 3.14: Postfit distributions for the SoftDrop mass of the leading (a) and second (b) jet. Signal and background distributions are normalized to the fitted yields.

3.4.4 Inclusive measurement

The inclusive $t\bar{t}$ cross section can be computed starting from the signal yield extracted from data in eq. 3.8 as

$$\sigma_{t\bar{t}} = \frac{N_{sig}}{\epsilon_{ext} \cdot L}, \tag{3.9}$$

where ϵ_{ext} is the MC efficiency for the extended selection (eq. 3.6a) and $L = 37.0$ fb^{-1} is the integrated luminosity of the data sample.



(a)

(b)

(c)

(d)

Figure 3.15: Postfit distributions for p_T (upper) and η (lower) of the leading (left) and second (right) jet. Signal and background distributions are normalized to the fitted yields in the reduced selection phase space.

By using the values reported before, the measured inclusive cross section is

$$\sigma_{t\bar{t}} = 572 \pm 14(\text{stat})\ \text{pb}. \tag{3.10}$$

This value has to be compared to the theoretical value of

$$\sigma_{t\bar{t}}^{theor} = 832^{+20}_{-29}(\text{scale}) \pm 35(\text{PDF} + \alpha_s)\ \text{pb}, \tag{3.11}$$

as computed with the Top++ program [24] at next-to-next-to-leading-order (NNLO) in perturbative QCD, assuming $m_t = 172.5$ GeV. As it can be noticed, the measured inclusive cross section is quite off the theoretical estimate (by ≈ 30%). As it will be explained in the next section though, the inclusive cross section measurement is affected by many sources of systematic uncertainties which need to be taken into account when comparing the result to the theoretical estimate.

3.4.5 Systematic uncertainties

The cross section measurement is affected by many sources of systematic uncertainties. In this analysis the most important ones have been considered:

- *Jet Energy Scale* (JES) and *Jet Energy Resolution* (JER): in the MC signal sample the energy scale and resolution of jets have been corrected in order to better describe the data. However, the uncertainties on these corrections have impact on the cross section measurement. To evaluate these sources of uncertainty the jet energy and resolution have been shifted and smeared up and down by one JES and JER standard deviation respectively. These corrections impact both on the shape of the signal templates and on the efficiency of the selection. The inclusive cross section is thus computed again obtaining

$$\begin{aligned}\sigma_{t\bar{t}}^{\text{JES-up}} &= 562 \pm 14 \text{ pb},\\ \sigma_{t\bar{t}}^{\text{JES-down}} &= 581 \pm 14 \text{ pb},\\ \sigma_{t\bar{t}}^{\text{JER-up}} &= 577 \pm 14 \text{ pb},\\ \sigma_{t\bar{t}}^{\text{JER-down}} &= 575 \pm 14 \text{ pb}.\end{aligned}$$

  The systematic uncertainty is determined by taking the semi-difference between the extreme values with respect to the average, as in

$$\frac{|\sigma_{t\bar{t}}^{\text{JES-up}} - \sigma_{t\bar{t}}^{\text{JES-down}}|}{\sigma_{t\bar{t}}^{\text{JES-up}} + \sigma_{t\bar{t}}^{\text{JES-down}}} = 0.017, \tag{3.12a}$$

$$\frac{|\sigma_{t\bar{t}}^{\text{JER-up}} - \sigma_{t\bar{t}}^{\text{JER-down}}|}{\sigma_{t\bar{t}}^{\text{JER-up}} + \sigma_{t\bar{t}}^{\text{JER-down}}} = 0.002. \tag{3.12b}$$

  Thus the JES uncertainty amounts to about 1.7%, while the JER uncertainty is estimated to be 0.2%.

- *MC tuning*: in the nominal MC sample (POWHEG + PYTHIA 8), the systematic uncertainty in matching the matrix-element to the parton shower is determined by varying the parameter h_{damp}, which regulates the high-p_T radiation by damping real emission generated in POWHEG, within its uncertainties. The parameter is set to $h_{damp} = 1.58^{+0.66}_{-0.59}$ multiplied by the mass of the top quark in the CUETP8M2T4 tune. The parameters controlling the underlying event in the CUETP8M2T4 tune are also varied to estimate the uncertainty in this source. The inclusive cross section is computed again for each variation (Tune-up, Tune-down) and the following values are obtained

$$\sigma_{t\bar{t}}^{\text{Tune-up}} = 596 \pm 16 \text{ pb},$$
$$\sigma_{t\bar{t}}^{\text{Tune-down}} = 574 \pm 15 \text{ pb},$$

The systematic uncertainty is determined by taking the semi-difference between the extreme values with respect to the average, as in

$$\frac{|\sigma_{t\bar{t}}^{\text{Tune-up}} - \sigma_{t\bar{t}}^{\text{Tune-down}}|}{\sigma_{t\bar{t}}^{\text{Tune-up}} + \sigma_{t\bar{t}}^{\text{Tune-down}}} = 0.019, \tag{3.13a}$$

$$\tag{3.13b}$$

and thus amounts to 1.9%.

- *Initial State Radiation* (ISR) and *Final State Radiation* (FSR): during a collision at the LHC, the initial and final state particles may emit radiation in the form of gluons, which will then hadronize resulting in jets of particles. The average production of initial and final state radiation can be estimated, but it will be inevitably affected by an uncertainty, which will in turn become a source of systematic uncertainty in the cross section measurement. This contribution can be evaluated by increasing and decreasing by one standard deviation the amount of radiation produced in the initial and final state and then compute the cross section in each case. The following values are obtained:

$$\sigma_{t\bar{t}}^{\text{ISR-up}} = 578 \pm 15 \text{ pb},$$
$$\sigma_{t\bar{t}}^{\text{ISR-down}} = 564 \pm 15 \text{ pb},$$
$$\sigma_{t\bar{t}}^{\text{FSR-up}} = 744 \pm 66 \text{ pb},$$
$$\sigma_{t\bar{t}}^{\text{FSR-down}} = 492 \pm 12 \text{ pb}.$$

As for the JES and the JER uncertainties, the ISR and FSR contributions are computed as

$$\frac{|\sigma_{t\bar{t}}^{\text{ISR-up}} - \sigma_{t\bar{t}}^{\text{ISR-down}}|}{\sigma_{t\bar{t}}^{\text{ISR-up}} + \sigma_{t\bar{t}}^{\text{ISR-down}}} = 0.012, \tag{3.14a}$$

$$\frac{|\sigma_{t\bar{t}}^{\text{FSR-up}} - \sigma_{t\bar{t}}^{\text{FSR-down}}|}{\sigma_{t\bar{t}}^{\text{FSR-up}} + \sigma_{t\bar{t}}^{\text{FSR-down}}} = 0.204. \tag{3.14b}$$

This shows that the FSR contribution amounts to 20%, much higher than the ISR contribution which is about 1.2%.

- *Integrated luminosity*: the systematic uncertainty related to the integrated luminosity measurement is determined by x-y beam scans as in [25]. For the 2016 data taking period it amounts to 2.5%.

Other sources of systematic uncertainty have not been evaluated in this analysis. Some of them are associated to colour reconnection, parton distribution functions, generator modeling and parton shower matching scales.
The main sources of uncertainty are summarized in Table 3.4, where the total systematic uncertainty is computed as the sum in quadrature of all the contributions.

Source	%
Jet energy scale	1.7
Jet energy resolution	0.2
MC Tuning	1.9
Initial state radiation	1.2
Final state radiation	20.4
Total systematic unc.	20.6
Statistical unc.	2.5
Integrated luminosity	2.5

Table 3.4: Fractional uncertainties affecting the inclusive $t\bar{t}$ production cross section measurement.

By including the systematic uncertainties, the final measurement of the inclusive $t\bar{t}$ production cross section is

$$\sigma_{t\bar{t}} = 572 \pm 14(\text{stat}) \pm 118(\text{syst}) \pm 14(\text{lumi})\ \text{pb}. \tag{3.15}$$

As it is possible to notice, the measured inclusive cross section is quite off the theoretical estimate of eq. 3.11. It should be remarked, anyway, that some sources of systematic uncertainties have not been estimated in this analysis and thus the value reported here could be quite underestimated. Furthermore, the inclusive cross section measurement depends on the capability of a MC program to correctly model the development of the parton shower and hadronization. Different models may lead to differences in the signal yield and in the selection efficiency. In this analysis it was found that the choice of the MC signal sample greatly influences the selection efficiency: variations in the measured inclusive cross section as large as 30% are obtained by switching from the nominal to the POWHEG + HERWIG MC sample. This is the hint that some systematic effects are still unknown and deeper investigations should be carried on.
By switching from NLO to NNLO simulations, improvements in the signal modeling could also be achieved, leading to a more accurate selection efficiency and to a measured inclusive cross section closer to the theoretical value.

3.5 Differential cross section measurement

In this section, the differential $t\bar{t}$ cross section measurement is presented and performed, both at *detector-level* and *parton-level*.

3.5.1 Detector-level differential cross section

The detector-level $t\bar{t}$ cross section is defined, for each bin of the spectrum of a variable of interest x, as

$$\frac{d\sigma_{t\bar{t}}}{dx} = \frac{S(x)}{L \cdot \Delta x}, \tag{3.16}$$

where L is the integrated luminosity, Δx is the bin width and $S(x)$ is the signal yield for each bin of the variable of interest x. The signal yield is computed by using

$$S(x) = D(x) - R_{yield} N_{bkg} R_{shape}(x) B(x), \tag{3.17}$$

where $D(x)$ is the distribution of x, extracted from the data by applying the reduced selection, N_{bkg} is the fitted background yield obtained in the inclusive measurement by fitting the leading jet m_{SD} distribution, and $B(x)$ is the normalized background distribution of the variable x extracted from data by selecting the control region. R_{yield} is a correction factor, computed on data, needed to switch from the phase space where N_{bkg} is measured (extended selection) to the one in which the differential measurement is performed (reduced selection). R_{yield} is thus the ratio between the number of entries in the x spectrum, obtained by asking the reduced selection on the control sample and the number of entries in the leading jet m_{SD} distribution, obtained by asking the extended selection on the control sample. Since a closure test proved that R_{yield} is rather independent from b-tagging, it is safe to compute it on the control sample (0 btag). However, the $R_{shape}(x)$ correction can be applied to switch from the control sample to the signal phase space, by correcting the differences between the shapes of the distributions. This correction factor is computed on the QCD MC samples as the ratio between the x variable spectrum obtained by asking the reduced selection and the same distribution obtained by asking the control sample selection. The ratio is fitted with a linear function which is finally used as a correction factor on the $B(x)$ distribution.

Fig. 3.16 shows the comparison between the detector-level differential cross section as a function of the first and second jet p_T, measured from the data and the theoretical distributions. These have been obtained from two different $t\bar{t}$ MC samples (POWHEG + PYTHIA 8 and POWHEG + HERWIG) by using

$$\left(\frac{d\sigma_{t\bar{t}}}{dx}\right)_{MC} = \frac{1}{N_{gen}} \frac{\sigma_{t\bar{t}}^{theor}}{\Delta x} S_{MC}(x), \tag{3.18}$$

where $S_{MC}(x)$ is the x spectrum extracted from the MC $t\bar{t}$ sample by applying the reduced selection, N_{gen} is the total number of generated events in the MC sample, $\sigma_{t\bar{t}}^{theor} = 832$ pb is the theoretical inclusive $t\bar{t}$ production cross section.
As it is possible to notice, a shift is observed between the simulations and the data distribution, with the POWHEG + HERWIG MC better describing the scale of the measured spectrum. As already seen in the inclusive cross section measurement, this effect may be related to the selection MC efficiency. The choice of the simulated sample can determine variations as large as 30% in the signal selection efficiency, which in turn may lead to a shift in the absolute normalization of the simulated differential cross section distributions.

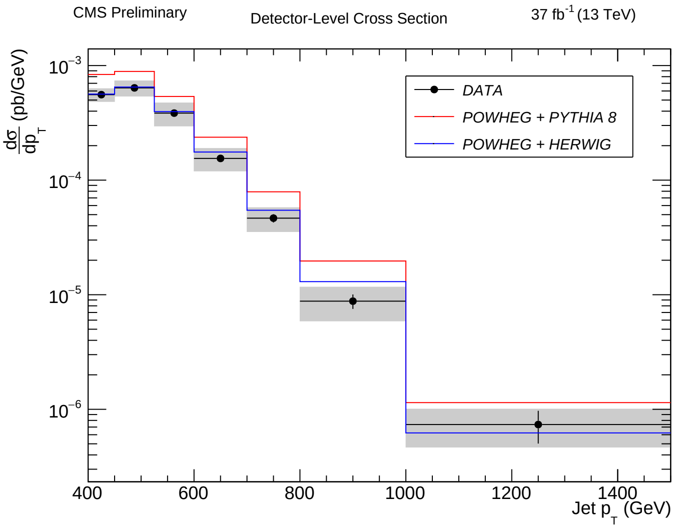


Figure 3.16: Measured detector-level differential cross section as a function of the first and second jet p_T, compared to the theoretical predictions obtained from the POWHEG + PYTHIA 8 and POWHEG + HERWIG simulated samples. The systematic uncertainty in each bin is shown as a shaded region.

In order to compare the shape of the detector-level differential cross section distribution with the simulations, a scale correction can be applied to the MC samples. Fig. 3.17, 3.18, 3.19 show the differential cross section obtained from data as a function of some variables of interest x, compared to the rescaled MC distributions, which have been computed by using

$$\left(\frac{d\sigma_{t\bar{t}}}{dx}\right)_{MC} = \frac{1}{N_{gen}} \frac{\sigma_{t\bar{t}}^{meas}}{\Delta x} S_{MC}(x), \tag{3.19}$$

where, this time, $\sigma_{t\bar{t}}^{meas}$ is the measured inclusive $t\bar{t}$ cross section assuming a certain MC sample as theoretical model. As it is possible to notice, the shape of the data distribution is in good agreement with the rescaled MC predictions, obtained from both simulated samples.

Each source of systematic uncertainty is evaluated separately in each bin via a variation of the corresponding aspect of the simulation setup, as in the inclusive measurement. For each variation, the theoretical differential cross section distribution is recalculated, and the corresponding systematic uncertainty is evaluated as for the inclusive measurement. The total systematic uncertainty is calculated separately in each bin by adding the individual contributions in quadrature. Fig. 3.20 shows the differential cross section theoretical distributions as a function of the leading and second jet p_T, obtained for each variation of the simulation setup. This plot gives an idea of the generation of the systematic error bars.

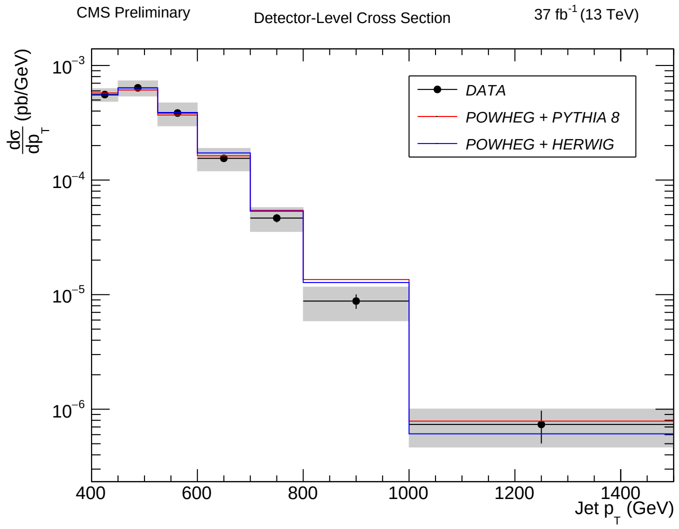


Figure 3.17: Measured detector-level differential cross section as a function of the first and second jet p_T, compared to the theoretical predictions obtained from the simulations and rescaled to the measured inclusive cross section. The systematic uncertainty in each bin is shown as a shaded region.

3.5.2 Parton-level cross section

The aim of a particle physics analysis is to estimate the true value of a certain variable or distribution, given the same quantity measured on data. This is done by removing from the measured distributions the effects of the detector, so that one can compare results obtained with a different experimental apparatus. The detector effects interfering with a measurement arise from the limited acceptance and resolution: each detector can observe particles only on a limited phase space region (geometrical acceptance) and with a limited resolution that results in a loss of events due to trigger, reconstruction and

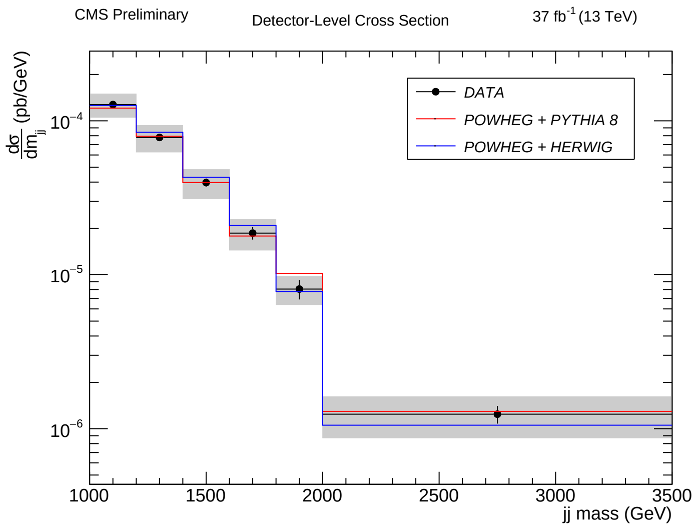


Figure 3.18: Measured detector-level differential cross section as a function of the invariant mass of the two leading jets, compared to the theoretical predictions obtained from the simulations and rescaled to the measured inclusive cross section. The systematic uncertainty in each bin is shown as a shaded region.

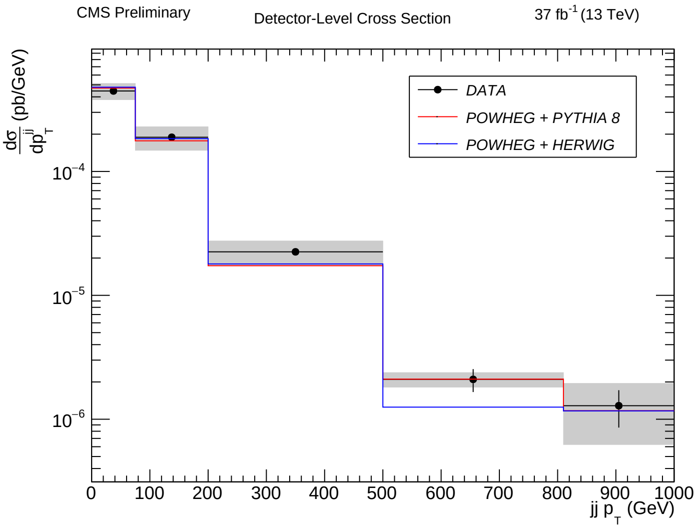


Figure 3.19: Measured detector-level differential cross section as a function of the total transverse momentum of the two leading jets, compared to the theoretical predictions obtained from the simulations and rescaled to the measured inclusive cross section. The systematic uncertainty in each bin is shown as a shaded region.

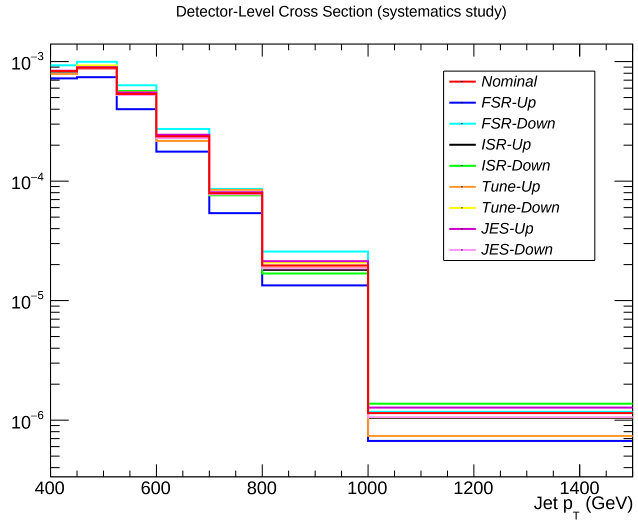


Figure 3.20: Theoretical detector-level differential cross section distributions as a function of the first and second jet p_T, obtained for each variation of the simulation setup, compared to the nominal MC prediction. This plot is shown to give an idea of the generation of the systematic error bars.

selection efficiencies. In the previous section the detector-level differential cross section has been computed on data as a function of some variables of interest x. In this section the detector effects are removed through the *unfolding* process and the true distribution is estimated, the so called *parton-level* differential cross section. In simulations, parton-level quantities are defined as the top quark related variables after QCD radiation has been simulated but before the top quark decay. The aim of this part of the analysis is thus to extrapolate the measured detector-level differential cross section to this phase space and to compare it to the theoretical distibutions coming from simulation.

Response matrix Event by event, the true value of a certain variable of interest x_{true}, generated in a certain bin, can be reconstructed, due to detector effects, into another bin. This effect is called *migration*, and its study is of utmost importance in order to perform the unfolding. If $\vec{x}_{true}$ is the vector of the bin contents of the true x spectrum, then the bin contents $\vec{x}$ of the measured spectrum can be expressed as

$$\vec{x} = A\vec{x}_{true}, \tag{3.20}$$

where A is called *response matrix* (or *migration matrix*) and it is used to take into account the bin-to-bin migration effect. This matrix needs to be evaluated on the MC signal sample. The bin-to-bin migrations are characterized by two quantities, the *stability* s^i and the *purity* p^i defined as

$$s^i = \frac{N^i_{true\&reco}}{N^i_{true}}, \tag{3.21a}$$

$$p^i = \frac{N^i_{true\&reco}}{N^i_{reco}}, \tag{3.21b}$$

namely, the stability s^i denotes the ratio between the number of events generated and correctly reconstructed in a given bin i and the events generated in that bin but reconstructed anywhere, while the purity is given by the ratio between the number of events generated and correctly reconstructed in a given bin i and the events reconstructed in that bin but generated anywhere. A response matrix can be normalized to 1 by row or column in order to explicitly show the purities and stabilities respectively. Fig. 3.21 shows the response matrices used in this analysis for some variables of interest. The true variable is on the y axis, while the reconstructed variable is on the x axis. For each matrix, the bins width has been chosen in order to have a value larger than 60% in the diagonal, both if the matrices are normalized by row or by column. The fact that A is usually not diagonal means that migration effects occur. The response matrices are computed on the MC signal sample by requiring both a *reco-level cut*, namely the reduced selection introduced in the previous sections, and a *parton-level cut*, that selects events in which the leading generated top quark p_T is greater than 400 GeV.

The unfolding process The unfolding process is the solution to the inverse problem: from the detector level distribution obtained from data, one wants to determine the best estimate of the true (parton-level) distribution by solving

$$\vec{x}_{true} = A^{-1}\vec{x}. \tag{3.22}$$

The inversion of a finite system of equations rarely admits an exact solution, so several techniques calculating approximate solutions have been developed. In this analysis the root *TUnfold* program [26] has been used. In this algorithm, the best $\vec{x}_{true}$ matching the measurement $\vec{x}$ is determined by minimizing the “Lagrangian” $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, where

$$\mathcal{L}_1 = (\vec{x} - A\vec{x}_{true})^T V_{meas}(\vec{x} - A\vec{x}_{true}), \tag{3.23a}$$

$$\mathcal{L}_2 = \tau^2(\vec{x}_{true} - f_b\vec{x}_0)^T R^T R(\vec{x}_{true} - f_b\vec{x}_0), \tag{3.23b}$$

where $\mathcal{L}_1$ is the least square minimization, $\vec{x}_{true}$ is the unfolding result and V_{meas} is the correlation matrix of the measured spectrum $\vec{x}$.

The quantity $\mathcal{L}_2$ defines the regularization, which damps non-physical fluctuations in $\vec{x}_{true}$. Such fluctuations arise from the statistical fluctuations of $\vec{x}$, which can be seen as a collection of random variables, normally distributed. These fluctuations are amplified when determining the stationary point of $\mathcal{L}$. The parameter τ gives the strength of the regularisation and it is considered as a constant while minimizing $\mathcal{L}$. The matrix R is called *regularisation matrix* and it has n columns and n_R rows, corresponding to n_R regularisation conditions. The bias vector $f_b\vec{x}_0$ is composed of a normalisation factor f_b and a vector $\vec{x}_0$. In the simpliest case, one has $f_b = 0$, $n_R = n$ and R is the unity matrix. In this case, $\mathcal{L}_2$ simplifies to $\tau^2||\vec{x}||^2$, effectively suppressing large deviations of $\vec{x}$ from zero.

When unfolding, the strength of the regularisation, τ^2, is an unknown parameter. In TUnfold a simple version of the L-curve method is implemented to determine the best value of τ. The idea of the L-curve method is to look at the graph of two variables L_x^{curve} and L_y^{curve} and locate the point where the curvature is maximal. These variables are defined as

$$L_x^{curve} = \log \mathcal{L}_1, \tag{3.24a}$$

$$L_y^{curve} = \log \frac{\mathcal{L}_2}{\tau^2}. \tag{3.24b}$$

L_x^{curve} tests the agreement of x with the data, while L_y^{curve} tests the agreement of x with the regularisation condition. Once the best value of τ is found, the unfolding is performed and the $\vec{x}_{true}$ spectrum is obtained.

In this analysis, the detector-level differential cross section distributions shown in Fig. 3.17, 3.18, 3.19 are unfolded to the parton phase space ($x_{true} = x_{parton}$). Thus, for each variable of interest, the unfolding algorithm is applied to the signal yield $S(x)$ so that the parton-level signal distribution $S_{unfold}(x_{parton})$ is obtained.

Since the parton-level differential cross section must not depend on detector design, it is necessary to take into account the effects introduced by the event selection. An *acceptance correction* $A(x_{parton})$ in then introduced in order to extrapolate the distributions to the full parton phase space. This correction is defined, starting from the $t\bar{t}$ simulation, as the ratio between the x_{parton} distribution after the reco+parton selection and the same quantity prior to any selection.
The parton level differential cross section is then computed as

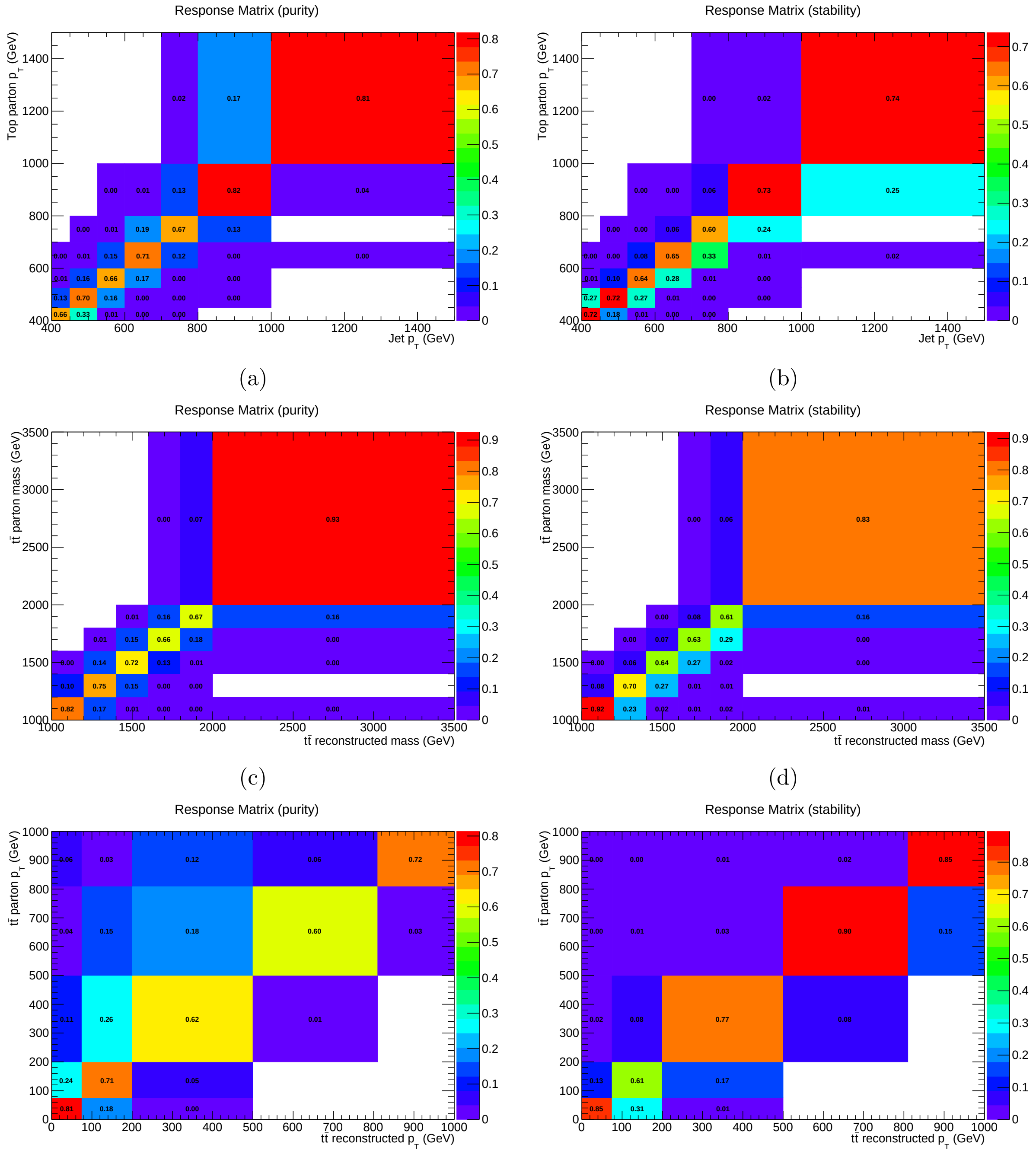
$$\frac{d\sigma}{dx_{parton}} = \frac{S_{unfolded}(x_{parton})}{L \cdot A(x_{parton}) \cdot \Delta x_{parton}}. \tag{3.25}$$

Fig. 3.22 shows the parton-level $t\bar{t}$ differential cross section as a function of the leading and second top quark p_T, compared to the theory expectations, computed on the POWHEG + PYTHIA 8 simulated sample as

$$\left(\frac{d\sigma_{t\bar{t}}}{dx_{parton}}\right)_{theor} = \frac{1}{N_{gen}} \frac{\sigma_{theor}}{\Delta x_{parton}} S_{MC}(x_{parton}), \tag{3.26}$$

where $S_{MC}(x_{parton})$ is the theoretical spectrum of any variable of interest x at parton-level.

As it is possible to notice in Fig. 3.22, the measured distribution appears to be quite off the theoretical prediction. As already seen before, this effect may be a consequence of an overestimation of the acceptance correction, which is computed with the nominal POWHEG + PYTHIA 8 simulated sample. It is then possible to rescale the measured parton-level distribution so that it does not depend anymore on the selection efficiency, multiplying it by the ratio between the theoretical and measured inclusive cross sections. In Fig. 3.23, 3.24, 3.25 the rescaled differential parton-level cross section distributions are shown as functions of some variables of interest. By rescaling the measured distribution, it is thus possible to observe a good agreement between the shapes of the measured and theoretical distributions.



(a) (b)

(c) (d)

(e) (f)

Figure 3.21: Response matrices used in this analysis to perform the unfolding. The ones on the right are normalized by row, thus showing the purities, the ones on the left are normalized by column, thus showing the stabilities.

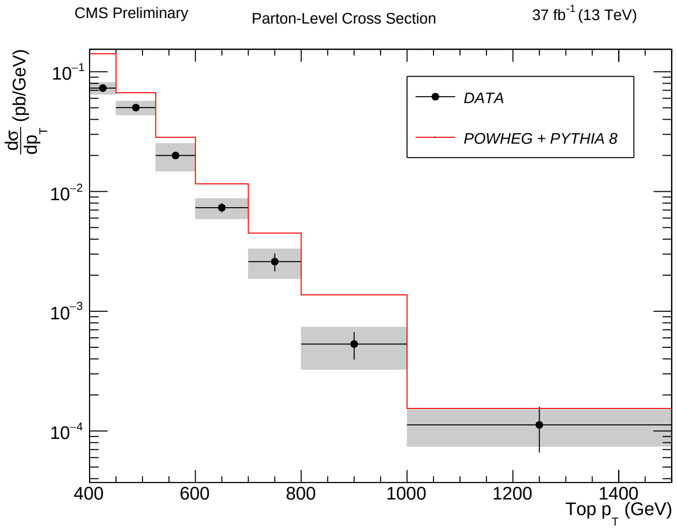


Figure 3.22: Measured parton-level differential cross section as a function of the leading and second top quark p_T, compared to the theoretical predictions obtained from the POWHEG + PYTHIA 8 simulated sample. The systematic uncertainty in each bin is shown as a shaded region.

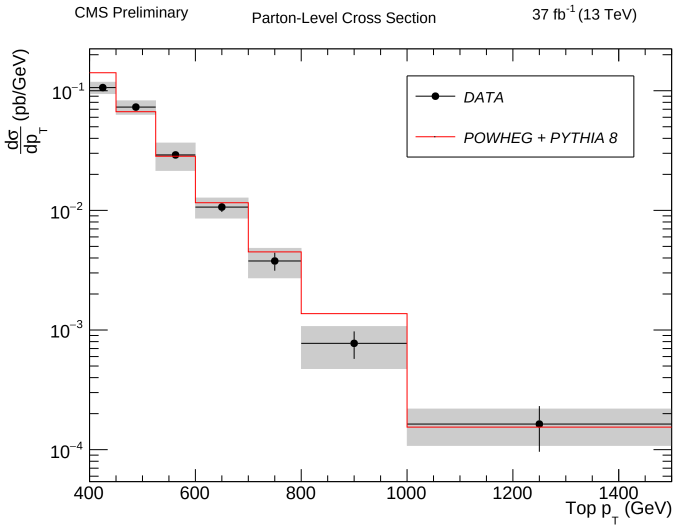


Figure 3.23: Rescaled parton-level differential cross section as a function of the leading and second top quark p_T, compared to the POWHEG + PYTHIA 8 prediction. The systematic uncertainty in each bin is shown as a shaded region.

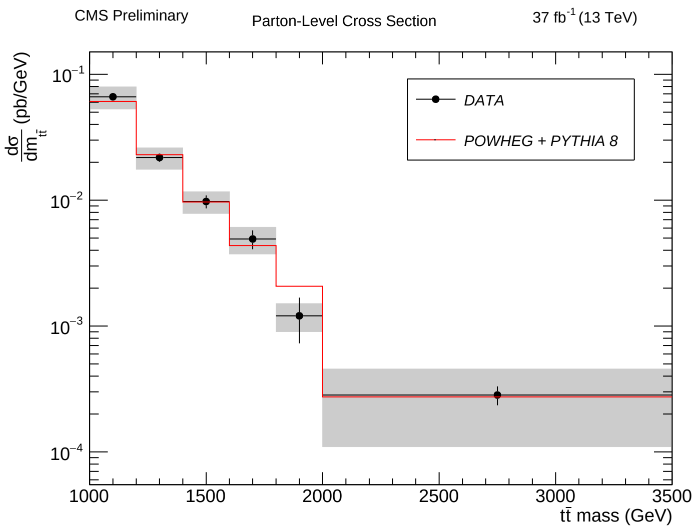


Figure 3.24: Rescaled parton-level differential cross section as a function of the $t\bar{t}$ mass, compared to the POWHEG + PYTHIA 8 prediction. The systematic uncertainty in each bin is shown as a shaded region.

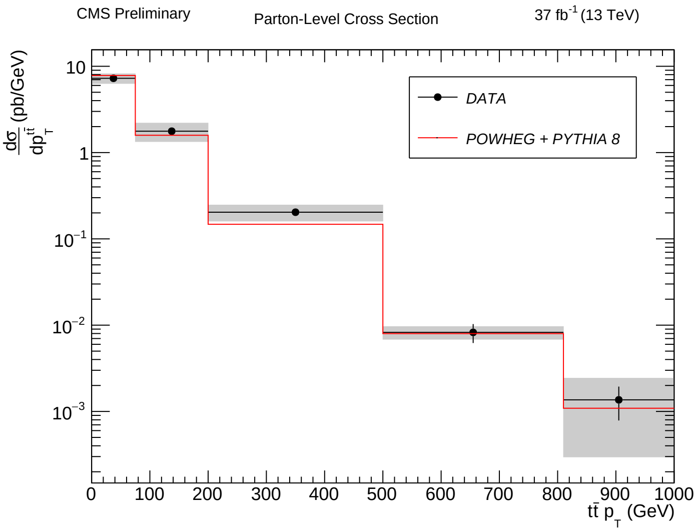


Figure 3.25: Rescaled parton-level differential cross section as a function of the $t\bar{t}$ p_T, compared to the POWHEG + PYTHIA 8 prediction. The systematic uncertainty in each bin is shown as a shaded region.

Chapter 4

Conclusions

In this work it has been presented a measurement of the inclusive and differential $t\bar{t}$ production cross section, using all-hadronic $t\bar{t}$ events with boosted topology. The data sample used in this analysis was collected in 2016 by the CMS experiment at the LHC and it amounts to an integrated luminosity of 37.0 fb^{-1}.
In order to extract the signal events from the data, an *extended* selection was applied to the event passing the signal trigger path HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV, which essentially requires two jets with transverse momentum larger than 280 GeV and 200 GeV, respectively, and at least one b-tagged jet. For each event, this selection required at least two AK8 jets, both having $p_T > 400$ GeV, $|\eta| < 2.4$, $m_{SD} > 50$, with the leading jet mass $m_{SD}^{(1)} < 300$ GeV. Moreover a *lepton veto* was required, thus selecting only events with all-hadronic final state. By using the CSV b-tagging algorithm, the two leading jets were required to contain at least one b-tagged subjet. Finally, starting from the two leading jet n-subjettiness τ_1, τ_2, τ_3, a neural network was constructed and trained on MC samples to help rejecting the background events.

Signal and background template distributions were then evaluated by fitting the leading jet SoftDrop mass distribution with parametric functions. The templates were obtained independently for 1-btag and 2-btag category events, namely events with one and two jets with a b-tagged subjet. The signal and sub-dominant background templates were evaluated by a fit on simulated samples, while the QCD background template was fitted on a QCD-enriched *control sample* extracted from data. Finally, by simultaneously fitting these templates on 1-btag and 2-btag data samples, the signal yield was obtained, which, together with the selection efficiency extracted from the $t\bar{t}$ simulated sample, allowed to compute the inclusive $t\bar{t}$ production cross section, which turned out to be $\sigma_{t\bar{t}} = 572 \pm 14(\text{stat}) \pm 118(\text{syst}) \pm 14(\text{lumi})$ pb. Despite this value being lower than the theoretical estimate, variations as large as 30% have been observed when changing the MC model. In particular, the nominal MC sample used in this analysis (POWHEG + PYTHIA 8) appears to overestimate the selection efficiency by about 30%, when compared to other simulations.

Then the differential cross section was measured by asking a *reduced selection*, namely a signal-dominated, more tight selection obtained from the extended one by applying a cut on the two leading jet masses: $140 \text{ GeV} < m_{SD}^{(1)} < 200 \text{ GeV}, \;\; 140 \text{ GeV} < m_{SD}^{(2)} < 200 \text{ GeV}$. As in the inclusive measurement, the measured *detector-level* cross section is found to be quite off the theoretical prediction because of the overestimation of the selection efficiency. However, by rescaling the simulated distribution to the measured inclusive cross section, a good agreement is found between the shapes of the measured and theoretical differential cross section.

Finally, an unfolding procedure was implemented in order to account for detector effects (such as detection efficiencies, measurement resolutions and systematic biases) and obtain the *parton-level* differential cross section, extrapolated to the full partonic phase space by using an acceptance correction evaluated on the MC sample. The distributions obtained from data are found to be quite off the theoretical predictions. However, if a possible overestimation of the acceptance correction is taken into account, the shapes of the distributions are found to be in good agreement.

Bibliography

[1] C. Patrignani *et al.* (Particle Data Group), *Review of Particle Physics*, Chin. Phys. C **40** 100001 (2016).

[2] CDF Collaboration, *Observation of top quark production in $p\bar{p}$ collisions with the Collinear Detector at Fermilab*, Phys. Rev. Lett. **74** (1995) 2626, doi: 10.1103/PhysRevLett.74.2626, arXiv:hep-ex/9503002.

[3] DØ Collaboration, *Observation of the Top Quark*, Phys. Rev. Lett. **74** (1995) 2632, doi: 10.1103/PhysRevLett.74.2632, arXiv:hep-ex/9503003.

[4] CDF Collaboration, *Measurement of the $t\bar{t}$ production cross section in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV using events with large missing transverse energy and jets*, Phys. Rev. D **84** (2011), doi: 10.1103/PhysRevD.84.032003, arXiv:1105.1806.

[5] DØ Collaboration, *Measurement of the $t\bar{t}$ production cross section using dilepton events in $p\bar{p}$ collisions*, Physics Letters B **704** (2011), doi: 10.1016/j.physletb.2011.09.046, arXiv:1105.5384.

[6] *LHC the guide*
http://cds.cern.ch/record/1165534/files/CERN-Brochure-2009-003-Eng.pdf.

[7] http://home.web.cern.ch/topics/large-hadron-collider.

[8] CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST **3** (2008) S08004.

[9] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[10] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[11] T. Sjöstrand, S. Mrenna, and P. Skands, *PYTHIA 6.4 physics and manual*, JHEP **05** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[12] T. Sjöstrand, S. Mrenna, and P. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[13] CMS Collaboration, *Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of $t\bar{t}$ at $\sqrt{s} = 8$ and 13 TeV*, CMS Physics Analysis Summary CMS-PAS-TOP-16-021, 2016.

[14] J. Alwall *et al.*, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[15] S. Agostinelli *et al.*, *Geant4 a simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250, doi:10.1016/S0168-9002(03)01368-8.

[16] M. Cacciari *et al.*, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, doi: 10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[17] F. Beaudette (the CMS Collaboration), *The CMS Particle Flow Algorithm*, Proceedings of the CHEF2013 Conference - Eds. J.C. Brient, R. Salerno, and Y. Sirois - p295 (2013), ISBN 978-2-7302-1624-1, arXiv:1401.8155.

[18] J. Thaler, K. Van Tilburg *Identifying Boosted Objects with N-subjettiness*, JHEP **1103** (2011) 015, doi: 10.1007/JHEP03(2011)015, arXiv:1011.2268.

[19] A. Larkoski *et al.*, *Soft Drop* JHEP **05** (2014) 146, doi: 10.1007/JHEP05(2014)146, arXiv:1402.2657.

[20] CMS Collaboration, *Top Tagging with New Approaches*, CMS Physics Analysis Summary CMS-PAS-JME-15-002, 2016.

[21] CMS Collaboration, *Identification of b quark jets at the CMS experiment in the LHC Run2 Startup*, CMS Physics Analysis Summary CMS-PAS-BTV-15-001, 2016.

[22] *ROOT TEfficiency class reference*,
URL : https://root.cern.ch/doc/master/classTEfficiency.html.

[23] *Toolkit for Data Modeling with ROOT (RooFit)*, available at root.cern.ch/roofit.

[24] M. Czakon, A. Mitov, *Top++: a program for the calculation of the top-pair cross-section at hadron colliders*, Comput. Phys. Commun. 185 (2014) 2930, doi: 10.1016/j.cpc.2014.06.021, arXiv:1112.5675.

[25] CMS Collaboration, *CMS Luminosity Measurements for the 2016 Data Taking Period*, CMS Physics Analysis Summary CMS-PAS-LUM-17-001, 2017.

[26] S. Schmitt, *TUnfold: an algorithm for correcting migration effects in high energy physics*, doi: 10.1088/1748-0221/7/10/T10003, arXiv:1205.6201.