Data-driven Background Estimation for Top-pair Searches with the CMS Detector

von
Till Michael Arndt

Bachelorarbeit in Physik

vorgelegt der
Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen

im
August 2012

an gefertigt im
III. Physikalischen Institut B

bei
Prof. Dr. Achim Stahl
Ich versichere, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie Zitate kenntlich gemacht habe.

Aachen, den
# Contents

1. **Introduction** ............................................. 1  

2. **The Standard Model** ..................................... 3  
   2.1 The Top Quark ........................................... 3  
   2.2 Backgrounds for Top-pair Events ....................... 5  

3. **The CMS Experiment at the LHC** ....................... 9  
   3.1 The Compact Muon Solenoid .............................. 10  

4. **Collision Data, Event Simulation and Selection** ...... 13  
   4.1 Collision Data and Event Simulation ................... 13  
   4.2 Event Selection and Reconstruction .................... 15  

5. **Data-driven Background Estimation** ................... 17  
   5.1 Discriminating Variables ............................... 18  
   5.2 Software ................................................ 21  
   5.3 Likelihood Fit with ROOT ............................... 21  
   5.4 Likelihood Fit with RooFit ............................. 23  
   5.5 Measurement of the Ratio of the W + jets and Top-pair Cross Sections ..................................... 23  
   5.6 Estimation of Uncertainties: Ensemble Study .......... 25  

6. **Summary and Outlook** ................................... 29  

7. **Appendix** ................................................... 31  
   A Additional Plots .......................................... 31  
   B Summary of the Files for Data and Simulation .......... 31  
   C Acknowledgements ......................................... 34  

**Bibliography** .................................................. 35
Chapter 1

Introduction

Since the end of 2009 the Large Hadron Collider (LHC) near Geneva, Switzerland has been acquiring collision data, aiming at improving the understanding of particle physics and especially the standard model. One remaining question was the existence of the Higgs boson, expected to explain the mass of the W and Z gauge bosons. This question seems to have been answered by the results of the CMS and ATLAS experiments, which were presented on July 4th, 2012: A candidate for the Higgs boson was found at a mass of about 125 GeV [1] [2].

One of the elementary particles in the standard model is the top quark, discovered in 1995 at the Fermilab [3], which has the highest mass of all discovered elementary particles with $m_{\text{top}} = 173.5 \pm 0.6 \pm 0.8 \text{ GeV}$ [4].

The top quark decays too fast to form hadrons and almost exclusively decays into a W boson and a bottom quark. This study is focused on the semi-muonic decay channel.

One important part of any analysis is to separate the signal events, in this case top-pair candidates, from background events that leave a similar signature in the detector. It is essential to estimate the amount of remaining background events in the final selected candidate collection. This is done by using theoretical predictions for the cross sections of the background processes. These are often known with little accuracy and suffer from uncertainties in the simulation. Alternatively, the background fraction can be estimated from data which leads to a more robust analysis. In this study a method for simultaneous fitting of templates of discriminating variables is tested by determining the amount of background in semi-muonic top-pair events. The results are described and briefly discussed. This study was performed together with Florian Lenz [5].
System of Units

This analysis uses natural units ($\hbar = c = 1$). Units of common observables in this convention are

$$\begin{align*}
[\text{Energy}] &= [\text{Momentum}] = [\text{Mass}] = [\text{Time}]^{-1} = [\text{Length}]^{-1} = \text{eV}.
\end{align*}$$
Chapter 2

The Standard Model

The standard model describes the elementary parts of matter as fermions (particles with spin 1/2) and three fundamental interactions by gauge bosons. The group of fermions consists of 6 leptons and 6 quarks, which can be arranged into three generations as shown in Table 2.1.

Table 2.1: Fermions of the standard model [6]. $T_3$ denotes the appropriate equivalent to strong isospin for the second and third generation quarks.

| Fermions | I         | II        | III       | $Q|e|$ | $T_3$ | $Y$ | Color |
|----------|-----------|-----------|-----------|-------|-------|-----|-------|
| Leptons  | $\nu_e$,$e^-_L$ | $\nu_\mu,\mu^-_L$ | $\nu_\tau,\tau^-_L$ | 0     | 1/2   | -1  | -     |
|          | $\nu_\mu,R$ | $\mu^-_R$  | $\nu_\tau,R$ | -1    | -1/2  | -1  | -     |
|          | $e^-_R$    | $\mu^-_R$  | $\tau^-_R$  | 0     | 0     | 0   | -     |
|          | $\nu_e,R$  | $\nu_\mu,R$| $\nu_\tau,R$| -1    | 0     | -2  | -     |
| Quarks   | $u$,$d_L$  | $c$,$s_L$  | $t$,$b_L$  | $2/3$ | 1/2   | 1/3 | rgb   |
|          | $u_R$      | $c_R$      | $t_R$      | $-1/3$| -1/2  | 1/3 | rgb   |
|          | $d_R$      | $s_R$      | $b_R$      | 2/3   | 0     | 4/3 | rgb   |
|          |            |            | $-1/3$     | 0     | 2/3   |     |       |

2.1 The Top Quark

The existence of the top quark was predicted in 1973 [7], to complete the fermionic content of the standard model, as listed in Table 2.1. Among the six quarks, the top quark’s existence was established experimentally in 1995 [3].

At the LHC the dominant top pair production processes are gluon-gluon fusion ($\sim 90\%$) and quark-anti-quark annihilation ($\sim 10\%$) as shown in Figure 2.1[8]. The probabilities mentioned here are for a center of mass energy of $\sqrt{s} = 14$ TeV that has not yet been reached at the LHC. At $\sqrt{s} = 7$ TeV gluon-gluon fusion makes up about 80 % of the top-pair production [9].
Both top quarks decay into a b quark and a W boson. W bosons can either decay hadronically into two quarks, or leptonically into a lepton and a corresponding neutrino. The fully hadronic decay channel has the benefit of an unambiguous kinematic reconstruction, as the final state objects (six hard jets) are completely determined. Moreover, this channel has a high branching ratio ($\sim 46\%$). However, the events are hard to differentiate from other multijet events, especially QCD multijet events. There is also a high uncertainty in the measurement of the jet-energy scale (JES) and an additional problem with the event reconstruction due to combinatorial jet ambiguities.

The dileptonic channel on the other hand has a clear signature due to the two isolated leptons, but its branching ratio is lower ($\sim 10\%$) and the two neutrinos are hard to reconstruct after the measurement, due to the fourfold ambiguity. The semi-leptonic channel can be used as a compromise with a branching ratio of ($\sim 44\%$) and a unique signature. This study only uses the semi-muonic decay channel (see Figure 2.2). The branching ratio of this channel is approximately $15\%$ [11]. It results in two b quarks jets and two quarks that are measured as jets in the detector. Additionally, there is an isolated muon, a neutrino and thus missing transverse energy (MET) in the event. The transverse energy of all particles has to be balanced. If this is not the case, it can be assumed particles were not detected.

Figure 2.3 shows a typical reconstructed $t\bar{t}$-event candidate from data collected in 2011 by the CMS detector.
2.2 Backgrounds for Top-pair Events

The relevant backgrounds for this study are single-top decays, W boson events with jets (W + jets), Z boson events with jets (Z + jets/ Drell-Yan) and QCD multijet events. The cross section of these processes can be found in Figure 2.8. Top-pair events have a significantly lower cross section than the background processes. Especially QCD multijet events are relevant, because of the high cross section.
A QCD multijet event, depicted in Figure 2.4, naturally results in multiple measured jets. If one of those jets is misidentified as an isolated muon or a muon from a jet is assumed to be isolated and the detector additionally measures some MET, the event can fake a semi-muonic top pair decay.

![Feynman graph of QCD multijet event](image)

**Figure 2.4:** Feynman graph of QCD multijet event, "e" denotes a muon [10].

As shown in Figure 2.5, a W + jets event might be identified as a semi-muonic t\bar{t} decay if one of the jets results from a b quark. It might also be another quark, which is wrongly considered to be a b quark.

![Feynman graph of a W + jets event](image)

**Figure 2.5:** Feynman graph of a W + jets event [10].
2.2. BACKGROUNDS FOR TOP-PAIR EVENTS

A $Z + \text{jets}$ event, as shown in Figure 2.6, might be regarded as a semi-muonic $\bar{t}t$ event if one of the particles from the $Z$ decay is not identified correctly. This could result in a contribution to MET. Two additional jets would be needed, for example from gluon radiation.

Figure 2.6: Feynman graph of a $Z + \text{jets}$ event [10].

Single-top events have a lower cross section than the backgrounds mentioned before, as they are produced by weak processes. Nevertheless, their top-content produces a signature very similar to the $\bar{t}t$-signature, as shown in Figure 2.7.

Figure 2.7: Feynman graph of a single-top event [10].
Figure 2.8: Cross section of important processes at hadron colliders [12], reworked in [13].
Chapter 3

The CMS Experiment at the LHC

The LHC is operated by the European Organization for Nuclear Research (CERN) and is located near Geneva. It is a collider ring for protons and heavy ions, with a circumference of 26.7 km. There are currently four major experiments at the LHC, shown in Figure 3.1: The two multipurpose detectors ATLAS and CMS, LHCb, a detector specifically for b-quark measurements, and ALICE, focused on heavy-ion collisions.

![Overall view of the LHC experiments.](image)

Figure 3.1: The LHC ring and its four experiments [14].

It was designed for a maximum center of mass energy of $\sqrt{s} = 14$ TeV. The protons are injected into the collider after having been accelerated to the energy of 450 GeV. These protons are amassed in 2808 bunches with approximately $1.15 \times 10^{11}$ protons each, according to design, which leads to a collision every 25 ns [15]. In 2012, the LHC reached a center of mass energy of $\sqrt{s} = 8$ TeV. For the data from 2011, used in this analysis, the protons were accelerated to reach $\sqrt{s} = 7$ TeV. This step by step increase of the energy was introduced after problems with the cooling system lead to major damage to the collider in 2009.
CHAPTER 3. THE CMS EXPERIMENT AT THE LHC

3.1 The Compact Muon Solenoid

The CMS detector, shown on Figure 3.2, has a length of 21.6 m, a diameter of 14.6 m and a weight of 12500 t [16].

![Figure 3.2: An exploded view of the CMS detector [16].](image)

The inner tracking detector consists of silicon pixel and stripe detectors. By analyzing early results of the CMS experiment, the momentum resolution was found to lie at approximately 1% for 1 GeV and 5% for 1 TeV. The impact parameter resolution for high-momentum tracks is close to 10 µm [17].

The inner part of the electromagnetic calorimeter (ECAL) is a preshower calorimeter based on silicon strip detector layers, while the rest of the ECAL consists of scintillator crystals made of lead tungstate [18]. Its energy scale was found to agree with simulation within 1% in the barrel and within 3% in the endcaps [19].

The hadronic calorimeter (HCAL) encloses the ECAL. Its resolution was measured by testbeams. The resolution is described by the following formula over the entire range [20]:

\[
\frac{\sigma(E)}{E} = \frac{1}{\sqrt{E}} \oplus 0.045 \text{ (E in GeV)}
\]

The solenoid, situated between the calorimeter and the muon system, provides a magnetic field of approximately 4 T in order to facilitate precise measurements of momenta of charged particles by the bent of their tracks.

The muon system enables high-resolution reconstruction of muons, despite the small dimensions of the detector. Its resolution depends on the transverse momentum of the particle, which can be combined with the resolution of the calorimeters [21]:

\[
\frac{\Delta p_T}{p_T} = 0.045 \cdot \sqrt{p_T} \text{ (p_T in TeV)}
\]

The CMS experiment uses a three-tiered trigger system to decrease the amount of stored data: The first tier consists only of the detector electronics, while the L1 hardware triggers (second tier) have access to the data from calorimeter and muon system. They use basic
3.1. THE COMPACT MUON SOLENOID

definitions for particle candidates to select the most interesting events. The software-based high-level triggers (HLT) are simplified versions of the final reconstruction algorithms. Through these steps the rate of interactions written to storage media is decreased to approximately 300 Hz [22].

As of June 18th, 2012, the CMS experiment reached an integrated recorded luminosity of $\mathcal{L} = 6.65 \text{ pb}^{-1}$, which is already more than in 2011 [23].
CHAPTER 3. THE CMS EXPERIMENT AT THE LHC
Chapter 4

Collision Data, Event Simulation and Selection

4.1 Collision Data and Event Simulation

The data used in this study was recorded in 2011 at a center of mass energy of $\sqrt{s} = 7$ TeV. The corresponding luminosity is $\mathcal{L} = 4.6 \, \text{fb}^{-1}$. Detailed information on the datasets can be found in section 7.

For the background estimation, we used samples officially simulated by the CMS experiment (they were e.g. also used in [24]). The simulation chain employs several steps in order to achieve a full model of the recorded data. It starts with the matrix-element generator MADGRAPH [25] applying the leading-order parton distribution function set CTEQ6L1 [26]. TAUOLA was utilized to model the decay of tau leptons [27]. These generators are based on perturbation theory. The result is a list of the outcoming and incoming particles including their momenta and energies. Afterwards, PYTHIA is used to model the hadronization and fragmentation process, the initial and final state radiation. They are based on phenomenological models, as they cannot be calculated from perturbation theory [28]. The MLM Matching procedure [29] is used to prevent double counting in the showers from matrix element generation and PYTHIA showering. In order to improve the generated events $Z2$ has been used [30] [31]. For the simulation of the detector output GEANT4 was employed [32].

Because of pileup in the data events, the simulated events have to be overlayed with a random number of pileup vertices. Subsequently, they have to be re-weighted (see also section 7).

Besides the background samples mentioned in chapter 2.2 a $t\bar{t}$ sample is generated. The background events other than the single-top events were generated with PYTHIA. The single-top events were generated with POWHEG [33] [34].
The following cross sections were the result of calculation (for more details on the used files see section [7]):

**Table 4.1:** Cross sections obtained from simulation.

<table>
<thead>
<tr>
<th>Sample</th>
<th>$t\bar{t}$</th>
<th>W + jets</th>
<th>Z + jets</th>
<th>$t_{\bar{s}}$</th>
<th>$\bar{t}_{\bar{s}}$</th>
<th>$t_{t\bar{W}}$</th>
<th>QCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma ,[pb]$</td>
<td>157</td>
<td>31300</td>
<td>3000</td>
<td>2.7</td>
<td>1.5</td>
<td>42.6</td>
<td>22.0</td>
</tr>
<tr>
<td>$\Delta \sigma ,[pb]$</td>
<td>24</td>
<td>1600</td>
<td>100</td>
<td>0.1</td>
<td>0.1</td>
<td>2.4</td>
<td>0.1</td>
</tr>
<tr>
<td>Order</td>
<td>NLO</td>
<td>NNLO</td>
<td>NNLO</td>
<td>NLO</td>
<td>NLO</td>
<td>NLO</td>
<td>NLO</td>
</tr>
</tbody>
</table>

The cross sections listed in Table 4.1 are generated by numerical methods using perturbation theory. The generated error is assumed to be an underestimation. It is only an approximation to a finite order of perturbation theory. Consequently, the errors for W + jets and Z + jets events can be assumed to be about 30%. For QCD multijet events the error is estimated at 50% [35].

At CMS the following results were measured [36] [35] [37]:

$\sigma_{W+\text{jets}} = 30.9 \pm 1.2 \,(\text{lumi.}) \text{nb}$ \hspace{1cm} $\sigma_{Z+\text{jets}} = 2.9 \pm 0.1 \,(\text{lumi.}) \text{nb}$

$\sigma_{\text{Singletop}} = 84 \pm 30 \,(\text{stat. + sys.}) \text{pb}$ \hspace{1cm} $\sigma_{t\bar{t}} = 154 \pm 17 \,(\text{stat. + sys.}) \text{pb}$

**Figure 4.1:** Stacked jet multiplicity (simulation scaled to data luminosity).

The stacked jet multiplicities, depicted in Figure 4.1, show a clear excess of simulated events in comparison to data. The simulated samples are scaled to data luminosity, so they should be in accordance. Therefore, a correction of the cross section for each of the single backgrounds is not possible. This excess can be explained with the high uncertainties of the simulated cross sections. In addition, the luminosity of the measured data has an uncertainty of about 10%. Finally, the uncertainties of the jet-energy scale of about 5 − 10% could lead to a different selection efficiency for measured data and simulated events.

1 Only the leading error is given.
It is also visible that there are hardly any QCD multijet events left after the selection steps. Moreover, it is important to note that our simulated $\bar{t}t$ sample contains all $\bar{t}t$-decay channels.

## 4.2 Event Selection and Reconstruction

For reconstruction, the particle flow algorithm [38] is used. Further criteria for the physics objects have to be utilized corresponding to the specific analysis. The selection and reconstruction applied in this study is based on a CMS top reference selection [24] [37].

The vertices at CMS are reconstructed from tracks [39]. The tracks are associated with the vertices through a fit. Besides a successful fit ($n_{dof} > 4$), a close proximity to the nominal interaction point ($|z| < 24\text{ cm}, \rho < 2\text{ cm}$) is required for an object to be assigned to the hard process. During additional reconstruction steps all charged objects connected to pileup vertices are rejected.

The selection requires muons to be identified in the muon system (at least one hit), as well as in the inner tracking system (at least ten hits) [40]. The muon candidate should be isolated to ensure that it does not originate from a jet. Combined isolation is defined as follows:

\[
\text{relIso} = \frac{\text{hadronIso} + \text{photonIso}}{p_T} < 0.125
\]

The variables hadronIso and photonIso denote the summed $p_T$ of all hadrons and photons respectively within a distance of $\Delta R < 0.4$. Furthermore, muons should have a distance to the next jet of $\Delta R > 0.3$. In this analysis the unprescaled muon trigger IsoMu24 is used. Thus the muon is required to have a $p_T$ larger than 26 GeV, to be in the plateau of the trigger turn on curve.

An additional veto on further muons, using the so called 'loose muon Id', serves to suppress background, e.g. Z boson decays. The 'loose muon Id' requires $p_T > 10\text{ GeV}, |\eta| < 2.5$ and relIso < 0.2. An electron veto is applied to suppress further background from dileptonic top-pair events. Events with at least one electron with $E_T > 15\text{ GeV}, |\eta| < 2.5$ and relIso < 0.2 are rejected. The reconstruction algorithm is described in detail in [41].

Quarks or gluons from the hard process are detected as high-energetic jets. A jet is regarded as an object of collimated hadrons and leptons originating from the original parton, which are being produced along the hadronization process. This allows to deduce from the jet observables some of the original parton properties. The non-isolated hadrons are sorted into jets by the particle-flow algorithm after all isolated particles have been reconstructed using the anti-$k_T$ algorithm (described in [42]). Due to uninstrumented parts of the detector or neutrino emission, the energy disposition in the hadronic calorimeter does not exactly match the energy of the original jet. For the purpose of correcting these effects, CMS uses a factorized approach (effects from non-linear calorimeter, pileup etc., described in [43]).
In order to distinguish from noise, this analysis requires the hadron energy fraction to be larger than zero. Consequently, the neutral hadron fraction should be smaller than 0.99. Moreover, the jets associated with the hard process should have a $p_T > 30$ GeV and $|\eta| < 2.4$. Jet-smearing corrections were not applied. This study does not use a $b$ tag, as a high percentage of background is important for the method of background estimation used here.
Chapter 5

Data-driven Background Estimation

For a precise measurement in the semi-muonic decay channel, the uncertainties of the backgrounds have to be minimized. Especially, the cross sections of the backgrounds suffer from large uncertainties. In contrast, the distribution of the kinematic properties are modeled with more accuracy. Therefore observables with a discriminating shape can be used to correct the cross section by fitting simulated templates to measured data. This study uses two variables, MET and M3, which are described in section 5.1.

To perform the fit the binned and extended maximum likelihood method is used:

$$L = e^\nu \cdot \nu^N \prod_{i=1}^{m} \frac{p_i(x_i, a_i)^{k_i}}{k_i!}$$  \hspace{1cm} (5.1)

$p_i(x_i, a_i)$ is the poisson-distributed probability for events to appear in bin $i$ with bin center $x_i$. $k_i$ denotes the number of entries in bin $i$, $m$ the number of bins, $N$ the total number of entries. In an extended, binned maximum-likelihood fit, the expected number of events $\nu$ is a free parameter. This method is explained thoroughly in [5].

The following function was used to obtain the value $a_i$ in the likelihood fit:

$$a_j[i] = \sum_k \beta_k \cdot \alpha_{jk}[i]$$  \hspace{1cm} (5.2)

Here $i$ denotes the bin, $j$ the variable, $\alpha$ the prediction through simulation in the respective bin. $\beta_k$ is the free parameter which can be interpreted as a direct contribution from a single process $k$. The value of $\beta_k$ can then be used to determine the cross section of the respective process.

The method is tested by fitting the stacked background templates to data. A fit for each of the templates is impossible because of the high deviations between simulated and measured data. Therefore, the ratio of the cross sections of W + jet and t+$\bar{t}$ events is calculated. W + jets events are the dominant background process and have a high uncertainty. QCD multijet events have an even higher cross-section uncertainty, but the amount of these events after the selection is to small to apply a data-driven approach. Details of the fitting procedure can be found in [5].

Some of the figures in the next section show ratio plots. The errors of all histograms depicted are taken into account for these plots. The label on the y-axis denotes the ratio between simulated data and measured events when reading 'Data/MC'.
5.1 Discriminating Variables

Motivated by reference analyses of the CMS collaboration, this study uses the variables M3 and MET [44]. MET denotes the missing transverse energy that is caused, amongst other effects, by neutrinos in the \( t\bar{t} \) events. It should discriminate between events with a neutrino and QCD multijet respectively \( Z + \) jets events where the MET originates from non-detected particles.

![Figure 5.1: MET: non-stacked simulated events.](image)

In Figure 5.1 it is clearly visible that MET, while hardly discriminating for \( W + \) jets and single top events, shows the expected difference between events with a neutrino and the ones without a neutrino. MET was reconstructed with the particle flow algorithm [45]. It discriminates very well for \( Z + \) jets background events.

![Figure 5.2: M3: non-stacked simulated events.](image)
5.1. Discriminating Variables

The second variable M3, as shown in Figure 5.2, is the invariant transverse mass of the sum of the three jets with the highest transverse momentum. It approximates the top mass for the semi-muonic $t\bar{t}$ events. Therefore, it should have a different shape for the non-top like background events. The discriminating power is certainly visible. The double peak structure for the simulated $t\bar{t}$ sample between 80 GeV and 200 GeV is suitable for discriminating between $t\bar{t}$ events and background events. The peak at $\sim 190$ GeV approximates the top mass, but the smaller peak at $\sim 90$ GeV is investigated. It is caused by events where the combinatorial jet association for M3 failed. Figure 5.3 investigates the effect using generator information:

![Figure 5.3: M3 for non semi-muonic decays from $t\bar{t}$ sample.](image)

The shape of M3 for non semi-muonic $t\bar{t}$ decays shows only the peak at $\sim 190$ GeV. These events make up $\sim 12\%$ of the $t\bar{t}$ sample.

Figure 5.4 shows M3 for semi-muonic $t\bar{t}$ decays where the three quarks with the highest $p_T$ all originate from the hadronically decaying top quark. We assume that the $p_T$ of the quark approximates the $p_T$ of the resulting jet. As expected, M3 has a very clear peak at $\sim 190$ GeV, but these events only make up $\sim 19\%$ of the whole $t\bar{t}$ sample.

![Figure 5.4: M3, quarks from hadronic top decay have highest $p_T$.](image)
Chapter 5. Data-driven Background Estimation

Figure 5.5: $M_3$, if the $b$ quark from the leptonic top decay contributes.

The remaining $t\bar{t}$ sample ($\sim 69\%$) consists of events where the $b$ quark from the leptonically decaying top quark belongs to the three quarks with the highest $p_T$. The shape of $M_3$ in Figure 5.5 shows the double peak and the plateau between them. As these events make up the majority of the $t\bar{t}$ sample they dominate the final shape.

Figure 5.6: Stacked templates (simulated samples scaled to data luminosity).

Figure 5.6 shows the templates for $M_3$ and MET. The simulated templates are stacked and normalized according to the data luminosity of $L = 4.6 \text{ fb}^{-1}$. The shapes are compared to data.

The shapes for measured data and simulated events do not agree. The most obvious deviation is the already mentioned excess of simulated events by about 15%. The double-peak structure is less clear in the measured data. This might be a smearing effect.

The values over 500 GeV are disregarded because they do not provide additional separation power. Nevertheless, the complete regime up to 800 GeV for $M_3$ is included in a separate fit to estimate systematic effects on the final results due to fit constraints.

In the end MET and $M_3$ provide sufficient discriminating power.
5.2 Software

ROOT is a framework for physics analyses developed at CERN. In this study two different ROOT packages are used for fitting simulated to measured templates. The native ROOT package allows fits to use the maximum-likelihood method, but the errors of the simulated template histograms are not taken into account. These errors are incorporated with an ensemble study. The package RooFit is optimized for template fits and is supposed to consider the errors of the simulated templates [46]. Both packages use the extended and binned maximum-likelihood method as described in equation 5.1.

5.3 Likelihood Fit with ROOT

The stacked simulated templates of the added backgrounds and $\text{t}\bar{\text{t}}$ are simultaneously fitted to measured data for M3 and MET. The procedure is described in more detail in [5].

![Figure 5.7: Simultaneous fit for M3 and MET using ROOT.](image)

The result is shown in Figure 5.7:

\[
\text{Background : } \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} = 0.93 \pm 0.02 \quad \text{t}\bar{\text{t}} : \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} = 0.75 \pm 0.02
\]

The fit is able to discriminate. However, the fitted templates and the data do not agree for M3 in the region below 150 GeV. The difference between fit result and measured data is significant. This might be due to the jet-energy scale already mentioned in section 5.1. For further investigation, a different binning has been used for M3 (see Figure 5.8).
Chapter 5. Data-driven Background Estimation

Figure 5.8: Simultaneous fit for M3 and MET using ROOT with different binning for M3.

The fit for a rebinned M3 shows the same excess of data in the region below 150 GeV. It also shows a systematic deviation (especially in the ratio plot). Consequently, this effect is not due to the binning or the resolution. It might be due to the uncertainty of the jet-energy scale. To investigate this we increased M3 by 5% as shown on Figure 5.9.

Figure 5.9: Simultaneous fit for M3 and MET using ROOT with different binning and 5% increase for M3.

The test shows that the deviations in the shapes of data and simulated events for the variable M3 might correspond to the jet-energy scale. Nevertheless, a more thorough investigation is needed to confirm or refute this assumption.

One of the remaining issues are the uncertainties of the template histograms, which have not been taken into account yet.
5.4 Likelihood Fit with RooFit

The simultaneous fit with RooFit gives the following result, as shown in Figure 5.10:

\[
\begin{align*}
\text{Background: } \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} &= 0.93 \pm 0.02 \\
\text{t\bar{t}: } \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} &= 0.75 \pm 0.02 
\end{align*}
\]

We are unable to completely trace the estimation of the uncertainties done by RooFit. Consequently, we decide not to trust the errors given by RooFit. A more detailed explanation can be found in [5].

Still, the results from ROOT and RooFit are consistent, thus the cross check was successful. The fit with ROOT is used for the estimation of the ratio of the W + jets and t\bar{t} cross section.

5.5 Measurement of the Ratio of the W + jets and Top-pair Cross Sections

In order to calculate the ratio of the W + jets and t\bar{t} cross sections, only the contributions of the templates for these two processes are fitted to data as a free parameter \( \beta_k \). The result is shown in Figure 5.11. The remaining background templates are scaled according to luminosity. Due to excess of the simulated events and the not exactly corresponding shapes, the templates for the backgrounds are not included into the fit as free parameters. As a test the remaining three backgrounds are stacked into a third template and included into the fit, but the result did not improve as shown in Figure 7.1. Furthermore, the remaining third template is scaled up, despite the clear excess of simulated events, so this result cannot be trusted.
The result with fixed backgrounds is, as shown in Figure 5.11:

\[
W + \text{jets}: \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} = 0.88 \pm 0.03 \quad \bar{t}\bar{t}: \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} = 0.75 \pm 0.02
\]

The result of a simultaneous fit with M3 restricted to 800 GeV, as shown in Figure 5.12, is:

\[
M3 : 500 \text{ GeV} \Rightarrow 800 \text{ GeV} \quad W + \text{jets}: \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} = 0.82 \pm 0.03 \quad \bar{t}\bar{t}: \frac{\sigma_{\text{meas.}}}{\sigma_{\text{simu.}}} = 0.78 \pm 0.02
\]
5.6. **Estimation of Uncertainties: Ensemble Study**

The resulting ratio is:

\[ M3 : 500 \text{ GeV} \Rightarrow 800 \text{ GeV} \quad \frac{\sigma_{W+\text{jets}}}{\sigma_{t\bar{t}}} = 209 \pm 9 \]

The ratio plot for M3 shows systematic deviations for higher energies, which justify the narrow restriction that was implemented. Nevertheless, the fit with the wider restriction for M3 will be used to estimate a systematic uncertainty of the ratio between the W + jets and t\bar{t} templates.

A dominant source of systematic uncertainty is due to the luminosity. This uncertainty can be cancelled in this analysis by using the ratio of cross sections. Consequently, the ratio of the cross section of the W + jets and t\bar{t} cross sections will be used for our result. Additionally, the ratio should mitigate the effects from the excess of simulated events described in section 4.1.

To calculate the ratio the cross-sections values obtained from simulation are used:

\[ \sigma_{t\bar{t}} = 157 \, \text{pb} \quad \sigma_{W+\text{jets}} = 31300 \, \text{pb} \]

The resulting ratio is:

\[ \text{Measurement : } \frac{\sigma_{W+\text{jets}}}{\sigma_{t\bar{t}}} = 235 \pm 10(\text{stat.}) \pm 26(\text{sys.}) \]

The systematic uncertainty due to the fitting range is estimated by considering the difference between the ratio of the fit where M3 has been restricted to a maximum of 500 GeV (see Figure 5.11) and the one where it has been restricted to a maximum of 800 GeV (see Figure 5.12). The statistic effects contributing to the difference between these two fits are disregarded in the estimation of the systematic error.

---

**Figure 5.13:** Schematic representation of the Neyman method.

In order to estimate the statistic and systematic uncertainties in a reliable approach an ensemble study has been performed. A graphic representation is shown in Figure 5.13. Pseudo experiments are generated by multiplying the W + jets template histograms with a
value $A$. The simulated templates are then stacked and the values for each bin are Poisson distributed. The result is a pseudo data histogram, to which the original templates are then fitted according to the method presented in section 5.5. This process is repeated 4000 times and the fit results are then put into a histogram, which is fitted by a Gaussian. To estimate the additional uncertainty due to the finite number of simulated events, the values for each bin are distributed according to a Gaussian, before being Poisson distributed. A sample Gaussian distribution is shown for the Poisson as well as for the convoluted distribution in Figure 5.14. As expected the gaussian for the convoluted distribution is broader than the other.

![Figure 5.14: Sample Gaussian distribution for $A = 1$.](image)

The results of the ensemble study are investigated by a Neyman approach. The results of the Gaussian fit from the ensemble study are plotted as a graph, with each mean of a Gaussian as a point and the $\sigma$ as its error. The graph allows to interpret the result for the ratio between the $W + \text{jets}$ and $t\overline{t}$ cross section from section 5.5.
5.6. **Estimation of Uncertainties: Ensemble Study**

The outcome as seen in Figure 5.15 is:

\[
\frac{\sigma_{W+jets}}{\sigma_t} = 235^{+17}_{-16} \text{ (stat.)}^{+39}_{-36} \text{ (stat.+sys.)}
\]

It is compatible with the simulated value:

\[
\frac{\sigma_{W+jets}}{\sigma_t} = 199 \pm 67
\]

The CMS experiment estimated the cross section to be [36] [37]:

\[
\sigma_{W+jets} = 30.93 \pm 0.09 \text{(stat.)} \pm 0.27 \text{(sys.)} \pm 0.30 \text{(th.)} \pm 1.23 \text{(lumi.) nb}
\]

\[
\sigma_t = 154 \pm 17 \text{ (stat. + sys.)} \pm 6 \text{ (lumi.) pb}
\]

The ratio is:

\[
\frac{\sigma_{W+jets}}{\sigma_t} = 201 \pm 21 \text{(stat. + sys.)}
\]

The result from this study is in agreement with the values measured by CMS and with the theoretical predictions from the standard model.
CHAPTER 5. DATA-DRIVEN BACKGROUND ESTIMATION
Chapter 6

Summary and Outlook

This study presents a method for data-driven background estimation for top-pair events at the CMS experiment. The semi-muonic decay channel is used at a center of mass energy of \( \sqrt{s} = 7 \) TeV. The data investigated corresponds to a luminosity of \( L = 4.6 \, \text{fb}^{-1} \).

The signature of the semi-muonic decay consists of a high-energy isolated muon, a neutrino causing missing transverse energy and four high-pT-jets. The relevant background processes which can produce a similar result in the detector are: W + jets, QCD-multijet, Z + jets and single-top events.

For the background estimation, simulated samples are used for t\( \bar{t} \) events and the individual backgrounds. The samples are official CMS samples. Correspondingly, the selection is a CMS reference selection, but no b tag is applied in order to enrich the background. After the selection, an excess of simulated events is clearly visible. This overestimation is not further investigated in this study, but it can be explained by the high uncertainties of the simulated cross sections.

The background estimation is implemented as a simultaneous fit for the two discriminating variables M3 and MET. M3 denotes the transverse invariant mass of the three jets with the highest transverse momentum and MET denotes the missing transverse energy.

We implement the simultaneous fit with ROOT and with the package RooFit. As a test the amount of background and t\( \bar{t} \) events in the data is fitted. The result from RooFit is used to check the result obtained with ROOT. The results for fitting the amount of t\( \bar{t} \) and background events are consistent, thus the result can be trusted. The uncertainties of the templates are not taken into account with this method.

For the final result, only the templates for t\( \bar{t} \) and W + jets events are fitted to data in order to estimate the ratio of the respective cross sections. The result is:

\[
\text{Measurement} : \frac{\sigma_{W+\text{jets}}}{\sigma_{t\bar{t}}} = 235^{+17}_{-16}(\text{stat.})^{+30}_{-36}(\text{stat. + sys.})
\]

The uncertainties are derived by an ensemble study. The result is consistent with other results from CMS and the theoretical prediction.

\[
\text{Theory} : \frac{\sigma_{W+\text{jets}}}{\sigma_{t\bar{t}}} = 199 \pm 67
\]

\[
\text{CMS Measurement} : \frac{\sigma_{W+\text{jets}}}{\sigma_{t\bar{t}}} = 201 \pm 21(\text{stat. + sys.})
\]

A better result could be obtained by fitting each of the backgrounds as a separate template. The excess of simulated events should also be investigated in order to improve the result.
An additional improvement could be achieved by including a third discriminating variable in the fit. Further discrimination could enable the fit to be performed after the whole selection including the b-tag. Alternatively, events with only three jets could be included in the selection to increase the amount of background events and the overall statistics, but the discrimination power of the variable M3 would have to be investigated again. Nevertheless, we are confident that the results validate the method used, which will hopefully be used in further analyses.
Chapter 7

Appendix

Additional Plots

Figure 7.1: Simultaneous fit of the W + jet, \( \bar{t} \bar{t} \) and the remaining background templates with ROOT.

Summary of the Files for Data and Simulation
Table 7.1: Datasets used for the analysis

<table>
<thead>
<tr>
<th>Filename</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simulated Samples:</strong></td>
</tr>
<tr>
<td>/TTJets_TuneZ2_7TeV-madgraph-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/DYJetsToLL_TuneZ2_M-50_7TeV-madgraph-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/QCD_Pt-20_MuEnrichedPt-15_TuneZ2_7TeV-pythia6/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/WJetsToNu_TuneZ2_7TeV-madgraph-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/T_TuneZ2_s-channel_7_TeV-powheg-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/T_TuneZ2_t-channel_7_TeV-powheg-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/T_TuneZ2_tW-channel-DR_7TeV-powheg-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/Tbar_TuneZ2_s-channel_7_TeV-powheg-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/Tbar_TuneZ2_t-channel_7_TeV-powheg-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td>/Tbar_TuneZ2_tW-channel_7_TeV-powheg-tauola/Summer11-PU_S4_START42_V11-v1/AODSIM</td>
</tr>
<tr>
<td><strong>Data Samples:</strong></td>
</tr>
<tr>
<td>/SingleMu/Run2011A-PromptReco-v6/AOD</td>
</tr>
<tr>
<td>/SingleMu/Run2011A-May10ReReco-v1/AOD</td>
</tr>
<tr>
<td>/SingleMu/Run2011A-PromptReco-v4/AOD</td>
</tr>
<tr>
<td>/SingleMu/Run2011B-PromptReco-v1/AOD</td>
</tr>
</tbody>
</table>
Table 7.2: json and pileup-reweighting files

<table>
<thead>
<tr>
<th>Filename</th>
<th>json-Files:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cert_160404-163869_7TeV_May10ReReco_Collisions11_JSON_v3.txt</td>
</tr>
<tr>
<td></td>
<td>Cert_160404-180252_7TeV_PromptReco_Collisions11_JSON.txt</td>
</tr>
<tr>
<td></td>
<td>Cert_170249-172619_7TeV_ReReco5Aug_Collisions11_JSON_v3.txt</td>
</tr>
<tr>
<td>Pileup-Reweighting Files 3D:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cert_160404-163869_7TeV_MayReReco_Collisions11_JSON_v3_pileupTruth_v2_finebin.root</td>
</tr>
<tr>
<td></td>
<td>Cert_165088-167913_7TeV_PromptReco_JSON.pileupTruth_v2_finebin.root</td>
</tr>
<tr>
<td></td>
<td>Cert_170249-172619_7TeV_ReReco5Aug_Collisions11_JSON_v2_pileupTruth_v2_finebin.root</td>
</tr>
<tr>
<td></td>
<td>Cert_172620-173692_PromptReco_JSON.pileupTruth_v2_finebin.root</td>
</tr>
<tr>
<td></td>
<td>Cert_175832-177515_PromptReco_JSON.pileupTruth_v2_finebin.root</td>
</tr>
<tr>
<td></td>
<td>Cert_177718-178078_7TeV_PromptReco_Collisions11_JSON.pileupTruth_v2_finebin.root</td>
</tr>
<tr>
<td></td>
<td>Cert_178098-180252_7TeV_PromptReco_Collisions11_JSON.pileupTruth_v2_finebin.root</td>
</tr>
<tr>
<td>Pileup-Reweighting Files 1BX:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cert_160404-163869_7TeV_MayReReco_Collisions11_JSON_v3_pileup_v2.root</td>
</tr>
<tr>
<td></td>
<td>Cert_175832-177515_PromptReco_JSON.pileup_v2.root</td>
</tr>
<tr>
<td></td>
<td>Cert_165088-167913_7TeV_PromptReco_JSON.pileup_v2.root</td>
</tr>
<tr>
<td></td>
<td>Cert_177718-178078_7TeV_PromptReco_JSON.pileup_v2.root</td>
</tr>
<tr>
<td></td>
<td>Cert_170249-172619_7TeV_ReReco5Aug_Collisions11_JSON.pileup_v2.root</td>
</tr>
<tr>
<td></td>
<td>Cert_178098-180252_7TeV_PromptReco_Collisions11_JSON.pileup_v2.root</td>
</tr>
<tr>
<td></td>
<td>Cert_172620-173692_PromptReco_JSON.pileup_v2.root</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to thank the numerous people who helped me complete this study. My thanks first of all to Prof. Dr. Achim Stahl for enabling me to write this interesting bachelor thesis.

I owe a great debt to Dr. Heiko Geenen, Yvonne Küssel, Heiner Tholen, Felix Höhle and Dr. Olliver Pooth for their abundance of patience in answering all our numerous questions.

The nice atmosphere in the top group and in the whole institute 3B made the task less strenuous.

My thanks go to Steffen Lindemann and Marten Weise for proofreading this thesis. Florian Lenz has been a tremendous co-worker during the last three months, for that I am very grateful.

At last I would like to thank my family for always supporting me.
Bibliography


[30] Tune Z2 is identical to the tune Z1 described in [31] except for the PDF. Z2 uses CTEQ6L [26] while Z1 uses CTEQ5L.


