Boosted top quarks in the ttbar dilepton channel: optimization of the lepton selection

DESY Summer School 2014

9 September, 2014

Author:
Ibles Olcina Samblàs*

Supervisor:
Carmen Diez Pardos

Abstract

A study of boosted top quark topologies in $t\bar{t}$ events in the dilepton decay channels was performed with the data recorded in pp collisions at $\sqrt{s} = 8$ TeV by the CMS detector during the LHC 2012 run period, which corresponds to an integrated luminosity of 19.7 fb^{-1}. The kinematics of the decay products of boosted top quarks were studied and the criteria to select the leptons was optimized. The optimized selection leads to a better agreement between Monte Carlo and data events, higher efficiency and a higher background rejection.

*Universitat de València, Spain

Contents

1 Introduction

1.1 Motivation

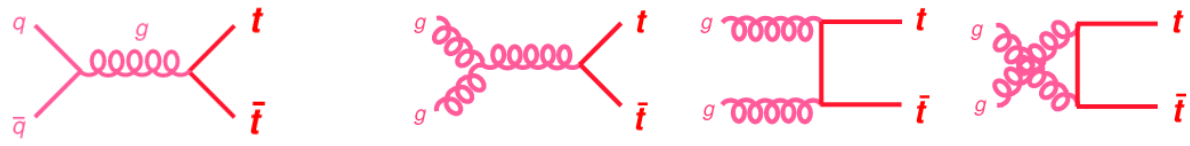
The top quark was discovered by the Tevatron experiments CDF and DØ in 1995, about 20 years after it had been predicted. It is the heaviest elementary particle, with a mass of $m_t = 173.34 \pm 0.27$ $(stat.) \pm 0.71$ $(syst.)$ GeV [1], which is of the same order of magnitude as the mass of a gold atom. Because of its huge mass, it is a good candidate for studies of the fermionic coupling to the Higgs boson. In addition, it has a lifetime which is considerably shorter than the average time for hadronization, which causes that it always decays to other fundamental particles before forming bound states. As a result, the top quark constitutes an important fundamental particle for precision measurements of the Standard Model (SM) as well as for searches of new physics beyond the SM. At the moment, the LHC can be regarded as a "top factory" since the large achieved luminosities results in large statistics of $t\bar{t}$ events.

The aim of this report is to outline the data analysis that was conducted within the top quark physics group of CMS-DESY as part of the DESY summerschool 2014. The main goal of the analysis was to study top quarks with very high momentum produced in pp collisions at the LHC and optimise the selection of the leptons produced in their decays. As part of the work, isolation efficiencies as well as trigger efficiencies were studied (each term is explained later in the text) and an improved agreement between Monte Carlo events and data events was found. The data set used for this analysis is the one recorded in 2012 by the CMS detector at the LHC, which corresponds to pp collisions at $\sqrt{s} = 8$ TeV and an integrated luminosity of 19.7 ± 0.9 fb^{-1}.

1.2 ttbar production and decay products

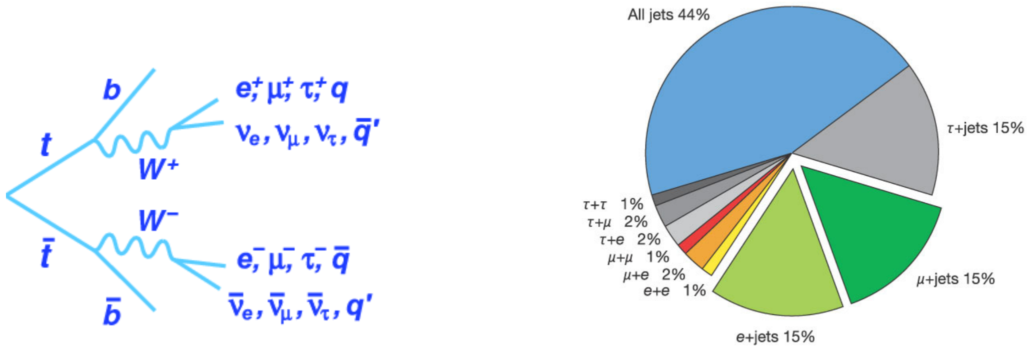
$t\bar{t}$ pairs of quarks are mainly produced through QCD processes. Such a pair of quarks can be produced from either a pair of gluons, Figure 1a, or a pair of quark and anti-quark, Figure 1b, being the first process the dominant one at the LHC. Once the top quark is produced, it decays in almost 100% of the cases to a W^+ boson and a b quark (or the opposite charges for the $\bar{t}$ quark). According to the decay products of the W boson, the $t\bar{t}$ events can be divided into three different categories (Figure 2):

- **Fully hadronic channel:** Both W bosons decay into a pair of quark and anti-quark of different flavour. As a result, 6 quarks are produced in total. This channel has the largest branching fraction, about 44%, but also has a very high background coming from QCD processes.
- **Semileptonic channel:** One W boson decays to a pair of quark and anti-quark and the other to one charged lepton and a neutrino of the same flavour. As a result, 4 quarks, 1 charged lepton and 1 neutrino are produced. The branching ratio of this channel is about 30% and it has a moderate background mainly coming from W production with additional jets (referred to as "W+jets"). Given the very short lifetime of the tau lepton, only electrons and muons are considered as possible lepton candidates.
- **Dilepton channel:** Both W bosons decay into 1 charged lepton and 1 neutrino. As a result, 2 quarks, 2 charged leptons and 2 neutrinos are produced. This channel has the lowest branching ratio, about 5% but also has a very low background coming mainly from $Z/\gamma^* - boson$ production with additional jets (referred to as "Z+jets") and QCD processes.



(a) (b)

Figure 1: Leading order Feynman diagrams for strong production of top quark pairs from (a) quarks and (b) gluons. *Source:* DØ Collaboration.



(a) (b)

Figure 2: (a) Leading order Feynman diagram for the $t\bar{t}$ decay. (b) Top pair decay branching fractions. *Source:* DØ Collaboration.

1.3 Event selection

The dilepton channel was used for the analysis, which follows the same event selection and strategy as [3]. As a result, the relevant events for this analysis must contain 2 isolated, opposite-charged leptons with high transverse momentum, since they are formed from a W boson, which is a very massive particle, 2 energetic jets and large missing transverse energy, defined as $\not{E}_T = -\sum_{leptons} p_T - \sum_{jets} p_T$, due to the neutrinos. Specifically, the requirements for the event selection are the following:

- **Trigger:** Events must pass first a trigger based on dileptons.
- **Leptons:** Electron candidates are required to have a transverse energy $E_T > 20$ GeV and muon candidates to have a transverse momentum $p_T > 20$ GeV, within a pseudorapidity $|\eta| < 2.4$. Lepton candidates are also required to fulfil an isolation criterion, which will be explained in Section 3.
  It is also required that $m^{ll} > 20$ GeV, where m^{ll} is the invariant mass of the selected lepton pair. Furthermore, most of the Z+jet background is removed for the $\mu^+\mu^-$ and e^+e^- channels by rejecting events in the Z^0 mass region, which means rejecting events in the energy region 76 GeV $< m^{ll} <$ 106 GeV.
- **b-jets:** It is required to have at least two jets with $p_T > 30$ GeV within a pseudorapidity $|\eta| < 2.4$. A b-tagging algorithm is applied later to infer which of the jets came from a b quark. This is done using

the property that B hadrons have a large decay length and therefore jets originated from a b quark tend to show displaced tracks or many secondary vertices. After updating the status of the selected jets, only one b-tagged jet is required to avoid losing too many statistics.

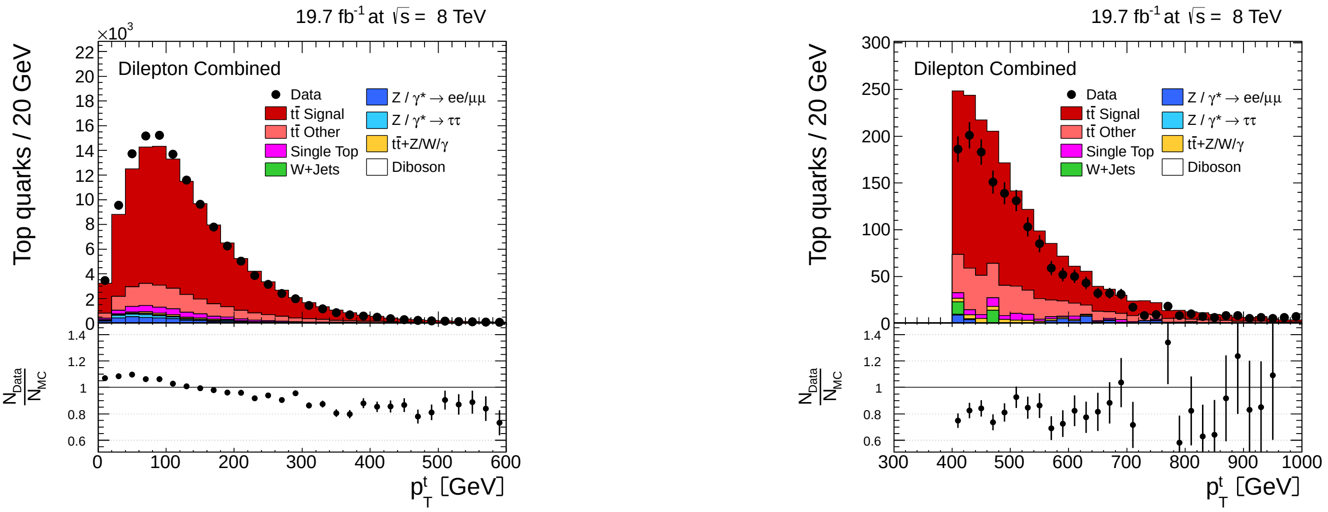
- **Missing transverse energy:** It is required that $\not{E}_T > 40$ GeV in the dimuon and dielectron channels to reject a large fraction of background from "Z+jets" processes.
- **Kinematic reconstruction:** The kinematics of the full $t\bar{t}$ event can be totally reconstructed by solving the kinematic equations, which are required to have at least one solution. The system of equations is constructed using the four momenta of the 2 leptons, of the 2 jets and the missing transverse energy. Then, the following assumptions are made to avoid having an unconstrained system of equations:
  - The missing transverse energy is entirely caused by the neutrinos.
  - The lepton and the neutrino from the same top branch have an invariant mass equal to the W mass, 84.4 GeV.
  - The reconstructed top and anti-top quark masses are fixed to a nominal value of 172.5 GeV.

  After applying these constrains, the system of equations can have up to 4 solutions. The ones with more b-tagged jets and with the best reconstructed neutrino energy with respect to Monte Carlo generated spectrum are preferred.

2 Kinematics of boosted topologies

Figure 3a shows the distribution of the transverse momentum of the top quarks, which is obtained from the fully kinematic reconstruction of the $t\bar{t}$ system. To study the boosted regime, only top quarks with a momentum higher than 400 GeV are selected ("boosted tops" in the following), as can be seen in Figure 3b. Table 2 shows the comparison of the number of events and the signal to background ratio of the boosted and the not boosted regime. There is an excess of MC events over data events for the boosted regime, which implies a worse agreement between them. Also, the ratio of signal to background decreases. The main goal of this project was to improve the signal selection efficiency and the signal to background ratio for boosted topologies by optimizing the event selection.

First, it was studied how the kinematic distributions of the decay products of the top quarks were modified in the boosted regime. In Figure 4, the distribution of the angular distance between a lepton and the corresponding b-jet, defined as $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, is represented for the full phase space and the boosted region. In these two plots it can be observed that the distribution is shifted towards smaller angular values for the boosted region. This is expected, since the decay products of a top quark will be closer to each other if the top quark has a high momentum.
Figure 5 shows the distributions of the transverse momentum of the leptons and the b-jets as well as the missing transverse energy for the full phase space and the boosted region. In order to compare each pair of distributions better, the ratio between the $t\bar{t}$ signal in the boosted regime and the full phase space can be seen in the third column of Figure 5. It can be observed that a large fraction of leptons with high momentum are produced from boosted tops while this fraction is much lower for b-jets. Also, all the distributions in the boosted regime show an excess of MC events, as already observed in Table 2.



(a) (b)

Figure 3: Transverse momentum distribution of the top quarks for the full phase space (left) and the boosted regime (right). Both plots correspond to the combined results of the three dilepton channels ee, $e\mu$ and $\mu\mu$. $t\bar{t}$ events from the MC simulation are represented in red while all the other contributions are part of the background.

	All data set		Boosted regime	
	Events	Ratio over MC (%)	Events	Ratio over MC (%)
Total MC	64757		1023	
Total Data	64819	100.1	841	82.2
Total Background	14620	22.6	301	29.4
$t\bar{t}$ signal	50137	77.4	722	70.6

Table 1: Total number of MC and data events as well as the fractions of $t\bar{t}$ signal and background for the full phase space and the boosted regime for the combined dilepton channel. In the boosted regime, there is an excess of MC events and the proportion of background events is larger than for the full phase space.

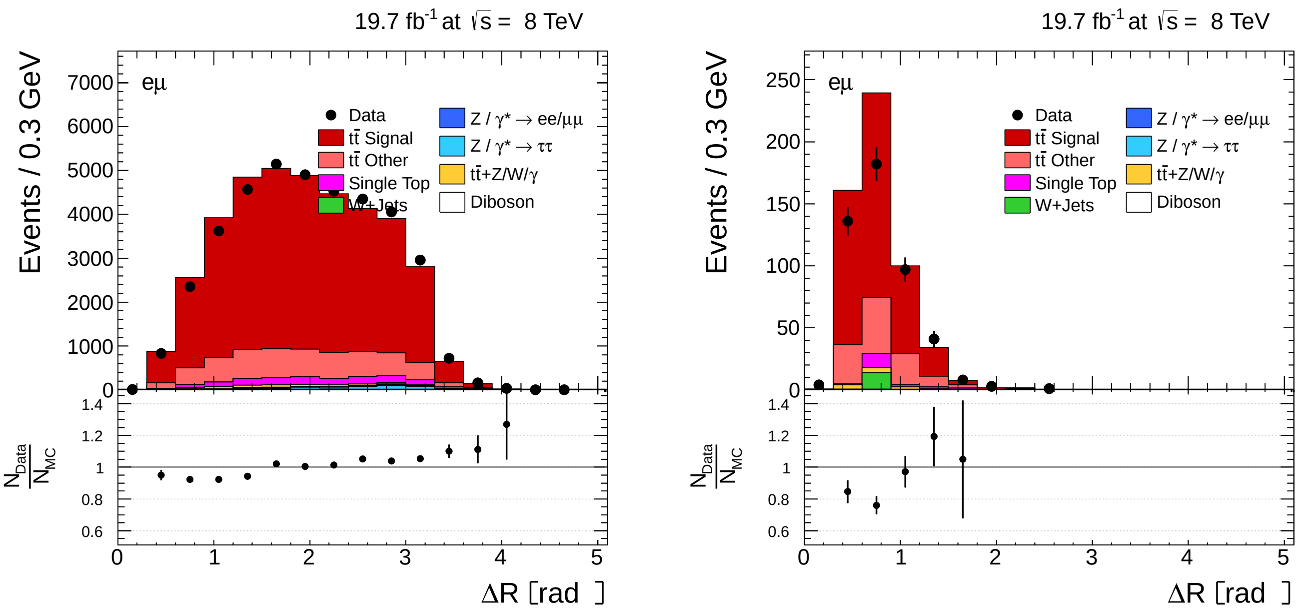


Figure 4: Distributions of ΔR between a lepton and the corresponding anti-b jet for the full phase space (left) and the boosted regime (right) for the $e\mu$ channel.

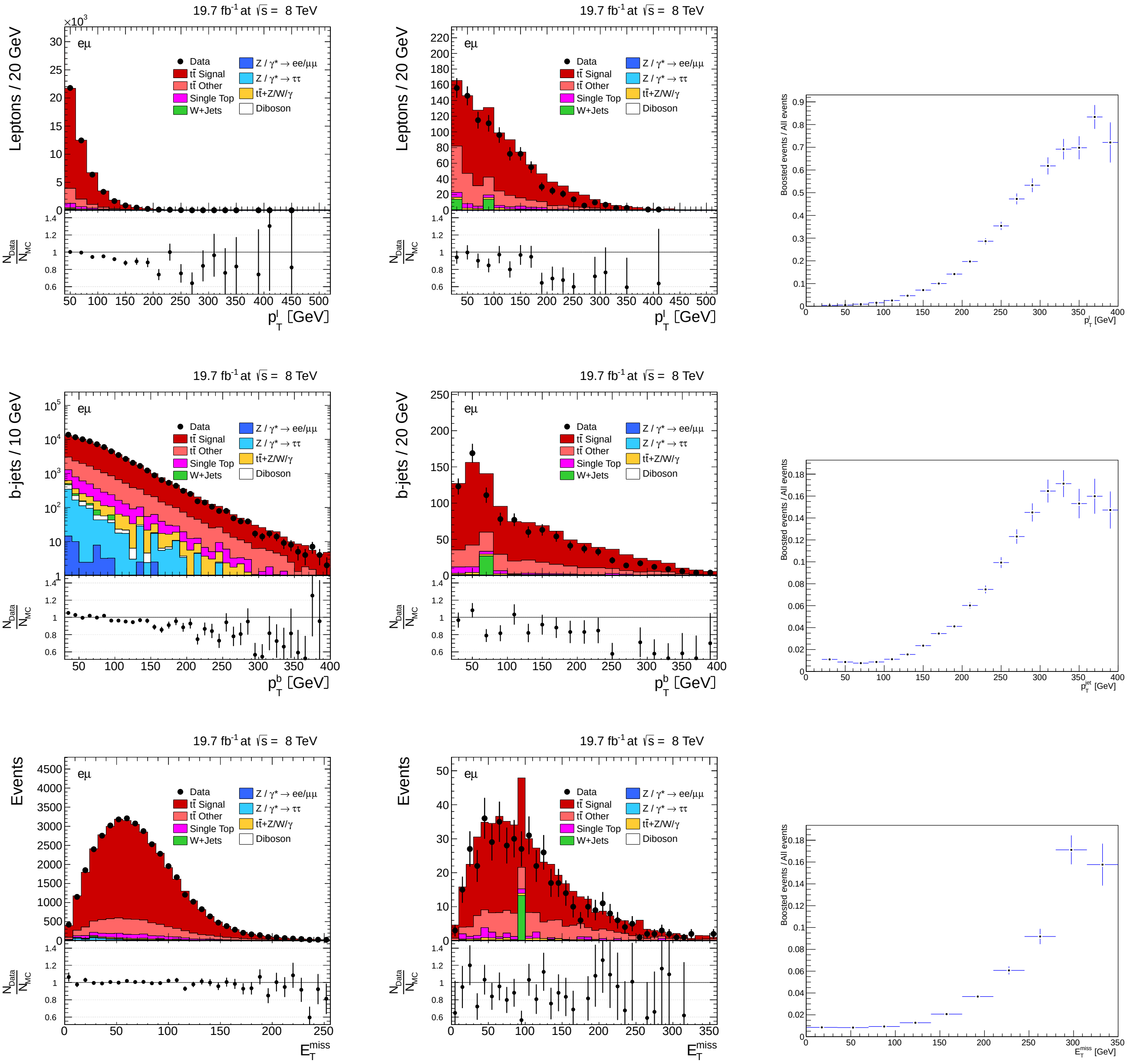


Figure 5: Distributions of some representative variables of the top decay products for all the events (left column) and for the boosted regime (central column) for the $e\mu$ channel. The ratio of boosted ttbar signal events over the total of $t\bar{t}$ signal events is shown in the right column. The first and second rows correspond to the transverse momentum of the leptons and the b-jets respectively while the third row is the missing transverse energy.

3 Optimization of lepton selection for boosted topologies

In the previous section it was seen that the agreement between MC and data events was worse for boosted topologies, as well as that the proportion of background events was larger. In this work, these statistics were improved by optimizing the selection of leptons in boosted topologies. Two optimizations were devised:

1. The first row of Figure 5 shows that almost half of the leptons in the boosted regime with $p_T < 40$ GeV comes from background events. Therefore, the cut on the transverse momentum of the leptons could be changed from 20 to 40 GeV to reject a considerable amount of background. Furthermore, from the third plot of the first row it is observed that a very low fraction of boosted leptons would be lost because of this new cut.

2. It was concluded from Figure 4 that the decay products of the top quarks are closer to each other in the boosted regime. For this reason, the isolation criterion for boosted topologies for the leptons should be revised.

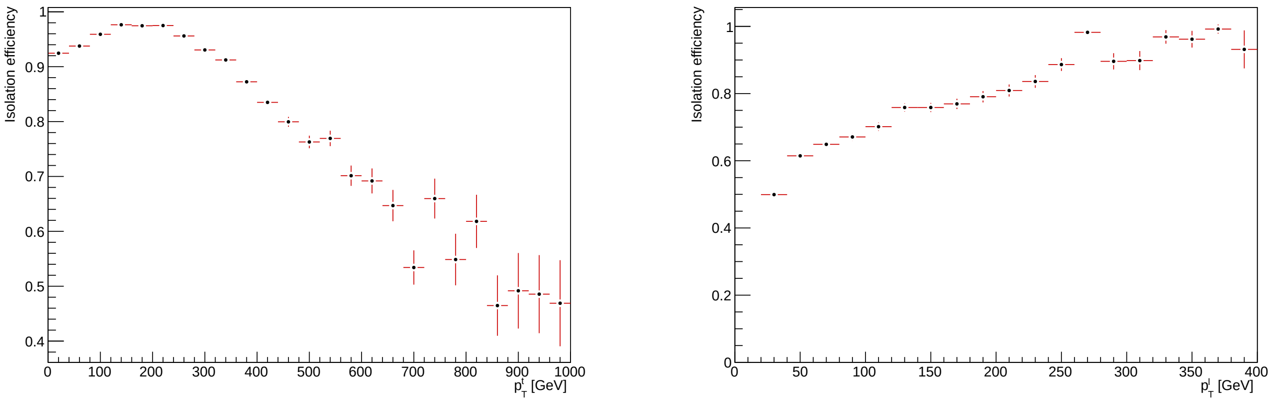
3.1 Standard isolation criterion

The standard isolation criterion used for the leptons is the following: *accept event if* $I^{pf}_{Rel} < 0.15$ *inside a cone in* $\eta - \phi$ *space of* $\Delta R < 0.3$ *around the lepton, where* I^{pf}_{Rel} *is defined as:*

$$I^l_{rel} = \sum_{i=had,\gamma} E^i_T / p^l_T \tag{1}$$

This means that the sum of transverse energy deposits from charged and neutral hadrons and photons relative to the transverse momentum of the lepton inside a certain cone around the lepton is small, so most of the energy inside that cone comes from the lepton.

Figures 6a and 6b show the isolation efficiencies as a function of the transverse momentum of the top quarks and leptons, respectively, where the isolation efficiency is defined as the number of events that were required to fulfil an isolation condition by the total number of events. Figure 6a shows that the isolation efficiency is reasonably good for low momentum values of the top quarks but it decreases to values of almost 50% for the highest momentum values. The opposite happens for the momentum of the leptons, as can be seen in Figure 6b, where the worst efficiencies are obtained for the lowest values of the momentum. The decreased isolation efficiencies obtained with this isolation criterion motivated the search for another isolation criterion.



(a) **(b)**

Figure 6: Isolation efficiency as a function of the transverse momentum of the top quarks and the leptons. The full data set was considered for the first plot while the boosted regime was used for the second one.

3.2 2D cut

A new isolation criterion was studied, which will be referred as "2D cut", states: *accept the event is* $\Delta R(l, closest\ jet) > 0.5$ ***or*** $|p_T^{rel}(l, closest\ jet)| >$ *25 GeV, where* $p_T^{rel}(l, closest\ jet)$ *is the magnitude of the transverse component of the momentum of the lepton relative to the momentum of its closest jet.*

Consequently, events will be rejected if they do not satisfy both conditions at the same time. This is shown in Figure 7, where there are no events in the region of $\Delta R(l, closest\ jet) < 0.5$ and $|p_T^{rel}(l, closest\ jet)| < 25$ GeV. The comparison of the isolation efficiency between the two isolation conditions can be seen in Figure 8 as a function of the transverse momentum of the top quarks, the leptons, the b-jets and the missing transverse energy. The efficiencies obtained with the 2D cut shows a clear improvement in the boosted region. Table 2 summarises the number of events lost and the efficiencies obtained with the full data set and the boosted regime. The isolation efficiencies are similar between the two isolation conditions for the full data set, whereas the isolation efficiency of the 2D cut condition is 87.7%, about 15% higher that the standard criterion for the boosted regime.

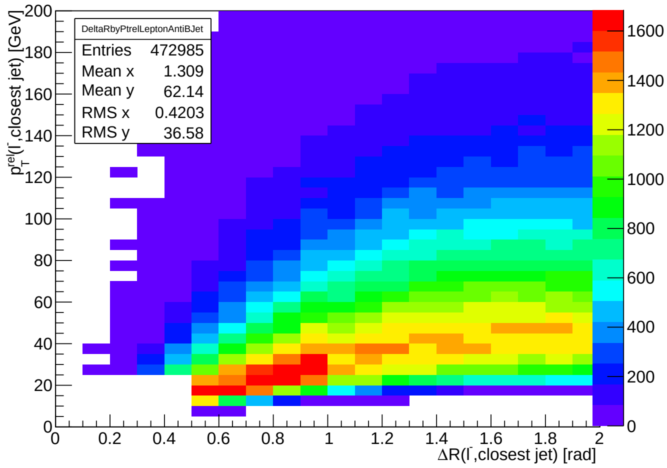


Figure 7: 2D histogram to illustrate the working principle of the 2D cut condition. The $x-$axis is the ΔR between a lepton and its closest jet and the $y-$axis is the magnitude of the transverse component of the lepton relative to the momentum of its closest jet.

	No Isolation Cond.	Regular Isolation Cond.	2D cut
All data sample			
Number of events	469277	444696	442814
Loss of events		24581	26463
Efficiency (%)		94.8	94.4
Boosted regime			
Number of events	7651	5595	6708
Loss of events		2056	943
Efficiency (%)		73.1	87.7

Table 2: Comparison of the total isolation efficiency between two different isolation criteria. The new criterion has a 15% higher efficiency for the boosted regime.

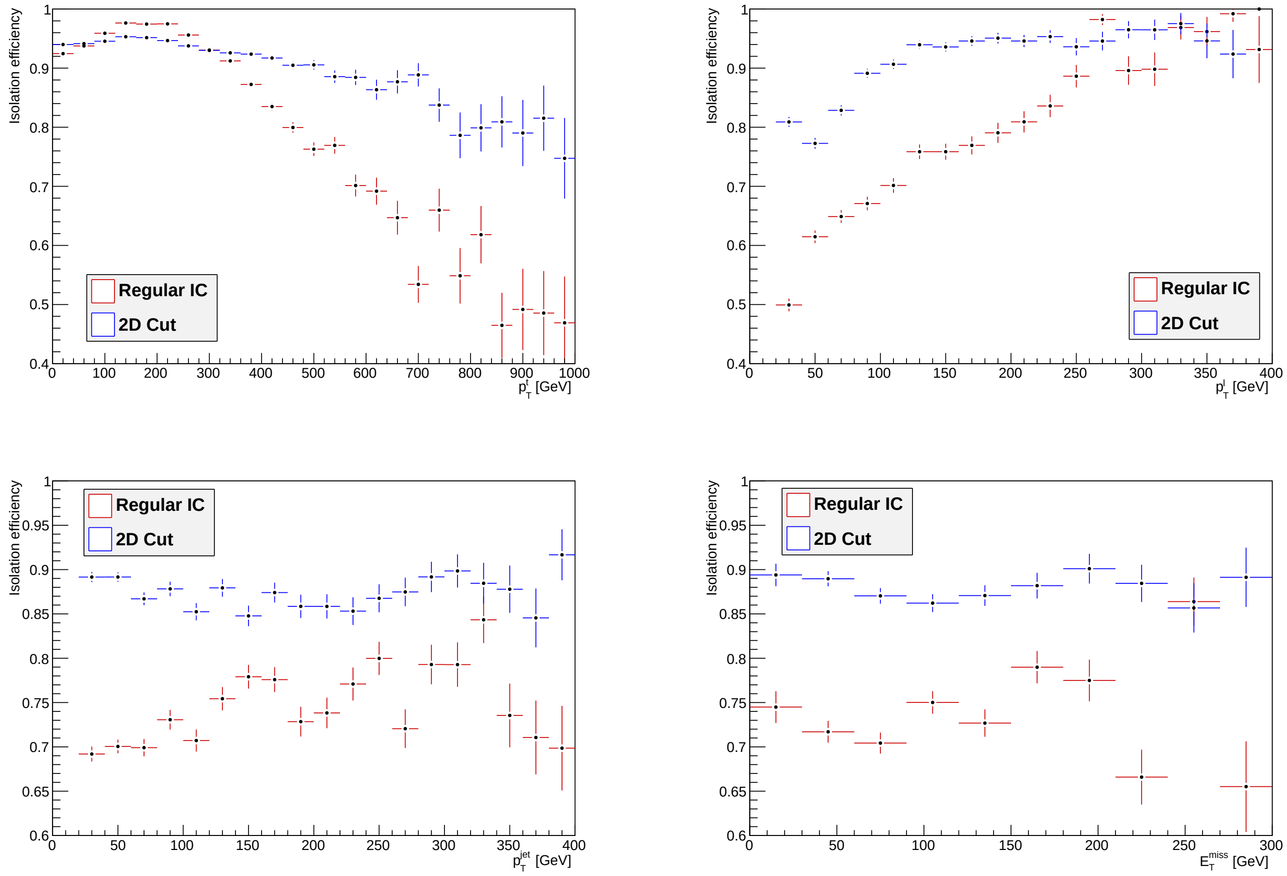


Figure 8: Comparison of isolation efficiencies for the two different isolation criteria as a function of the transverse momentum of the top quarks (top left), the leptons (top right), the b-jets (bottom left) and the missing transverse energy (bottom right). The first plot is for the full data sample while the other three are for the boosted regime.

Better isolation efficiencies are obtained with the new isolation condition. Nevertheless, given the importance of the trigger in the selection of the leptons, trigger efficiencies with this new isolation condition were also checked and compared to the ones obtained with the regular isolation condition. These efficiencies are shown in Figure 9 as a function of the transverse momentum of the top quarks, the leptons, the b-jets and the missing transverse energy. In general, the trigger efficiencies with the 2D cut are about 4% lower, which is a much smaller decrease compared to the gain in isolation efficiency.

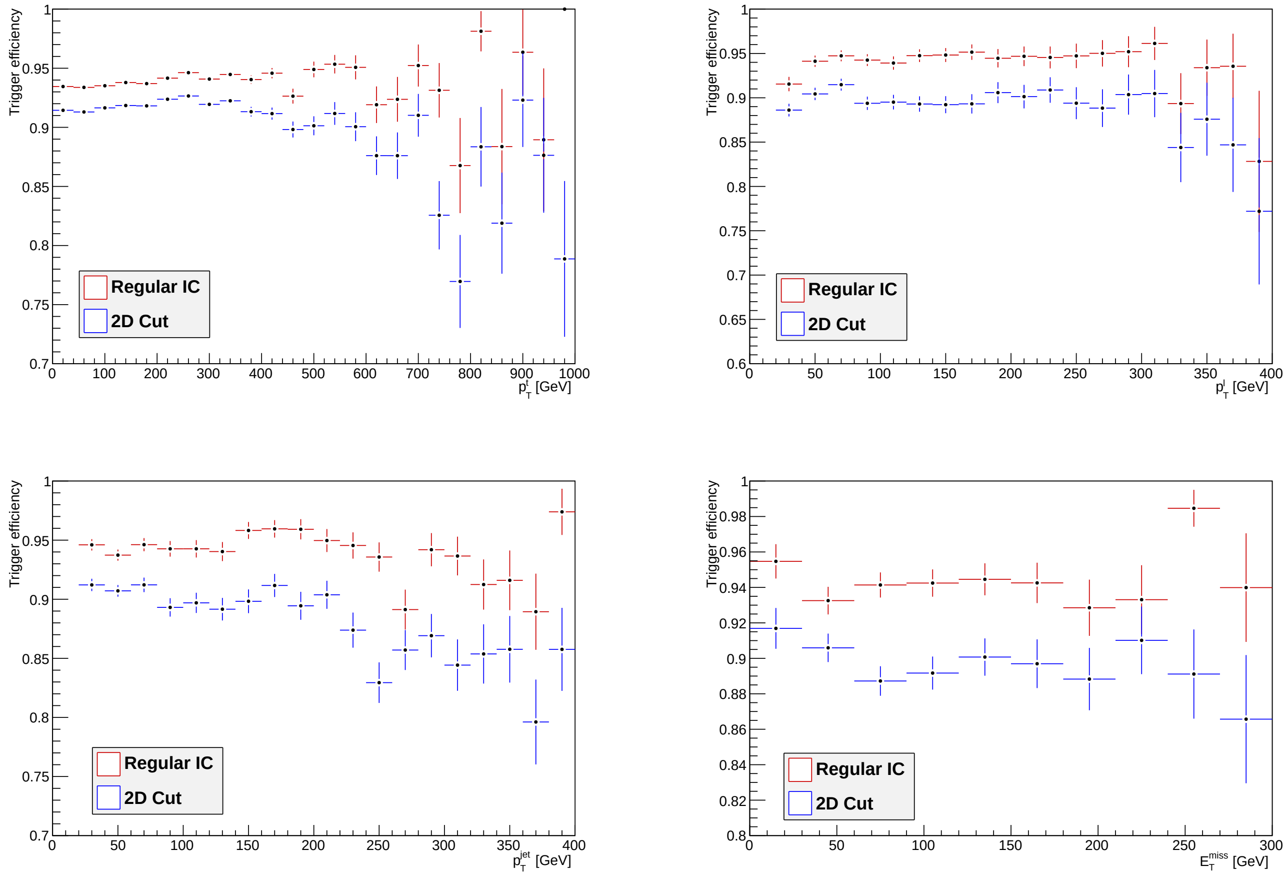


Figure 9: Comparison of trigger efficiencies for the two different isolation criteria as a function of the transverse momentum of the top quarks (top left), the leptons (top right), the b-jets (bottom left) and the missing transverse energy (bottom right). The first plot is for the full data set while the other three are for the boosted regime.

4 Results with the optimized selection

In this section, the new distributions using the two optimizations explained in the previous section are shown. Figure 10 shows the transverse momentum distribution of the top quarks for the full data set and the boosted regime and in both cases a good agreement between MC and data events is observed. Table 3 summarises the changes in the agreement between MC and data events as well as the proportions of background and $t\bar{t}$ signal for each of the steps in the optimization process followed in this analysis. With the new isolation condition, the efficiency is higher the agreement between MC and data events improves slightly and it does not affect the fraction of background over signal obtained for the boosted regime. Changing also the cut in the transverse momentum of the leptons improves the agreement between MC and the proportion of background is much reduced. As a result, the optimized selection have similar proportions of background and $t\bar{t}$ signal for the boosted regime to the ones of the original full data set, clearly better than the default selection.
Figure 11 shows the fully optimized distributions of ΔR between a lepton and the corresponding b-jet, the transverse momentum of the leptons and the b-jets as well as the missing transverse energy for the boosted region in the $e\mu$ dilepton channel. All the distributions show a small excess of MC events over data events, as already seen in Table 3, but in general there is a clear improvement between them.

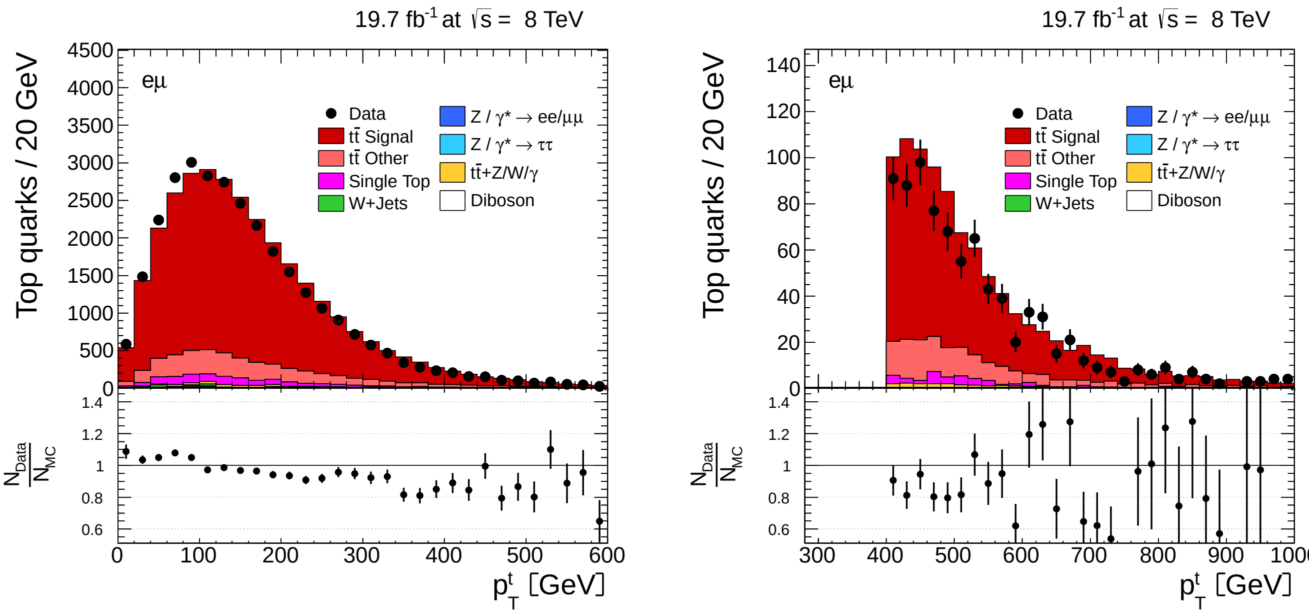


Figure 10: Transverse momentum distribution of the top quarks for the full data set (left) and the boosted sample (right) for the $e\mu$ channel. $t\bar{t}$ events from the MC simulation are represented in red while all the other contributions are part of the background.

Regular IC	**All data set**		**Boosted regime**	
	Events	Ratio over MC (%)	Events	Ratio over MC (%)
Total MC	38396		548	
Total Data	38239	99.6	472	86.2
Total Background	7708	20.1	157	28.7
$t\bar{t}$ signal	30688	79.9	391	71.3

2D cut	**Boosted regime (20 GeV cut)**		**Boosted regime (40 GeV cut)**	
	Events	Ratio over MC (%)	Events	Ratio over MC (%)
Total MC	672		482	
Total Data	742	110.4	434	90.0
Total Background	204	30.3	106	22.1
$t\bar{t}$ signal	468	69.7	376	77.9

Table 3: Total number of MC and data events as well as the fractions of $t\bar{t}$ signal and background for the regular isolation criterion (first table) and the 2D cut criterion (second table) for the $e\mu$ channel. With the new isolation condition and the change of the cut to the transverse momentum of the leptons from 20 to 40, a similar proportions of background and $t\bar{t}$ signal to the ones of the original full data set are obtained for the boosted regime (upper table, first column and lower table, second column).

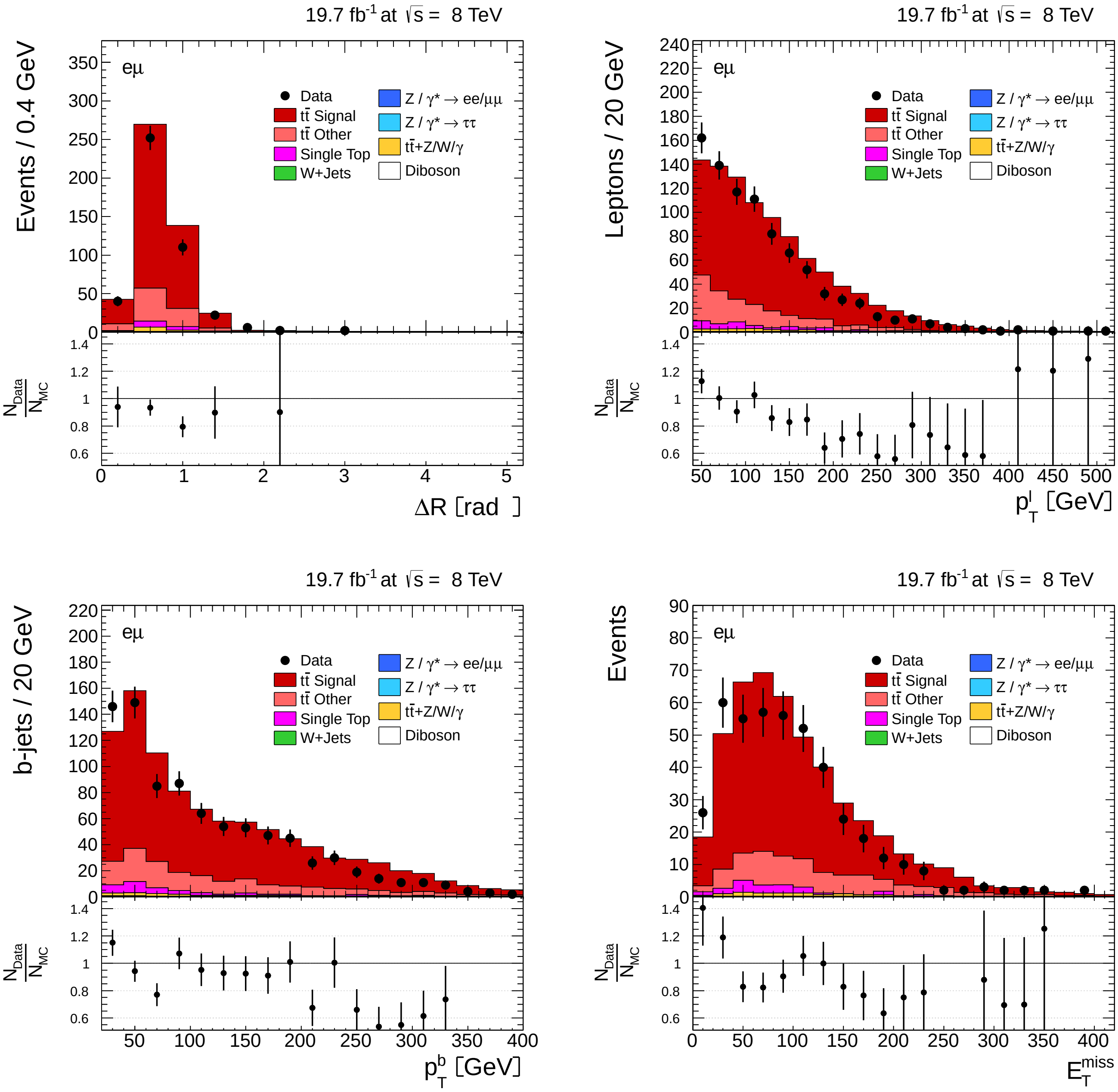


Figure 11: Distributions of some representative variables of the top decay products for the boosted regime and for the $e\mu$ channel. From left to right and from top to bottom: ΔR between a lepton and the corresponding b-jet, transverse momentum of the leptons, b-jets and transverse missing energy.

5 Conclusions

A study of top quarks with very high momentum was performed in the $t\bar{t}$ dilepton decay channels. The consequences of having boosted tops are that the distributions of the angular distance between a lepton and the corresponding b-jet were shifter towards smaller values, as well as the transverse momentum of the decay products of the top quark took higher values. Applying the standard criteria to select the events, an excess of MC events over data events for the boosted regions and a higher proportion of background events was found. In order to improve the selection, two optimizations were studied. First, the cut on the transverse momentum of the leptons was changed from 20 to 40 GeV to reject a larger fraction of background without losing too many signal events in the boosted regime. Second, an alternative isolation criterion was considered, which showed an improvement of approximately 15% in the isolation efficiency. The trigger efficiencies with the new isolation condition were about 4% worse than the standard isolation condition but this is a much smaller decrease compared to the gain in isolation efficiency. With these improvements, there was only an excess of the 10% of data over MC events. Furthermore, the proportion of background was reduced considerably and it was similar to the one that in the full phase space.

6 Acknowledgments

I would like to thank Carmen Diez Pardos, the supervisor of this project, for her guidance and help during this summer programme. I found very interesting the work done together and I hope to have the chance to collaborate with her again in the future. I am convinced that the experience gained here will help me in any future project that I undertake.
I am also grateful to Iván Asin, who kindly answered some of the many questions that I had the first weeks here and also gave me support and advice for my project. Finally, a special thanks goes to all the people that makes this summer programme possible. I think that this is an incredible opportunity for any student to extend his/her formation as a scientist.

References

[1] ATLAS, CDF, CMS and D0 Collaborations, *First combination of Tevatron and LHC measurements of the top-quark mass*, (2014). arXiv:1403.4427.

[2] CMS Collaboration, *Measurement of differential top-quark pair production cross sections in pp collisions at* $\sqrt{s} = 7$ *TeV*, EPJC 73 (2013) 2339. arXiv:1211.2220v3.

[3] CMS Collaboration, *Measurement of differential top-quark pair production cross sections in pp collisions at* $\sqrt{s} = 8$ *TeV*, (CMS PAS TOP-12-028).