Selection of $t\bar{t}$ events at the ATLAS experiment using multivariate analysis methods

Philippe d'Argent, Deutsches Elektronen-Synchrotron, Germany

September 11, 2014

Abstract

In this report the selection of top quark pair events in the decay modes $t\bar{t} \to (be^+\nu_e)(\bar{b}e^-\bar{\nu}_e)$ and $t\bar{t} \to (b\mu^+\nu_\mu)(\bar{b}\mu^-\bar{\nu}_\mu)$ is presented. It is investigated how the utilization of multivariate analysis techniques, instead of rectangular cuts, could lead to a more powerful discrimination between signal and background. By making use of a boosted decision tree in order to classify signal and background events, the signal efficiency is enhanced from $\epsilon_S = 49\%$ to $\epsilon_S = 80\%$ at a background efficiency of $\epsilon_B = 2.4\%$ for the ee channel and from $\epsilon_S = 49.5\%$ to $\epsilon_S = 79\%$ at a background efficiency of $\epsilon_B = 2.2\%$ for the $\mu\mu$ channel.

Contents

1 Introduction

The fundamental building blocks of matter are quarks and leptons. There are six known leptons in nature, the electron, muon, tau and their associated neutrinos. Among the six known quarks (up, down, charm, strange, top, bottom), the top quark is by far the heaviest. The top quark is essential to study the properties of the Higgs boson since its Yukawa coupling to the Higgs boson, which is proportional to the mass, is the largest. Furthermore, the top quark is the only quark that decays before it hadronizes, such that top decays provide an unique window on the bare quark. A precise understanding of the top quark properties is also crucial in order to probe various beyond the Standard Model extensions. New physics may give rise to additional top production or decay mechanisms. Additionally, $t\bar{t}$ production is a major source of background in almost all searches for physics beyond the Standard Model [1]. The aim of the analysis presented in this report is to improve the selection of $t\bar{t}$ events in the dilepton decay mode by making use of multivariate analysis techniques.

This report is organized in the following way: In Section 2 a brief overview of the ATLAS experiment is presented, followed by a discussion of the top quark production mechanisms and its decay modes. The cut based selection of $t\bar{t}$ candidates is summarized in Section 3. Section 4 adresses the utilization of multivariate analysis tools in order to improve the efficiency of the signal selection. Finally, Section 5 gives a conclusion and an outlook.

2 The ATLAS experiment

The ATLAS (A Toroidal LHC Apparatus [2]) experiment is one of the four major experiments at the Large Hadron Collider (LHC) at CERN located in Geneva. The Large Hadron Collider (LHC) is a proton-proton collider with a circumference of $27\,km$. In the following Section the setup of the ATLAS detector and the production mechanism of top quarks at the LHC are briefly discussed. In the end, the simulated data sets used in this analysis are introduced.

2.1 The ATLAS detector

The cylindrical geometry of the ATLAS detector provides an almost full solid angle coverage around the interaction point. The inner detector, which covers a pseudorapidity range $|\eta| < 2.5$ [1], consists of multiple layers of silicon pixel and microstrip detectors and a straw-tube transition radiation tracker (TRT). It is responsible for the reconstruction of charged tracks and vertices. A superconducting solenoid provides the inner detector with a 2 T axial magnetic field.

The solenoid is surrounded by high-granularity lead/liquid-argon electromagnetic (EM) sampling calorimeters and an iron/scintillating-tile hadronic calorimeter. With the aid of the calorimeter system electrons, photons and jets can be reconstructed.

The outermost part of the detector is dedicated to reconstruct muons. The muons spectrometer is instrumented with separate trigger and high-precision tracking chambers. An overview of the ATLAS detector and its subsystems is shown in Figure 1.

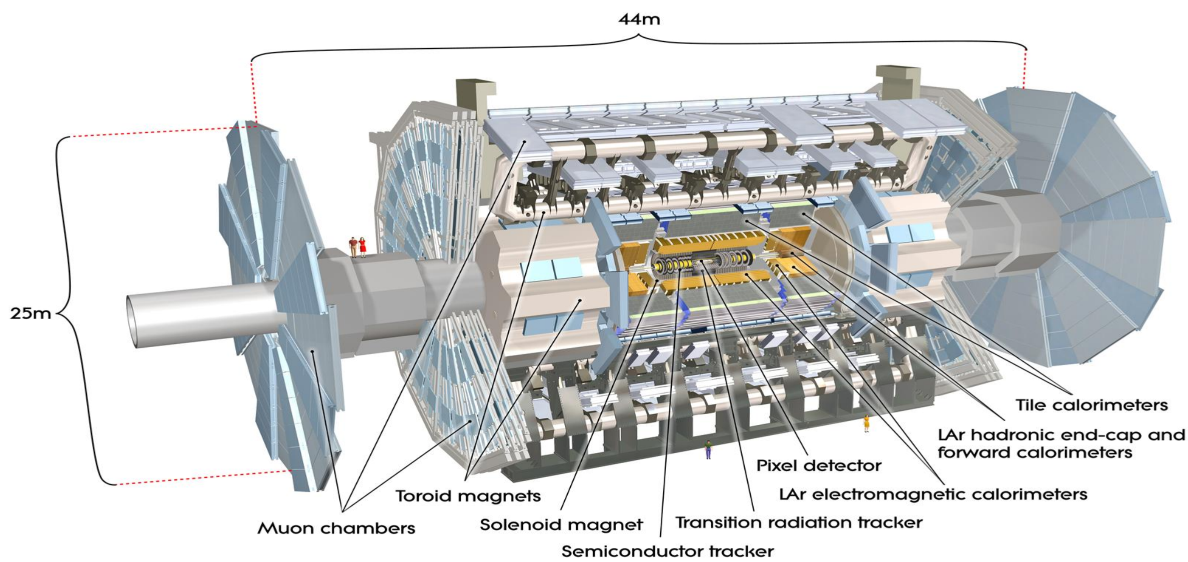

Figure 1: The ATLAS detector and subsystems [2].

[1] The pseudorapidity η is defined as $\eta = -ln[tan(\theta/2)]$, where the polar angle θ is measured with respect to the LHC beam-axis.

2.2 Top pair production and decay modes

The main production mechanism of top quarks at the LHC is the top-pair production through gluon-gluon fusion ($\approx 90\%$ at the design center of mass energy $\sqrt{s} = 14\,TeV$ [1]) followed by quark anti-quark annihilation ($\approx 10\%$ [1]). The corresponding Feynman diagrams are depicted in Figure [2]. Since the top quark decays almost exclusively into a bottom quark and a W boson ($Br(t \to bW) \approx 100\%$ [1]), the final state topology is labeled according to the W boson decay mode. The "alljet" decay mode has the largest branching fraction (see Figure [3]) but suffers from large multijet background due to the high event multiplicity at the LHC. In contrast, the "dilepton" decay modes leave cleaner signatures. In this analysis the two decay modes :

$$t\bar{t} \to (be^{+}\nu_e)(\bar{b}e^{-}\bar{\nu}_e) \tag{1}$$

$$t\bar{t} \to (b\mu^{+}\nu_\mu)(\bar{b}\mu^{-}\bar{\nu}_\mu) \tag{2}$$

are used, denoted as ee and $\mu\mu$ channel in the following.

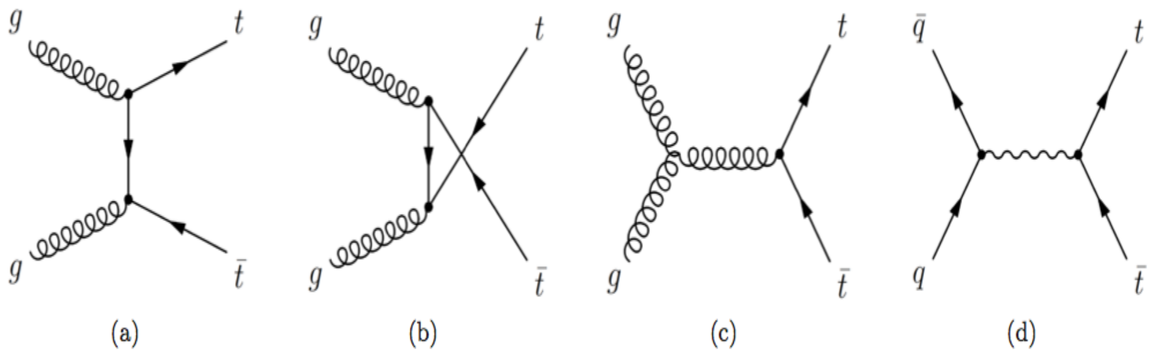

Figure 2: Feynman diagrams for top production processes at lowest order: (a), (b), and (c) gluon-gluon fusion, (d) quark-antiquark annihilation [3].

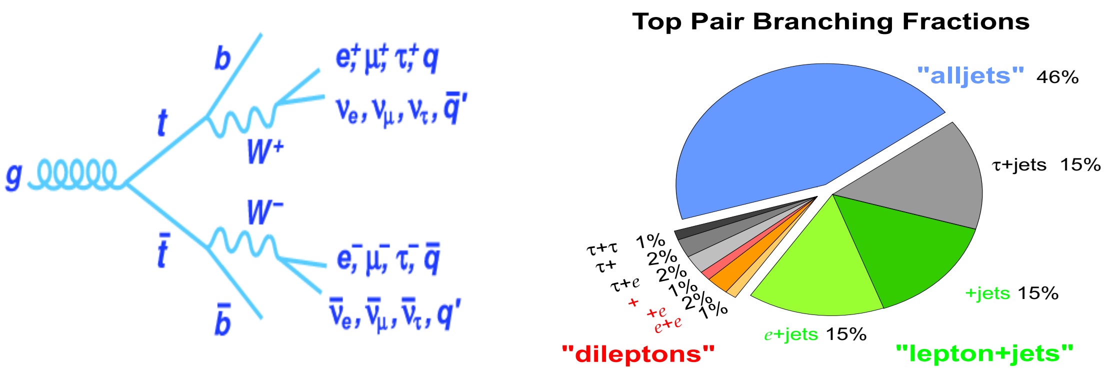

Figure 3: Top pair decay modes (left) and corresponding branching fractions (right) [4].

2.3 Simulated event samples

Monte-Carlo (MC) simulation samples are used to study the signal and background efficiencies of the $t\bar{t}$ selection described in Section 3 and 4.

The $t\bar{t}$ signal events are generated with POWHEG [5] and PYTHIA [6].
Single top quark background, which arises from the associated Wt production, is modeled with MC@NLO v4.01 [7]. Background due to diboson production is generated using Alpgen v2.13 [8] interfaced with Herwig [6]. Finally, Drell-Yan events ($Z \to e^+e^-, \mu^+\mu^-$ plus jets) are modeled using the Alpgen v2.13 generator including leading-order matrix elements with up to five additional partons. In Section 3.1 further details about the background composition are given.

3 Cut based selection of top pair events in the dilepton decay channel

The $t\bar{t} \rightarrow (b\bar{l}\nu_l)(\bar{b}l\bar{\nu}_l)$ event signature involves two b-jets, two leptons and two neutrinos, which leave the detector without being detected. The selection exploits this event topology in order to separate true $t\bar{t}$ events from several processes which could mimic the same signature. In this Section a cut based selection is presented (Section 3.1) and the corresponding cut efficiencies are discussed (Section 3.2).

3.1 Selection criteria

Events with exactly two oppositely charged leptons of the same flavor are selected (either e^+e^- or $\mu^+\mu^-$). Furthermore, at least two jets are required, where at least on of them is tagged as a b-jet. For this purpose, jets originating from b-quarks are identified by exploiting the long lifetime of b-hadrons ($\approx 1.5\,ps$) which leads to a displaced secondary vertex [9]. The working point of the neural network algorithm, which combines information about the secondary vertex and displaced tracks associated with the jet, is chosen such that the algorithm identifies b-jets from top quark decays with 70% efficiency [10].

The top mass is much higher than the rest mass of the decay products leading to high momentum leptons and jets. Therefore, a transverse momentum of at least $p_T = 25\,GeV$, i.e. the momentum in the plane perpendicular to the beam axis, of the decay products is required in order to suppress low momentum background.

The dominant background originates from $Z + jets$ events, with Z decaying into ee or $\mu\mu$, as shown in Figure [4]. Hence, events where the dilepton invariant mass is consistent with the nominal Z boson mass are discarded (Z veto: $|m(ll) - m(Z)| > 10\,GeV$). In addition, the diboson (WW, WZ, ZZ) or single top production, as shown in Figure [4], could be wrongly reconstructed as a $t\bar{t}$ decay. The missing transverse energy (E_T^{miss}) can be used as an indicator of unobserved neutrinos. Since the signal decay mode involves two neutrinos the missing transverse energy is on average much larger for signal than for background events, as can be seen on Fig. [5]. Consequently the cut $E_T^{miss} > 60\,GeV$ is applied. All selection criteria are summarized in Table 1.

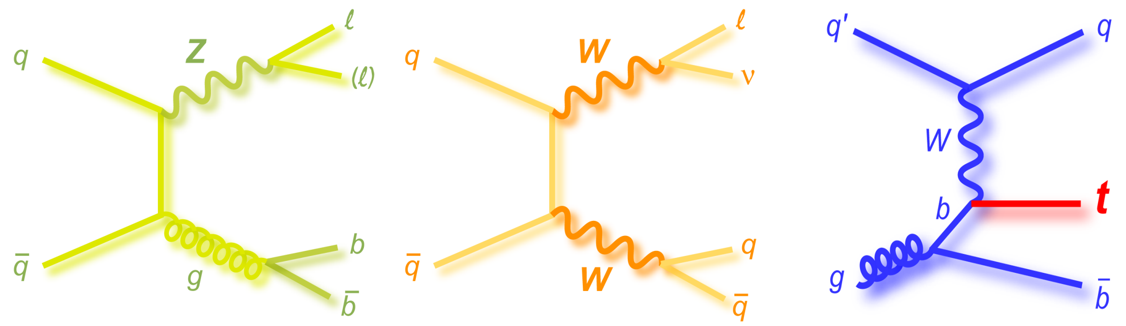

Figure 4: Background sources: Z+Jets (left), diboson (middle) and single top (right) [4].

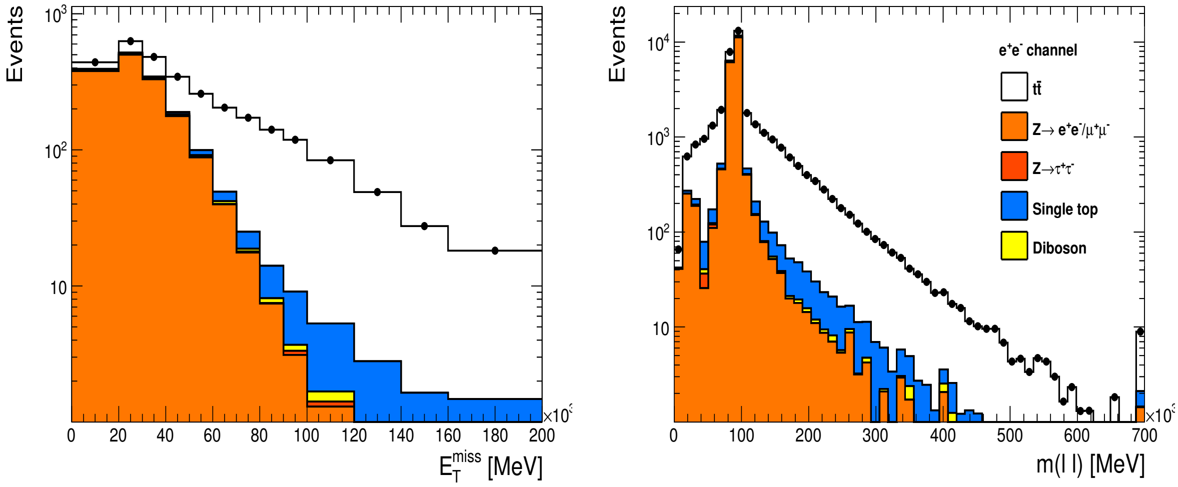

Figure 5: Distribution of E_T^{miss} (left) and $m(ll)$ (right) for $t\bar{t}$ signal and various background sources.

Variable	Cut
$N_{Leptons}$	$= 2$
N_{Jets}	≥ 2
N_{bTags}	≥ 1
$p_T(leptons, jets)$	$> 25\,GeV$
$m(ll)$	$> 15\,GeV$
$\lvert m(ll) - m(Z)\rvert$	$> 10\,GeV$
E_T^{miss}	$> 60\,GeV$

Table 1: Selection of top pair events in the dilepton decay channel

3.2 Cut efficiencies

After applying all selection criteria, the signal to background ratio (S/B) is 15.67 (14.5) for the ee ($\mu\mu$) channel. By removing one particular cut of the selection (but all others are applied), it can be checked how powerful or efficient the respective requirement is (see Figure [6]). The constraint on the missing transverse energy as well as the b-tagging have the largest impact on the signal to background ratio, i.e. removal of these requirements lead to a considerably higher background level, as show in Figure [6]. In particular the requirement $E_T^{miss} > 60\,GeV$ decreases significantly the signal efficiency from $\epsilon_S = 1.05\%$ ($\epsilon_S = 1.15\%$) to $\epsilon_S = 0.62\%$ ($\epsilon_S = 0.68\%$) for the ee ($\mu\mu$) channel. Altogether the cuts on E_T^{miss} and $m(ll)$ discard more than half ($\approx 51\%$) of the signal events . In Section 4, the utilization of multivariate analysis tools is studied, instead of rectangular cuts on E_T^{miss} and $m(ll)$, in the hope of devising a more efficient selection.

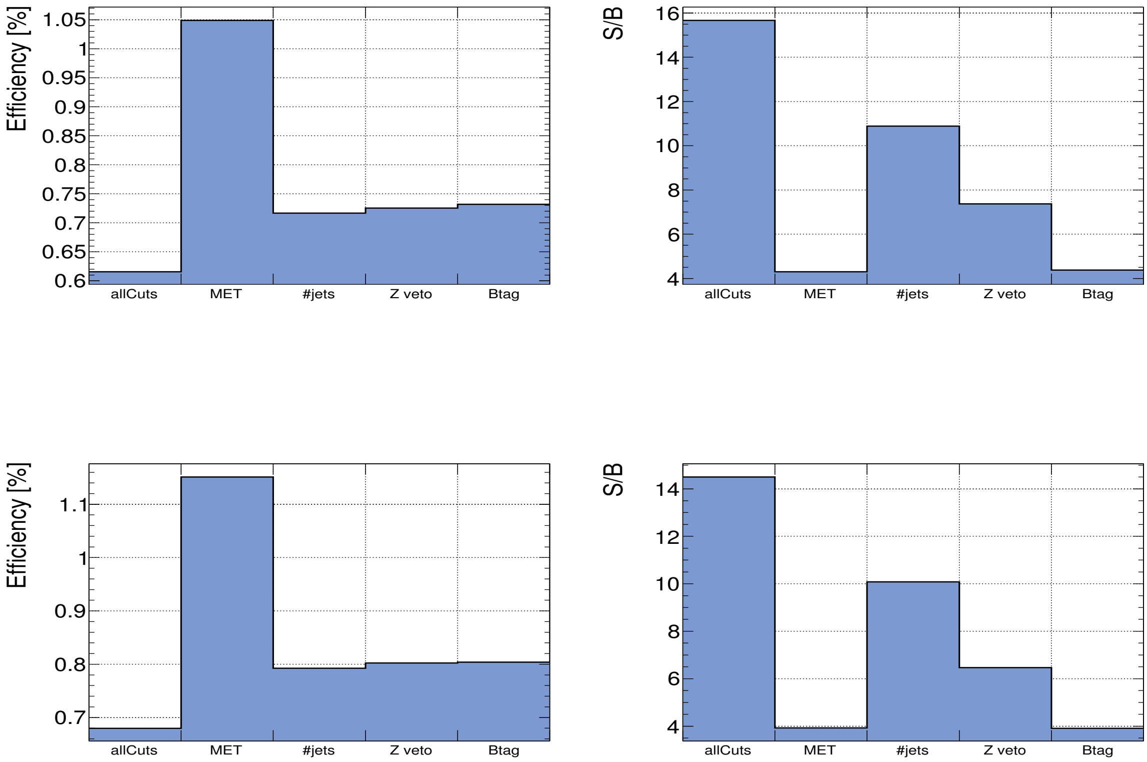

Figure 6: Signal efficiency (left) and signal to background ratio (right) after applying all cuts except one particular cut. Top: ee channel. Bottom: $\mu\mu$ channel.

4 Multivariate analysis

Instead of a simple cut-based selection, multivariate analysis tools can be used to discriminate signal from background. A multivariate classifier maps the n-dimensional space of the observable variables $\vec{x} = \{x_1, x_2, ..., x_n\}$ to an one dimensional output called the classifier response :

$$t(\vec{x}) : \mathbb{R}^n \to \mathbb{R} \tag{3}$$

The classifier response combines the information of the input variables, including their correlation, into one powerful discriminator.

Multivariate analysis techniques are based on supervised machine learning algorithms, which make use of clean signal and background samples in order to find the mapping function $t(\vec{x})$. These samples can be taken from MC simulation. Due to the limited statistics of the training samples the classifier could be overtrained, which means that the machine learning does not pick up actual signal or background properties, but statistical fluctuations. The performance of an overtrained classifier is better on the training sample than on any statistical independent data sample, therefor overtraining can be detected by comparing the performance between the training and a independent test sample [2]. Finally, the trained classifier can be applied to a data sample with unknown composition.

In Section 4.1 boosted decision trees (BDT), the multivariate analysis technique used in this analysis, are described. The training phase of the BDT is discussed in Section 4.2. Finally, the performance of the BDT is compared to the cut based selection (Section 4.3). The Toolkit for Multivariate Data Analysis (TMVA [11]) is used to train and evaluate the multivariate classifier.

4.1 Boosted decision trees

Decision trees (DT) are a natural extension of simple cuts but instead of discarding all events that fail a certain cut, wrongly classified events get a second chance to be classified correctly. Therefor, a much higher signal efficiency can be achieved.

A decision tree categorizes the events of a data sample based on a successive application of binary splits, as sketched in Figure [7]. Starting from the root node, a sequence of cuts divide the data into signal- and background-like subsamples. At each node of the DT the discrimination variable which provides the best separation power [3] is used to determine the optimal cut criterion. The division is repeated until a node has reached a certain minimum number of events (3 % of the total events in this analysis) or the maximum tree depth (equal to three in this analysis) is reached. These final "leaf" nodes are classified to be either signal or background-like according to the majority of the events inside the respective leaf. Consequently the discrete valued response function of the DT returns $DT(\vec{x}) = +1$ ($DT(\vec{x}) = -1$) if an event $\vec{x}$ ends up in a signal (background) leaf.

A shortcoming of decision trees is their instability with respect to statistical fluctuation in the training sample (e.g. the decision of the optimal cut criterion at a certain node may

[2] Half of the events are used for training, the other half for testing

[3] The quality of separation is defined by the so called Gini index: $p \cdot (1 - p)$, where $p = \frac{S}{S+B}$.

be influenced by a statistical fluctuation in the training sample), i.e. they are sensitive to overtraining. To stabilize the DT response and significantly improve the performance a so called boosting is applied. The principle behind boosting is that misclassified events from the training sample are given a larger weight than events which are in the correct leaf node. The adaptive boost algorithm (AdaBoost [11]) re-weights misclassified events with the common boost weight:

$$\alpha = \frac{1 - err}{err}, \tag{4}$$

where err is the fraction of misclassified events. The resulting reweighed training sample is then used to train a new decision tree. Repeating the boosting procedure several times (500 times in this analysis) leads to a set of decision trees (called a "decision forest"), where each tree learns from the errors of the previous ones. In the end, the boosted classifier response is given by the weighted average of the individual ones :

$$BDT(\vec{x}) = \frac{1}{N_{trees}} \sum_{i}^{N_{trees}} ln(\alpha_i) \cdot DT_i(\vec{x}) \tag{5}$$

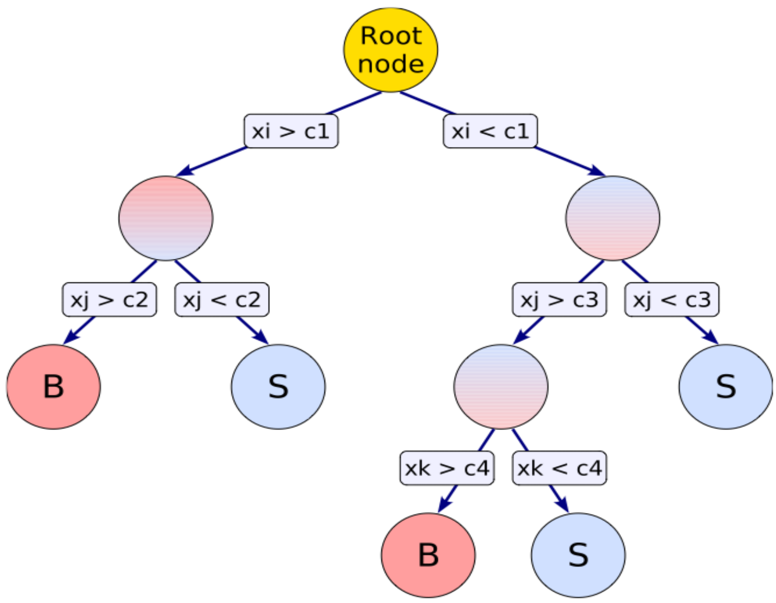

Figure 7: Schematic view of a decision tree. Starting from the root node, a sequence of binary splits using the discriminating variables $\vec{x}$ is applied to the data. [11]

4.2 Training of the multivariate classifier

The BDT is trained using the MC samples already described in Section 2.3. All cuts except the requirements on E_T^{miss} and $m(ll)$ are applied. In addition to E_T^{miss} and the absolute value of the difference of the dilepton invariant mass with respect to the nominal Z boson mass $\Delta m(ll) = |m(ll) - m(Z)|$, the transverse momenta of the two leptons and of the two leading jets and the angle between the two leptons $\Delta\phi$ are used as discrimination variables. A further input variable is the MV1 of the leading jet, which is the output of the multivariate classifier, already described in Section 3.1, dedicated to identify b Jets. The signal and background distributions of the training variables are shown in Figure [8] and [9]. In Table [2] the input variables are ranked according to their separation $\langle S^2 \rangle$ calculated by [11]:

$$\langle S^2 \rangle = \frac{1}{2} \int \frac{(\mathcal{P}_S(x) - \mathcal{P}_B(x))^2}{\mathcal{P}_S(x) + \mathcal{P}_B(x)} dx \,, \tag{6}$$

where $\mathcal{P}_S(x)$ and $\mathcal{P}_B(x)$ are the signal and background probability functions of the classifier x. For identical signal and background shapes the separation is zero and one in case of no overlap at all. The classifier response, which has a much better separation power of $\langle S^2 \rangle = 76\%$ than the individual training variables, is plotted in Figure [10].

Variable	Separation [%]
$m(ll)$	56.9
E_T^{miss}	39.8
Leading $p_T(jet)$	4.7
$MV1$	3.7
Sub-leading $p_T(jet)$	3.7
Sub-leading $p_T(lep)$	2.2
Leading $p_T(lep)$	2.0
$\Delta\Phi$	0.5

Variable	Importance [%]
$m(ll)$	23.0
E_T^{miss}	20.1
$\Delta\Phi$	11.2
$MV1$	11.1
Leading $p_T(lep)$	10.6
Leading $p_T(jet)$	9.8
Sub-leading $p_T(lep)$	7.3
Sub-leading $p_T(jet)$	6.8

Variable	Separation [%]
$m(ll)$	56.9
E_T^{miss}	40.1
Leading $p_T(jet)$	5.0
$MV1$	4.6
Sub-leading $p_T(jet)$	3.8
Leading $p_T(lep)$	2.1
Sub-leading $p_T(lep)$	2.0
$\Delta\Phi$	0.6

Variable	Importance [%]
E_T^{miss}	22.1
$m(ll)$	20.7
$\Delta\Phi$	11.3
$MV1$	10.6
Leading $p_T(lep)$	10.2
Leading $p_T(jet)$	8.7
Sub-leading $p_T(lep)$	8.3
Sub-leading $p_T(jet)$	8.0

Table 2: BDT input variables ranked according to their separation (left) and importance (right) for the ee channel (top) and $\mu\mu$ channel (bottom).

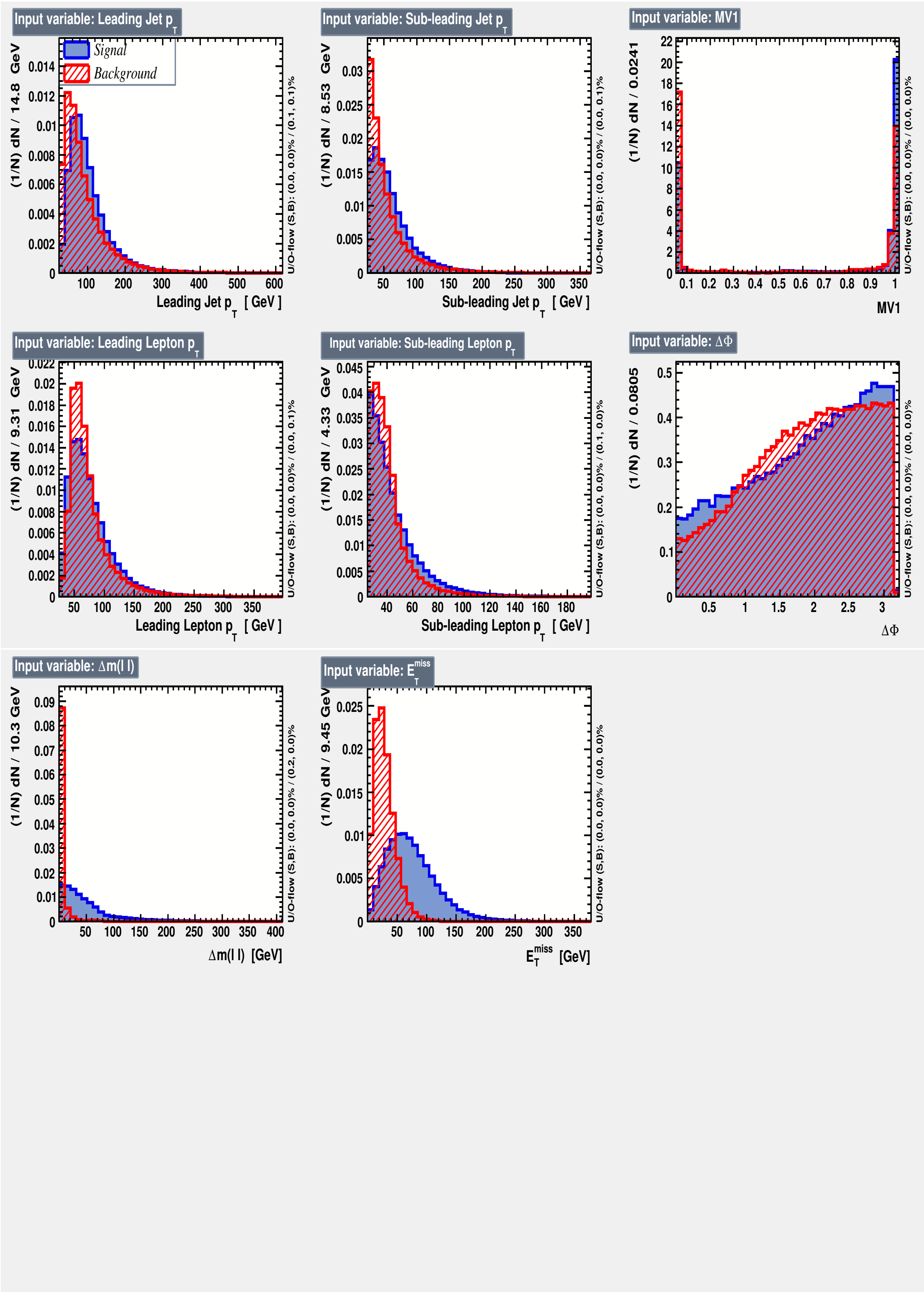

Figure 8: Input variables used to train the BDT (ee channel).

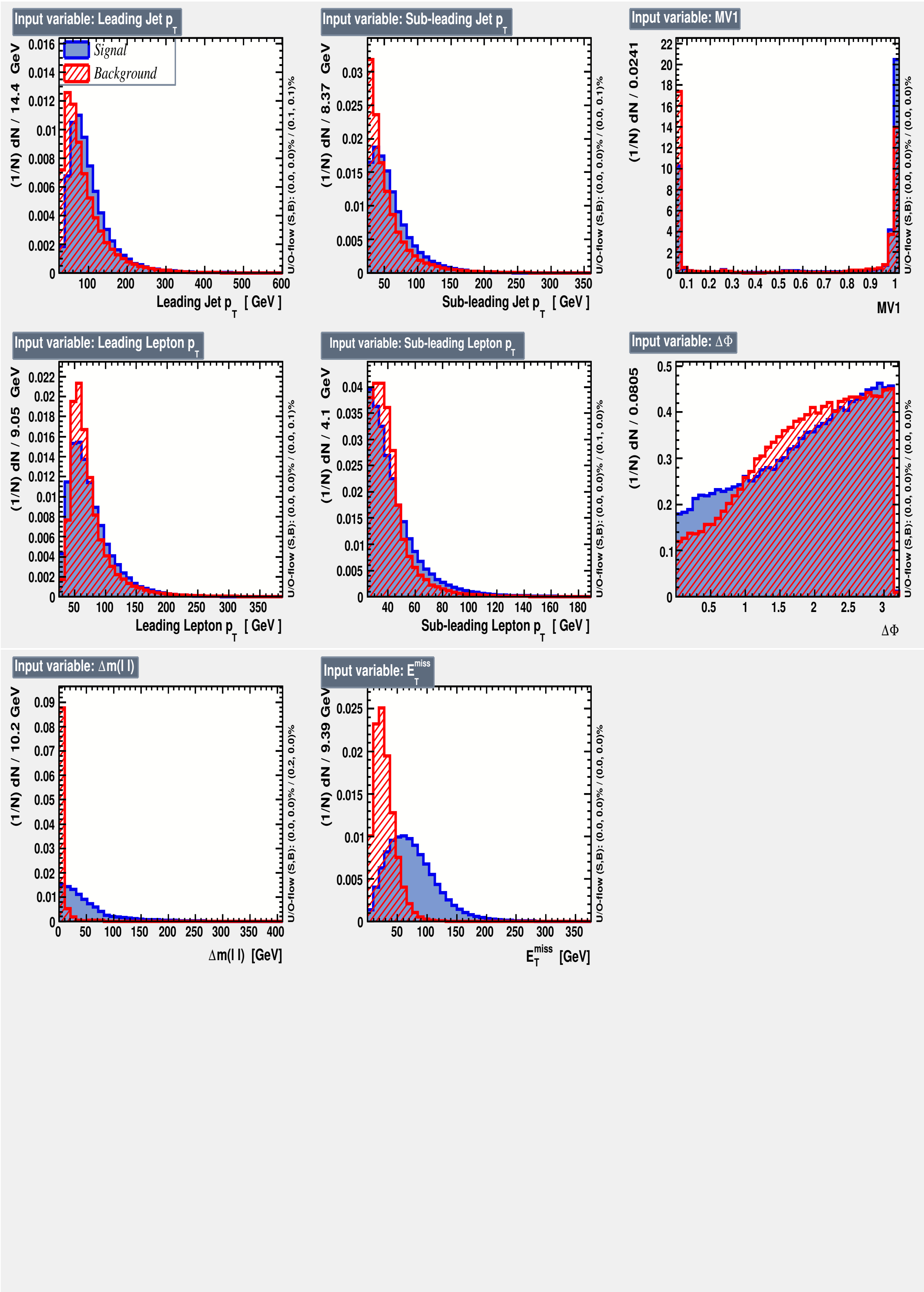

Figure 9: Input variables used to train the BDT ($\mu\mu$ channel).

A measure of the variable importance can be derived by counting how often the variable is used to split decision tree nodes, and by weighting each split occurrence by the separation gain-squared [4] it has achieved and also by the number of events in the node [11]. Note that this does not fully reflect the variable importance since removing one variable could be compensated by correlated variables and sometimes only combination of variables makes sense. For the latter, $\Delta\phi$ can serve as an good example since it is not immediately obvious why this discriminator takes third place at the importance ranking, shown in Table [2], although the separation power is very low. The reason for this becomes evident when inspecting the correlation between the training variables. In Figure [11], it is clear that there is a positive correlation between $\Delta\phi$ and $p_T(lep, jet)$ for signal events whereas the correlation is negative for background events. The training algorithm makes use of the different correlation to discriminate between signal and background. Consequently only the combination of $\Delta\phi$ and $p_T(lep, jet)$ is useful.

For the purpose of an overtraining check, the BDT response evaluated with the training sample and the independent test sample are plotted superimposed on Figure [10]. The BDT shows similar performance in both cases, hence, there is clearly no indication of overtraining.

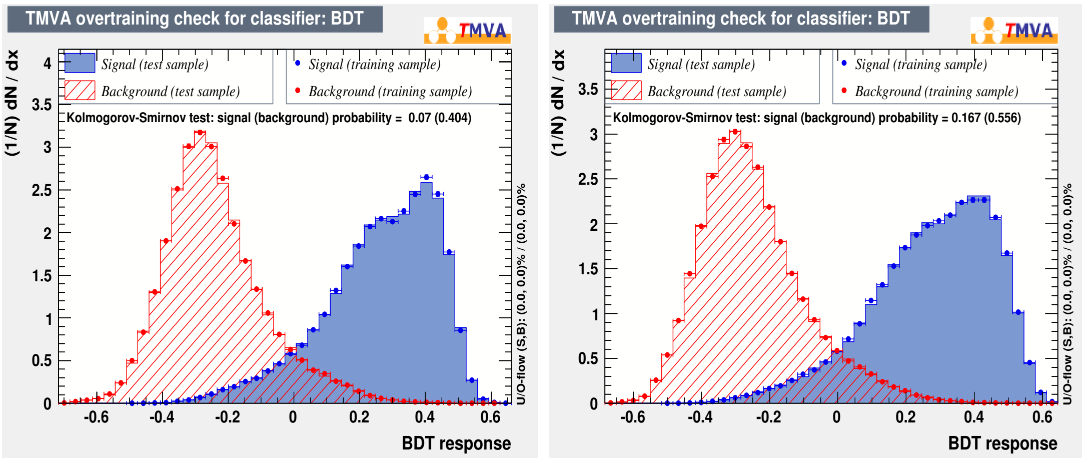

Figure 10: Signal and background distributions for the BDT response (test and training samples are superimposed to probe overtraining).
Left: ee channel. Right: $\mu\mu$ channel.

[4]The separation gain is defined as : $g(\text{parent node}) - g(\text{daughter node 1}) - g(\text{daughter node 2})$, where $g = p(1-p)$ is the Gini index.

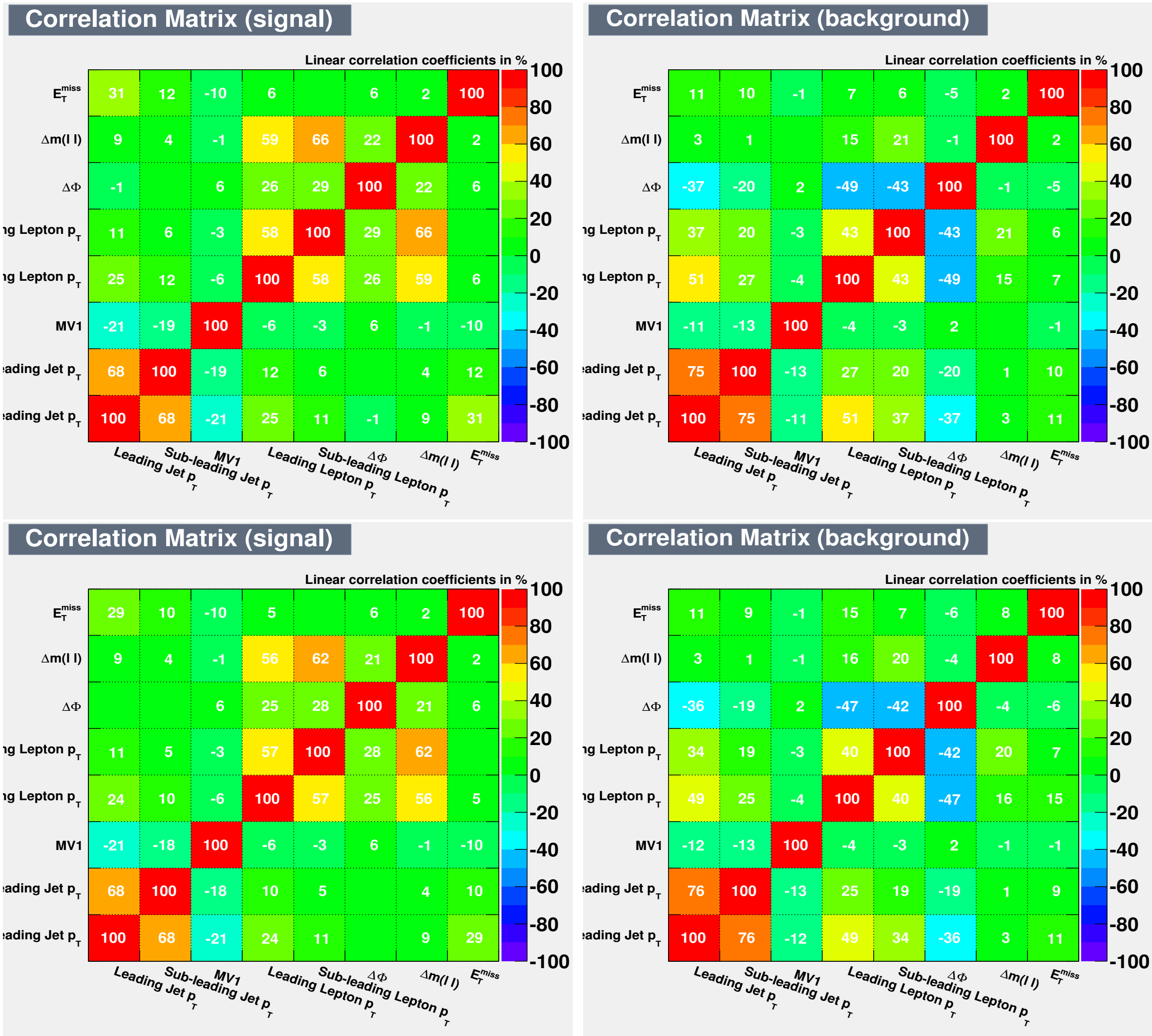

Figure 11: Correlation between input variables. Top: ee channel. Bottom: $\mu\mu$ channel.

4.3 Performance of the multivariate classifier

The purpose of the selection is to reduce the background level as much as possible while maintaining a high signal yield. It is desirable to maximize the signal significance $S/\sqrt{S+B}$, which is a measure of the statistical precision of the measurement. The corresponding cuts on the classifier response are $BDT > 0.026$ and $BDT > 0.020$ achieving the significance $S/\sqrt{S+B} = 113.1$ and $S/\sqrt{S+B} = 122.0$ for the ee and $\mu\mu$ channel, respectively. In contrast, the significance obtained with the cut based selection described in Section 3 is only $S/\sqrt{S+B} = 81.8$ ($S/\sqrt{S+B} = 84.5$) for the ee ($\mu\mu$) channel. In order to compare the BDT performance to the cut based selection properly, the signal efficiencies should be compared at a fixed background rejection or vice-versa (see Table [3]). At a background efficiency of $\epsilon_B = 2.4\%$ the application of the BDT enhances the signal efficiency from $\epsilon_S = 49\%$ to $\epsilon_S = 80\%$ for the ee channel ($\epsilon_S = 49.5\%$ to $\epsilon_S = 79\%$ at a background efficiency of $\epsilon_B = 2.2\%$ for the $\mu\mu$ channel).

Note that the performance of the two different lepton channels is almost equal. Although these channels are similar from a pure physical point of view (lepton universality) this is not immediately obvious given that the way how electrons and muons are reconstructed is completely different (calorimeters versus muon chambers).

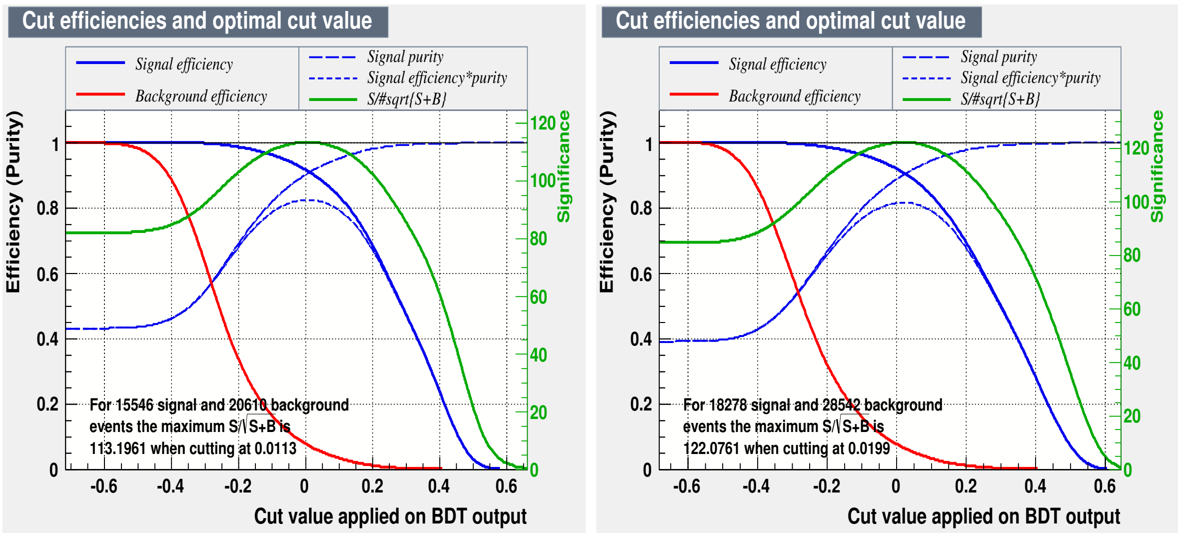

Figure 12: Signal and background efficiencies, significance and purity as a function of the BDT cut value. Left: ee channel. Right: $\mu\mu$ channel.

Selection	$\epsilon_S[\%]$	$\epsilon_B[\%]$	S/B	S/(S+B)	$S/\sqrt{S+B}$
Preselection	100	100	0.75	0.43	81.8
Cut	0.49	0.024	15.67	0.94	84.8
BDT > 0.026	0.90	0.062	10.9	0.916	**113.1**
BDT > 0.132	0.80	**0.024**	24.7	0.96	109.2
BDT > 0.291	**0.49**	0.003	114.7	0.99	87.2

Selection	$\epsilon_S[\%]$	$\epsilon_B[\%]$	S/B	S/(S+B)	$S/\sqrt{S+B}$
Preselection	100	100	0.64	0.39	84.47
Cut	0.495	0.022	14.5	0.94	92.0
BDT > 0.020	0.9	0.065	8.95	0.90	**122.0**
BDT > 0.138	0.79	**0.022**	23.1	0.96	117.7
BDT > 0.302	**0.495**	0.003	105.4	0.99	94.7

Table 3: Comparison of the multivariate classification with the cut based selection. Top: ee channel. Bottom: $\mu\mu$ channel.

Finally, the performance of the BDT is cross-checked using different multivariate classifiers. Among them are artificial neural networks , linear fisher discriminants and projective likelihood estimators (a detailed description of these multivariate analysis techniques can be found in Ref. [11]) . A good indicator of the performance of a multivariate classifier is the integral of the receiver operating characteristic (ROC) curve, which is given by the background rejection as a function of the signal efficiency, as shown in Figure [13]. None of the others classifiers have a better performance than the BDT.

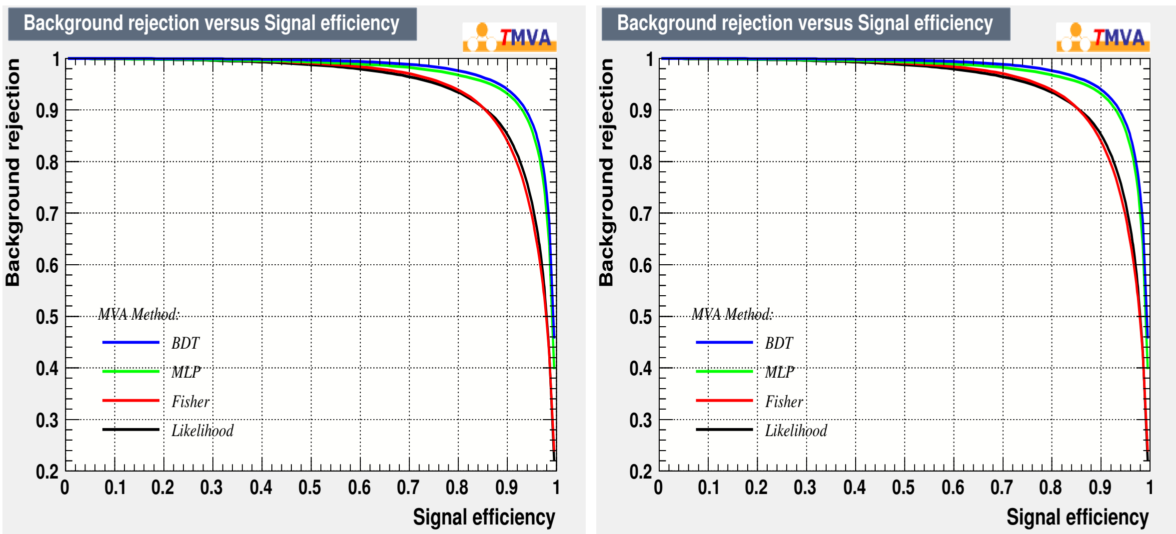

Figure 13: ROC curve: Signal and background efficiency as a function of the cut on the classifier output for boosted decision tree (BDT), artificial neural networks (MLP), linear fisher discriminants (Fisher) and projective likelihood estimator (Likelihood). Left: ee channel. Right: $\mu\mu$ channel.

5 Conclusions

In this analysis the cut based selection of $t\bar{t}$ candidates using the ee and $\mu\mu$ decay channels is studied. Especially, the requirements on the missing transverse energy ($E_T^{miss} > 60\,GeV$) and the dilepton invariant mass ($m(ll) > 15\,GeV$, $|m(ll) - m(Z)| > 10\,GeV$) discarding more than half of the signal events are found to be too tight. Consequently, these rectangular cuts are removed and a boosted decision tree is trained instead in order to discriminate signal from background. In doing so, the signal efficiency is enhanced from $\epsilon_S = 49\%$ to $\epsilon_S = 80\%$ at a background efficiency of $\epsilon_B = 2.4\%$ for the ee channel and from $\epsilon_S = 49.5\%$ to $\epsilon_S = 79\%$ at a background efficiency of $\epsilon_B = 2.2\%$ for the $\mu\mu$ channel. Furthermore, a maximum signal significance of $S/\sqrt{S+B} = 113.1$ and $S/\sqrt{S+B} = 122.0$ for the ee and $\mu\mu$ channel, respectively, can be achieved. In conclusion, the selection of $t\bar{t}$ events in the dilepton decay modes ee and $\mu\mu$ is significantly improved by using a multivariate classifier instead of rectangular cuts.

The selection of $t\bar{t}$ events in the $e\mu$ channel does not suffer from large Drell-Yan background, such that a cut based selection, similar to the one presented in Section 3, is already very powerful [10]. Nevertheless, the application of multivariate analysis techniques may lead to a gain in signal efficiency, too.

References

[1] A. Quadt, *Top quark physics at hadron colliders*. Eur. Phys. J. C 48, 8351000 (2006)

[2] The ATLAS Collaboration, *ATLAS Experiment at the CERN Large Hadron Collider*. JINST 3 (2008), S08003

[3] A. Gonzalez, *Search for flavor changing neutral currents in top quark decays using two-leptons+track selection*. CERN-THESIS-2012-153.

[4] `http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/top_feynman_diagrams.html`

[5] S. Alioli et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*. arXiv:1002.2581 [hep-ph].

[6] A. Buckley et al., *General-purpose event generators for LHC physics*. arXiv:1101.2599 [hep-ph].

[7] S. Frixione et al., *Single-top hadroproduction in association with a W boson* . JHEP 07, 029 (2008).

[8] M. L. Mangano et al., *ALPGEN, a generator for hard multiparton processes in hadronic collisions*. JHEP 07, 001 (2003).

[9] The ATLAS Collaboration, *Measurement of the top quark-pair production cross section with ATLAS in pp collisions at $\sqrt{s} = 7\,TeV$*. arXiv:1012.1792 [hep-ex].

[10] The ATLAS Collaboration, *Measurements of spin correlation in top-antitop quark events from proton-proton collisions at $\sqrt{s} = 7\ TeV$ using the ATLAS detector*. arXiv:1407.4314 [hep-ex].

[11] A. Hoecker et al., *TMVA - Toolkit for Multivariate Data Analysis*. arXiv:physics/0703039.