TMVA Study for $B'B'$ Search

J. Guo$^1$, D. Hu$^1$, J. Parsons$^1$, and K. Stockton$^2$

$^1$Columbia University  
$^2$Hamilton College

August 2, 2013

Abstract

Successfully separating background from signal is vital to finding a small signal amongst a large background in High Energy Physics. This can be very difficult when just using the standard cut-based separation, and computer-modeled multivariate analysis processes can be more efficient and useful. A list of 36 variables pertaining to both background and signal was narrowed down to 9 important variables using TMVA. These 9 variables were then examined for their accuracy when comparing Monte Carlo generated background and 2012 data in control regions. Though more tests should be undertaken, TMVA seems to be more efficient than the previous cut-based selection. TMVA could be a much more efficient and useful tool to differentiate between background and signal than cut-based selection to help with the $B'B'$ search.
1 Introduction

1.1 The Standard Model

The Standard Model is the most successful physical description of the particles and the interactions between those particles ever made [2]. Unfortunately, however, there are many physical interactions that the Standard Model does not touch upon, like gravity, or why there are no antiparticles that persisted after the Big Bang. Because of these inconsistencies thousands of scientists are searching for extensions and modifications to the Standard Model, often called Beyond Standard Model, or New Physics.

The current Standard Model is made up of matter particles, forces, and carrier particles. The matter particles are made up of 6 quarks and 6 leptons, and the force carriers are made up of bosons (gluons, photons, W and Z bosons, and the Higgs Boson)[2]. There are currently considered to be 3 generations of quarks, each generation having an up-type and a down-type quark. The visual representation of the Standard model can be seen in figure 1.

1.2 Motivation

One such search is looking for forth generation Vector-Like Quarks (VLQs), considered to be a natural extension of the Standard Model. The Higgs results from July 2012 have modified these searches to replace the search for chiral quarks with VLQs, though the previous search for chiral quarks can help guide the search for VLQs.
1.3 Decay Channel

If the previously searched-for chiral quark had a mass over $m_t + m_W$ then it would decay predominantly as $d_4 \rightarrow Wt$ [4]. The fourth-generation VLQ $B'$, however, would also decay to $Hb$ and $Zb$. The chiral $B'$ channel being used for analysis optimization, so we are focusing on the $B' \rightarrow Wt$ decay and thus are looking at at $B'B' \rightarrow ttWW \rightarrow WWbWWb$ [3].

Thus 4 $W$ bosons remain that can decay hadronically (to an up-type and down-type quark) or leptonically (to a lepton and neutrino). The single-lepton decay channel is predominantly examined because the branching ratio of the two-lepton decay is far lower than that of the single-lepton decay. This means that the main decay mode that we’re looking for is

$$B'B' \rightarrow ttWW \rightarrow WWbWWb \rightarrow (l\nu)(qq')b(qq')b.$$  \hfill (1)

As usual, the main challenge of determining if new physics exists is determining the accurate level of background produced. The major background levels come from $tt$ with jets from radiation, followed by $W +$ jets, which has high theoretical uncertainty due to lack of precise knowledge of gluon radiation. The backgrounds can be controlled and calculated correctly by examining background-dominated regions (which are henceforth referred to as 'control regions') and comparing collected data to Monte Carlo and data-derived background [3].
2 The ATLAS Detector at the LHC

2.1 The LHC at CERN

The Large Hadron Collider (LHC) at CERN (European Organization for Nuclear Research) in Geneva, Switzerland is a 27 km ring where two proton beams accelerated to 0.9 times the speed of light are made to collide. After collision different particles emerge and detectors placed around the LHC can detect what they are and where they emerge from the central beam line. The LHC is used to find New Physics because it runs at very high energy levels, and thus can produce higher-massed particles that were previously undiscovered.

We can’t just look at the LHC, however. We need some way to look at these particles. This brings us to the four major detectors situated around the LHC (ATLAS, LHCb, ALICE, and CMS). Each has its own focus, but all of the experiments have in common the fact that proton beams come together and collide inside their detecting area and release sprays of particles which leave tracks that can be identified through various tracking systems[1].

2.2 The ATLAS Experiment

ATLAS (A Toroidal LHC ApparatuS) is an experiment (detector) on the LHC beam that is half the size of Notre Dame (45 m long, more than 25 m high) and measures the results of these collisions and the importance of these results. It consists of four major sections: the inner detector, the calorimeters, muon spectrometer, and magnet system, all of which aid to track particles that result from proton-proton collisions. ATLAS is seeking to learn about the "basic forces that have shaped our Universe since the beginning of time and that will determine its fate"[10].

3 Analysis Overview

The main decay mode that is looked at is $B'B' \rightarrow ttWW \rightarrow WWbWWb \rightarrow (l\nu)(qq')(b\nu)(qq')b$, so we expect an electron or muon and $E_T^{miss}$ from the leptonic $W$ decay, 4-6 light jets from the hadronic $W$ decays, and 2 $b$-tagged jets. Due to time constraints only the electron channel was examined[5].

3.1 Object Selection

The hadronic $W$ bosons are expected to be very highly boosted and thus the two daughter jets from hadronic decays should be very close to each other, even occasionally overlapping. Therefore in order to reconstruct hadronic $W$ decays the analysis requires a single jet with a mass of that around that of the $W$ boson mass, or a pair of jets separated from each other by $\Delta R < 1$ with a combined mass of that of the $W$ boson mass. Electrons are selected with $|\eta_{cluster}| < 2.47$, but excluding the regions of $1.37 < |\eta| < 1.52$. They must have $E_T > 25$ GeV, with transverse energy defined by $E_T = E_{cluster}/\cosh(\text{track eta})$. Signal electrons
must pass the tight++ identification requirements, pass the standard Object Quality (OQ) cuts, and be isolated in the calorimeter and tracker. All electrons within an $\eta - \phi$ cone of 0.4 of any jet with $E_T > 25$ GeV are removed. Muons are selected with the Combined MuDrMuon algorithm and must have $|\eta| < 2.5$ and $p_T > 20$ GeV. They must pass the tight muon requirements and all muons within an $\eta - \phi$ cone of 0.4 of any jet with $E_T > 25$ GeV are removed. Jets are reconstructed from topological calorimeter clusters using $R = 0.4$ anti-$k_t$ algorithm. They have EM and JES calibrations applied and must satisfy $E_T > 25$ GeV and $|\eta| < 2.5$. The jet vertex fraction must satisfy JVF > 0.50. The closest jet to an accepted electron in an $\eta - \phi$ cone of 0.2 is removed[3].

3.2 Signal Region

The signal region is defined by having at least 6 jets, exactly one lepton, at least one b-tagged jet, at least one reconstructed $W$ boson, and an $H_T$ of at least 800 GeV. Due to combinatorics the $B'$ mass is very difficult to reconstruct, so instead of putting a lower limit on reconstructed mass it is put on $H_T$, which is defined as

$$H_T = \sum_{\text{jet}} p_T^i + p_T^{lep} + E_{miss}^T$$

and is fairly effective at distinguishing between higher mass $B'$ events and lower mass backgrounds[3].

3.3 Samples

The analysis applies the Good Run List (GRL) to the 2012 data set resulting in a total integrated luminosity of 20.3 fb$^{-1}$[5]. All Monte Carlo samples are developed using mc12 and re-weighted to match the pileup distribution of the data.

The signal $b' 800$ GeV sample was generated Atlfast (AF-II). The predominant background is $t\bar{t}$ with additional jets, followed by $W +$ jets, $t\bar{t} + V$, single top, $Z +$ jets where a lepton is missed, diboson production, and QCD background in which a jet is misidentified as a lepton [3]. Production of $t\bar{t}$ is modeled using Alpgen with up to 3 additional partons. $W +$ jets and $Z +$ jets are modeled by Alpgen with up to 5 light partons with $W +$jets normalized using scaling factors derived from data. $t\bar{t} + V$ is modeled using Madgraph and Pythia and diboson production is modeled using Herwig. Single top is modeled using MC@NLO and AcerMC. QCD background is determined from data using the matrix method[3]. Plots shown in this paper group $t\bar{t} + V$, diboson production, $Z +$jets and Single Top together as "Other backgrounds."

3.4 Preselection Cuts

The events must fit the following requirements during the preselection step [4]
- GRL and standard data quality and event cleaning cuts
- Single lepton trigger requirements with a trigger-matched lepton
- Exactly one electron or muon must pass the lepton selection requirements
- A primary vertex with at least 5 tracks
- At least one jet must pass the jet selection requirements
- Must not be one or more "LooseBad" jets in the AntiKt4TopoEMJets collection with $p_T^{jet} > 20$ GeV and $E^{jet} > 0$
- Must not have a selected electron that shares a track with a selected muon which passes all the requirements except the cut on the muon-jet separation
- "Triangle cuts" on $E_{T}^{miss}$ and transverse mass:
  - electron channel: $E_{T}^{miss} > 30$ GeV and $m_{T}^{W} > 30$ GeV
  - muon channel: $E_{T}^{miss} > 20$ GeV and $E_{T}^{miss} + m_{T}^{W} > 60$ GeV

3.5 Cut-Based Selection

The original method for differentiating between signal and background was to produce a nine-bin plot, which compares background and $b'$ signal in 9 different categories according to jet multiplicity ($n_{jet} = 6, 7, \geq 8$) and according to the number of hadronic $W$ candidates ($n_{W} = 0, 1, \geq 2$).

These regions allowed for comparison between the weighted number of events in both monte carlo generated signal and background in 9 different ways, showing how the shapes of the two samples (signal and background) differ. Unfortunately when comparing the actual number of weighted events the signal sample does not always end up being larger than the background samples, even in the signal region. This can be seen in figure 3, where there are two different $H_{T}$ cuts to show the difference when looking at different parts of the signal region.

4 TMVA (Toolkit for MultiVariate data Analysis with ROOT)

4.1 Introduction

TMVA is a program that is used through ROOT that uses machine learning to differentiate between signal and background. TMVA is a suite of programs, but the main program used in the current analysis is instated using Boosted Decision Trees[6].
Figure 3: The first plot has $H_T > 800$ and the second has $H_T > 1000$. These nine-bin cut-based plots show that the previous cut-based selection is not as discriminating as previously hoped.

4.2 Boosted Decision Trees (BDTs)

4.2.1 Decision Trees

When differentiating between background and signal it is necessary to first have Monte Carlo samples of both, in order to both train the analysis and to see if the training was correctly done. Decision trees use a number of selected variables (from the analysis) that help differentiate between background and signal and make cuts on each variable at each "node" of the tree. Cuts are made that will most highly differentiate between background and signal for each variable. Each event will reach a node and "go down" one of two branches depending on the value of the variable represented at the node. Eventually after multiple cuts the event will reach a "leaf" and, after multiple events have gone through these cuts, there will be many events on each leaf.

Each of these leaves will be given a score based on the number of background vs signal events that land on it. Leaves with a score under 1 (i.e. there are more background events than signal events) are labeled background and leaves with a score over 1 are labeled signal. This process is illustrated in figure 4.

A trained decision tree is left that can then be tested by sending more Monte Carlo-generated events through the cut process, and if an event ends up landing on a "signal" leaf it is labeled as signal, or as background if it lands on a "background" leaf[8].
4.2.2 Boosting

Unfortunately decision trees are considered to be quite unstable, where a small change in the training sample can cause a large change in the tree and the results[8]. This brings boosting onto the stage. In the beginning all events are given equal weights. When an event lands on a leaf it is given a score of -1 if it lands on a background leaf and 1 if it lands on a signal leaf. If an event (during testing) lands on a signal leaf when it is actually background (or vice versa) then its weight is boosted (increased) and a new tree is made with the new weights. This can be repeated hundreds of times to result in a highly trained boosted decision tree. The renormalized sum of all of the scores is summed for each event. This results in each event having a score from -1 to 1, where an event having a score of -1 means it is unequivocally background, and a score of 1 implies that it could not be mistaken for anything but signal[8].

Once the trees are trained then "unknown" Monte Carlo samples can be run over them to test whether the trees were overtrained or if they were trained correctly by comparing their given BDT scores to the BDT scores of the original samples[8]. An example of this can be seen in figure 6.

4.3 Procedure

To correctly train TMVA we start with all of the variables that could be important to differentiate between background and signal. We started from a list of 36 variables, all describing a different part of the event. 36 variables, however, is far too many to perfect for...
Figure 5: These matrices show all 36 variables, with the highly correlated variables shown by the concentrated spots of red and yellow. These are then used to narrow down the list of variables to a more select 9.

The list of variables needs to be narrowed down to a list of 9 variables, all of which must have a high differential between signal and background and which must be not too highly correlated to other high-ranked variables in order to have 9 completely independent variables.

TMVA it produces multiple plots as a result of training, including correlation plots for background and signal samples as shown in figure 5, showing which variables are correlated and it also produces a list of variables ordered by their "rank" and importance to TMVA. This list orders the variables by their separation between background and signal and by other processes chosen by TMVA. The variables are narrowed down three or four at a time starting by reducing the number of correlated variables and then by eliminating the least important variables from the ranking list. This results in a collection of nine important and independent variables that can be used to effectively differentiate between signal and background.

TMVA can be trained with any desired selection cuts. For the following results TMVA was trained just in the signal region defined in section 3.2. It was trained using the $b'800$ signal sample and the $t\bar{t}$ and $W+jets$ background samples.
4.4 Results with Final 9 variables

After running TMVA over the nine most important variables in the signal region TMVA was evaluated to see if it was correctly trained and efficient.

4.4.1 Efficiency

We can see in figure 6 the -1 to 1 system originally introduced in section 4.2.2, and we can also see that the TMVA training works well by the close correlation between the test and training samples for both signal and background. We can also see that there is a large difference between signal and background, and that it is possible to have a very large background rejection vs signal efficiency from figure 7 through using TMVA.
4.4.2 Correlation

After eliminating the highly correlated variables the correlation matrices at 9 variables are a good measure of how correlated the resulting variables. We can see in figure 8 that the resulting 9 variables are highly independent, with none of the variables having a correlation score of over 40.
Figure 8: These are two correlation matrices for 9 variables. There are no correlations between variables with a score above 40, which shows that all of the variables are independent and do not rely too highly on other variables. This means that they can be very useful for background and signal differentiation[7].

5 TMVA Results

5.1 Control Region Cuts

The nine important variables that result from the TMVA selection must also be well calibrated in the main control regions so that we know that they would be defined as well in the signal region after unblinding. All 9 plots and a list of the important variables are included in Appendix A.

There are four main control regions that are defined to help isolate the two main backgrounds (\(t\bar{t}\) and \(W+\)jets) and determine if they are correctly calculated in Monte Carlo simulations. The first two help determine if \(W+\)jets is correctly calculated, being 4 or 5 jets with 0 b-tagged jets, or \(\geq 6\) jets with 0 b-tagged jets. The second two are \(t\bar{t}\) control regions, with 4 or 5 jets with \(\geq 1\) b-tagged jet, and \(\geq 6\) jets with \(\geq 1\) b-tagged jets and \(H_T < 500\) GeV. None of these touch the signal region, to avoid unblinding.

The first variable to be looked at is number of jets in figure 9, which shows that though all four control regions are fairly well calibrated, the \(W+\)jets regions are slightly more underestimated than the \(t\bar{t}\) regions, implying that the \(W+\)jets calculation is off. The recalculation is in the process of being rectified, but was not done early enough to include in this analysis.
Figure 9: Plots reflecting the number of jets measured in all four control regions. The $t\bar{t}$ region seems to be better normalized than the $W+\text{jets}$ region.

One set of plots, however, does not reflect this underestimation. The plots showing the $\Delta R$ between the lepton and a reconstructed hadronic $W$ seem to be fairly consistent between data and Monte Carlo. This is seen in figure 12.
5.2 BDT Scores for Background and B’800

TMVA may be able to be used for the elimination of certain theoretical models as well. As shown in figure 11 we can see the expected limit set on the scale factor for the x-section for $b’800$. With no systematics included the limit is 0.759 (when it needs to be <1 for exclusion) and with Jet Energy Scale systematics the limit is at 1.03 with the electron channel alone. This shows that using TMVA $b’800$ can almost be excluded with only one channel. This limit is calculated with an $H_T$ limit of $H_T>500$ GeV.
Figure 11: This plot shows the BDT scores of background and the $b^\prime 800$ sample, with a zoomed-in version on the signal region\[5\]

In comparison the cut-based selection based off of both the 9-bin plot (an example of which is shown in section 3.5) and a 6-bin plot (6 bins of number of 2 and $\geq 3$ Ws and 6, 7 and $\geq 8$ jets) need cuts of $H_T>1400$ GeV on both the electron and muon channels to achieve much less dramatic limits. The 9-bin plot achieves a limit of 0.868 without systematics and a limit of 1.81 with JES systematics, and the 6-bin plot achieves a limit of 0.832 without systematics and a limit of 1.15 with JES systematics, neither of which is as low as the TMVA result on only one channel. More plots relating to this analysis can be seen in Appendix B.

5.3 BDT Scores in Control Regions

The true test of how effective TMVA is for future analysis is through testing its application on control regions defined in 5.1. The control region samples were given TMVA BDT scores from TMVA trained both with no cuts at all and with signal region cuts defined in section 3.2. Unfortunately these plots do not seem to have nearly as much agreement between data and Monte Carlo as the plots from section 5.1. There does, however, seem to be slightly better agreement when TMVA was trained using the entire sample ($B^\prime 800$ for signal and $t\bar{t}$ with $W$+jets for background) than when there were cuts on just the signal region (like in section 4.3), especially in the $t\bar{t}$ control regions. This implies that it may be useful to train TMVA over whatever region in which it is being tested. Unfortunately research time at CERN came to an end before all potential analysis paths could be examined.
(a) No cut: $W+\text{jets}$: 4,5 jets, 0 b-tagged jets
(b) Signal Region: $W+\text{jets}$: 4,5 jets, 0 b-tagged jets

(c) No cut: $W+\text{jets}$: $\geq 6$ jets, 0 b-tagged jets
(d) Signal Region: $W+\text{jets}$: $\geq 6$ jets, 0 b-tagged jets

(e) No cut: $t\bar{t}$: 4,5 jets, $\geq 1$ b-tagged jet
(f) Signal Region: $t\bar{t}$: 4,5 jets, $\geq 1$ b-tagged jet

(g) No cut: $t\bar{t}$: $\geq 6$ jets, $\geq 1$ b-tagged jet
(h) Signal Region: $t\bar{t}$: $\geq 6$ jets, $\geq 1$ b-tagged jet

Figure 12: BDT scores from TMVA training with and without cuts. In the $t\bar{t}$ control regions the BDT scores seem to be better calibrated without cuts than with the cuts to the signal region.
6 Conclusion

By using TMVA to narrow 36 variables to 9 important and independent variables we were able to see that they effectively reflected the Monte Carlo simulations vs the data in the control regions. Small modifications in the $W+\text{jets}$ calculations need to be calculated, but this analysis can be re-done with the new $W+\text{jets}$ normalizations to re-evaluate if these nine important variables are usable for later analysis. After applying TMVA to background and the $b'/800$ samples it emerges that the expected limit by using TMVA is not very sensitive to JES changes, which is a promising start to potentially continue to use TMVA in the $B'B'$ analysis. The control plots of the TMVA BDT scores show that there is still analysis to be done with regards to making cuts on TMVA in different control regions. These control plots will also most likely improve when the newer $W+\text{jets}$ normalization is used. TMVA is an effective form of multivariate analysis to continue to pursue in the $B'B'$ analysis.

7 Acknowledgments

Thank you so much to the Columbia REU program for affording me such a wonderful opportunity and such valuable experiences. Thank you specifically to Jun Guo and Diedi Hu, with whom I worked on a daily basis, and to Gustaaf Brooijmans and Emily Thompson for helping me with various parts of my research. I thank John Parsons especially for extending this offer and making this program possible.
References


A Control Plots for 9 Variables

The nine final variables are as follows:

- $H_T$ of all objects (HT all)
- Number of jets (nJets)
- Number of Ws (wN)
- Missing transverse energy (met)
- Transverse Momentum of the lepton (pt lep)
- Energy of the leading b-tagged jet (Elbjets)
- $\Delta R$ between the lepton and the leading b jet (dR lepLbjet)
- Minimum $\Delta R$ between the lepton and a hadronic W (dR min lepWhad)
- Average $\Delta R$ between hadronic Ws (aveDr hadW)

A.1 Control Plots for the 9 variables in the $W+$jets 4,5 jets, 0 b-tag control region
A.2 Control Plots for the 9 variables in the $W + \text{jets} \geq \text{jets}$, 0 b-tag control region
A.3 Control Plots for the 9 variables in the $t\bar{t}$ 4,5 jets, $\geq 1$ b-tag control region
A.4  Control Plots for the 9 variables in the $t\bar{t} \geq 6$ jets, $\geq 1$ b-tag, $H_T<500$ GeV control region

- Hit total
- Number of Jets
- number Ws
- Missing ET
- Leptonic pt
- Energy Leading b jet
B DDT scores for B’800 and Background (Provided by Jun Guo)

**Expected limit using 9 bins**

- Best expected limit is with HT>1400 GeV:
  - No systematics: ~ 820 GeV
  - With JES: ~ 700 GeV

**Expected limit using 6 bins**

- Best expected limit is with HT>1400 GeV:
  - No systematics: ~ 825 GeV
  - With JES: ~ 775 GeV