TMVA study for B’B’ Search

Jun Guo\(^1\), Diedi Hu\(^1\), John Parsons\(^1\), Kerkira Stockton\(^2\)

\(^1\)Columbia University, \(^2\)Hamilton College
Outline

1. Motivation
2. Analysis
3. Cut-Based Selection
4. TMVA
5. Control Plots
6. Conclusion
Motivation

- 3 Generations of quarks
- Before the Higgs discoveries 4th generation chiral quarks were considered to be a natural extension to the Standard Model
- Results from the Higgs searches have led to fourth generation Vector-like Quarks (VLQ) to be more favored than chiral quarks

Image Source: [http://files.myopera.com/easteinstein/usercss/estakhr_model_Quark_electric_Charge.jpg](http://files.myopera.com/easteinstein/usercss/estakhr_model_Quark_electric_Charge.jpg)  
[http://www.particlezoo.net/individual_pages/info/quark_pack.jpg](http://www.particlezoo.net/individual_pages/info/quark_pack.jpg)
Motivation

• Pair production of $B$ can result in three decay modes: $Wt$, $Zb$, and $Hb$

• We are looking at the $Wt$ decay mode, resulting in a predominant decay of

$$d_4 \rightarrow Wt \rightarrow WWb$$

• With 4 $W$ bosons and 2 $b$-quarks we are specifically looking for 3 hadronic decays and 1 leptonic decay due to a low branching ratio of more than one leptonic decay.

• We are using chiral $b'$ for analysis optimization
Analysis

- Preselection Cuts

  - One lepton (Electron or Muon)
    - Electron: $|\eta_{\text{cluster}}| < 2.47$ excluding $1.37 < |\eta| < 1.52$, $E_T > 25$ GeV
  
  - minimum $E_T$ miss value
  
  - Daughter jets from $W$ decay have a $\Delta R < 1$ with a dijet or single jet (if jets overlap) mass consistent with the $W$ boson mass
  
  - Triangle Cut to reduce QCD
    - electron channel: $E_T$ miss $> 30$ GeV and $W$ $m_T > 30$ GeV
  
  - GRL, Primary Vertex with at least 5 tracks

- Signal Region

  - Number of Jets $\geq 6$
  
  - Number of b-tagged jets $\geq 1$
  
  - $H_T \geq 800$ GeV

- Focusing on the electron channel
Analysis - Samples

- 2012 Data - Integrated Luminosity of 20.3 fb^{-1}
- B'800 Monte Carlo
  - Generated with Atlfast (AF-II)
- ttbar (dominant background)
  - Alpgen with up to 3 additional partons
- W+Jets
  - Alpgen with up to 5 light partons
  - Normalized using a scaling factors derived from data
- QCD
  - From Data: Matrix method
- ttbar + V (other)
  - Madgraph and Pythia
- Diboson Production (other)
  - Herwig
- Z+ Jets (Where a lepton is missed) (other)
  - Alpgen with up to 5 light partons
- Single Top (other)
  - MC@NLO and AcerMC
Cut-based Selection

- Monte Carlo seen with cuts for the Signal Region applied
- $H_T > 800$ GeV
- Needs tighter cuts in order to see the $b'800$ signal

- $H_T > 1000$ GeV
- Still quite difficult to see the signal, though background is diminishing
TMVA - Boosted Decision Trees

- Decision trees use cuts to determine whether an event is considered background or signal.
- Boosted Decision Trees are Decision Trees that are trained hundreds of times in order to give each event a weight based on whether it was correctly or incorrectly classified as background or signal.
- Trained using Monte Carlo background and signal samples.
- These weights are then applied to all of the scores (-1 or 1) that the event received and it is then given a score between -1 and 1, where -1 is fully background and 1 is fully signal.
TMVA - Procedure

- Uses Boosted Decision Trees to determine a good value for a cut on each variable between background and signal Monte Carlo samples
- Builds a list of the “most important” variables which can be used to limit the number of considered variables
- The correlation plots are also used to limit which variables are considered
- Trained using B’800 for the signal sample and ttbar and W+Jets for background
Example of a Correlation plot with all 36 variables. Used to narrow down the total number of variables.
### Final 9 Variables

#### Cuts

- $H_T > 500$ GeV
- $\geq 1$ $W$
- $\geq 1$ b-tagged Jet
- $\geq 6$ Jets

#### Input Variables and Plots

1. **Input variable: $H_T_{allobj}$**
   - $H_T_{allobj}$ vs. Signal and Background

2. **Input variable: $met$**
   - met vs. Signal and Background

3. **Input variable: $aveDr_{hadW}$**
   - $aveDr_{hadW}$ vs. Signal and Background

4. **Input variable: $pt_{lep}$**
   - $pt_{lep}$ vs. Signal and Background

5. **Input variable: $wN$**
   - $wN$ vs. Signal and Background

6. **Input variable: $jetN$**
   - $jetN$ vs. Signal and Background

7. **Input variable: $E_{lbjet}$**
   - $E_{lbjet}$ vs. Signal and Background

8. **Input variable: $dR_{min_{lepWhad}}$**
   - $dR_{min_{lepWhad}}$ vs. Signal and Background

9. **Input variable: $dR_{lepLbjet}$**
   - $dR_{lepLbjet}$ vs. Signal and Background
Test and Training samples are relatively on par

- Shows that TMVA was not overtrained

- Very good background rejection
- Shows TMVA could be a useful tool
- Variables are very highly uncorrelated, showing high independence
- This is useful for finding variables that can collectively help differentiate between background and signal
Control Plots

- Control Regions
- Plots of the 9 “important” TMVA variables plus number of b tagged jets
- All plots have certain pre-cuts
  - electron preselection, jet selection, triangle cut, ≥2 jets, HT>250 GeV
- 2 sets of control cuts
  - W+Jets control Region
    - 4,5 Jets with 0 b-tagged Jets
    - 6 Jets with 0 b-tagged Jets
  - TTbar control Region
    - 4,5 Jets with ≥1 b-tagged Jet
    - ≥ 6 Jets with ≥1 b-tagged Jet with Ht<500
Normalization is fairly spot-on but we can see when looking at the Number of Jets plot that Monte Carlo is slightly underestimated.
• We can see the same phenomena here, where the average point on the background/signal ratio plot is >1.
• We see here that at in the lower ranges Monte Carlo is consistent with data, but still slightly underestimated.
Minimum $\Delta R$ between the lepton and hadronic $W$ is far less consistent than in the 4, 5 Jets control region, but this is likely due to the statistical distribution of Monte Carlo samples. It also does not show the underestimation of Monte Carlo as dramatically as the other plots.

Number of $W$s shows, again, the slight underestimation of Monte Carlo
The 4,5 Jets $t\bar{t}$bar control region seems to be better normalized than the $W$+Jets regions, but they are still slightly underestimated.
Motivation

Analysis

Cut-Based Selection

TMVA

Control - ttbar

Conclusion

4,5 Jets, \geq 1 b-tagged Jet

We can see here that the 4,5 TTbar control region is far better normalized than the W+Jets control region, implying that the W+Jets normalization is slightly off.
The $\geq 6$ Jets ttbar control region seems to be better normalized than the W + Jets regions, but they are still slightly underestimated.
Even though the $\geq 6$ Jets $\bar{t}t$ control region seems not quite as consistent as the previous control region, $\Delta R$ between the leptonic and leading b-tagged jet seems to be fairly one-to-one, even at levels of only around 100 events.

We can see at even very low numbers of events that data and background are fairly consistent (taking in the possibility of statistical difference) in the $\Delta R$ between the lepton and Hadronic W plot.
BDT Scores for Background and B’800

- Expected limit set on scale factor on x-section for b’800
  - No systematics: 0.759 (<1 for exclusion)
  - With JES: 1.03, can almost exclude b’ 800 with e channel alone
- Compare with cut-based selection:
  Lower HT cut, currently using HT>500 GeV
  Much less sensitive to JES: compare limit factor for b’800

<table>
<thead>
<tr>
<th></th>
<th>No Systematics</th>
<th>With JES</th>
<th>Ratio of with-JES/no-systematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-bin (e and $\mu$ channels)</td>
<td>0.868</td>
<td>1.81</td>
<td>2.09</td>
</tr>
<tr>
<td>6-bin (e and $\mu$ channels)</td>
<td>0.832</td>
<td>1.15</td>
<td>1.39</td>
</tr>
<tr>
<td>TMVA (only e channel)</td>
<td>0.759</td>
<td>1.03</td>
<td>1.36</td>
</tr>
</tbody>
</table>
W+Jets
Control Region

4,5 Jets, 0 b-tagged Jets

≥ 6 Jets, 0 b-tagged Jets
<table>
<thead>
<tr>
<th>TTbar Control Region</th>
</tr>
</thead>
</table>

**Motivation**

- **No Cuts (Training)**
  - BDT scores
  - Events

- **Signal Region Cuts (Training)**
  - BDT scores
  - Events

- **TTbar Control Region**
  - No Cuts (Training)
  - Signal Region Cuts (Training)

**Analysis**

- **4,5 Jets, ≥ 1 b-tagged Jet**
  - ≥6 Jets, ≥ 1 b-tagged Jet, Ht < 500
Conclusion - Results

• **TMVA**
  - Using correlation plots and TMVA variable ranking narrowed 36 variables down to 9 important and independent variables
  - These variables are used to distinguish background from signal with the trained BDTs
  - Control plots based on TMVA weights would be a good indicator of weight consistency from Monte Carlo to Data
  - The expected limit by using TMVA is not very sensitive to JES changes compared to other methods, which is promising

• **Control Plots**
  - Control plots are important indicators of whether the TMVA-chosen variables are effective measures
  - Most plots seem fairly consistent, but the TTbar control regions seem to be more consistent than the W+Jets regions
  - Could lead to reevaluation of W+Jets normalization process
Sources


Thank you very much to Jun, Diedi, Emily, Gustaaf, and John!
Backup Slides
W+Jets: 4, 5 Jets, 0 btag

ave dR had W

dR leptonic and Lb jet

Missing ET
W+Jets: 4, 5 Jets, 0 btag

Energy Leading b jet

- 2012 data
- ttbar
- W + jets
- QCD
- Other bkgds

Events

10^5

0 100 200 300 400 500 600 700 800 900 1000

E leading b let (MeV)

Data / bkg

1.4

1.2

1.0

0.8

0.6

number b tagged Jets

- 2012 data
- ttbar
- W + jets
- QCD
- Other bkgds

Events

10^5

0 1 2 3 4 5 6

n h. jets

Data / bkg

1.4

1.2

1.0

0.8

0.6
W+Jets: $\geq 6$ Jets, 0 btag

ave dR had W

dR leptonic and Lb jet

number b tagged Jets
$W+\text{Jets: } \geq 6 \text{ Jets, 0 btag}$
**TTbar: 4,5 Jets, \geq 1 btag**

### ave dR had W

- **Energy Leading b jet**
- **min dR lepW had**

---

### dR lepton's and Lb jet

- **Events**
  - 2012 data
  - t\bar{t}bar
  - W + jets
  - QCD
  - Other bkgds

---

### Graphs

- **Events vs. dR (ave. dR had W)**
- **Events vs. dR (min dR lepW had)**
- **Energy Leading b jet**
- **min dR lepW had**

---

**33**
TTbar: $\geq 6$ Jets, $\geq 1$ btag, Ht<500 GeV

Ht total

Energy Leading b jet

number b tagged Jets
TTbar: ≥6 Jets, ≥1 btag, Ht<500 GeV
**Expected limit using 9 bins**

- **9 bins:**
  - 6 jets: 0W, 1W, >=2W’s
  - 7 jets: 0W, 1W, >=2W’s
  - >=8 jets: 0W, 1W, >=2W’s

- **Best expected limit is with HT>1400 GeV:**
  - No systematics: ~ 820 GeV
  - With JES: ~ 700 GeV
**Expected limit using 6 bins**

- 6 bins:
  - 6 jets: 2W’s, >=3W’s
  - 7 jets: 2W’s, >=3W’s
  - >=8 jets: 2W’s, >=3W’s

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