Advanced Analysis Techniques in the Search for Production of a Higgs boson in association with Top Quarks at CMS

Jason Slaunwhite
On behalf of the CMS collaboration
One of the biggest questions remaining in the standard model:

- Why do the electron and the top quark have such different masses?
- Top-Higgs coupling measurement is an important step in
- Accessible via $ttH$ production

\[ M_{\text{electron}} = 0.5 \text{ MeV} \]

\[ \text{Top Quark } M = 3 \times 10^5 M_{\text{electron}} \]
In this talk, we will see that TTH production is a challenging measurement because:
- Signal production rate is small compared to backgrounds
- Uncertainties are large
- No single variable gives great discrimination

We can overcome these issues using multivariate analysis techniques:
- To identify the objects associated with ttH decay with high efficiency and purity
- To distinguish ttH events from background
**Signal Process**

- **Production:** $t\bar{t}H$
- **Cross section:** 130 fb at $M=125$ GeV and 8 TeV
- **Focus on**
  - $H$ to $b\bar{b}$ (largest BR, 58%)
    - $\sigma \times \text{BR}(H \to b\bar{b}) = 75$ fb
- **Final state:**
  - $WWb\bar{b}bb\bar{b}$
- **We require $\geq 1$ W to e,\(\mu\)**
  - 1 lepton and up to 6 jets.
    - 4 jets come from $b$-quarks.
  - 2 leptons and up to 4 jets.
    - All 4 jets come from $b$-quarks.

J. Slaunwhite
Background Processes at 8 TeV

**WWbbbb: tt+bb**
- ~2-4 pb
- irreducible, ~24x larger than signal \( \sigma \times \text{BR}(H \to bb) \)

**WWbb+>=0jets: tt+jets**
- 234 pb
- fewer jets/ fewer tags, ~3000x larger than signal

**Single top, Dibson, W/Z+jets**
- Many fewer jets and tags
- Classify events according to jets and tags
**Object Definitions**

### Electrons from W
- **Tight**
  - $p_T > 30$ GeV
  - $\eta < 2.5$
  - Tight Isolation
  - **MVA ID**
- **Loose (main differences)**
  - $p_T > 15$ GeV
  - Loose Isolation

### Muons from W
- **Tight**
  - $p_T > 30$ GeV
  - $\eta < 2.1$
  - Tight Isolation
  - Tight ID
- **Loose (main differences)**
  - $p_T > 10$ GeV
  - Loose ID & Isolation

### Jets from W, t, H
- Anti-$k_T$ size 0.5
- $p_T > 40$ for jets 1,2,3
- $p_T > 30$ each other jet
- Loose ID requirements

### B-jets
- Pass all jet requirements
- **Combined Secondary Vertex**
  - (Medium operating point)
Ele performance compare

- MVA: Implemented with a Boosted decision tree
  - Trained for real vs fake electrons
- Ele MVA ID uses:
  - Tracking variables
  - Shower-shape variables
  - Geometric matching between track and calorimeter
  - Energy matching between track and calorimeter
- Has better efficiency for the same electron fake rejection

From DP-13-003
B-jets can be distinguished from other kinds of jets by looking for the decay of long-lived b-hadrons
- Vertexing
- Track impact parameter

Combined Secondary Vertex (CSV) uses both
- Overcomes vertexing efficiency

For the medium working point
- Efficiency: 65% per jet
- Fake rate: 1-1.5% per jet
  (tt+jets is 3000x larger than ttH)
- For the same fake rate, a tagger using vertex-only information would have 55% efficiency
**Event categorization**

- Background has fewer jets and tags, so classify events by num jets, and num tags
- Use all 9 categories in simultaneous fit

### S/B Ratio - 1 tight lepton

<table>
<thead>
<tr>
<th></th>
<th>4jets</th>
<th>5jets</th>
<th>&gt;=6jets</th>
</tr>
</thead>
<tbody>
<tr>
<td>2tags</td>
<td>x</td>
<td>x</td>
<td>0.0031</td>
</tr>
<tr>
<td>3tags</td>
<td>0.0027</td>
<td>0.0063</td>
<td>0.011</td>
</tr>
<tr>
<td>&gt;=4tags</td>
<td>0.028</td>
<td>0.037</td>
<td>0.040</td>
</tr>
</tbody>
</table>

### S/B Ratio - 2 lepton

<table>
<thead>
<tr>
<th></th>
<th>2jets</th>
<th>&gt;=3jets</th>
</tr>
</thead>
<tbody>
<tr>
<td>2tags</td>
<td>0.0001</td>
<td>x</td>
</tr>
<tr>
<td>&gt;=3tags</td>
<td>x</td>
<td>0.015</td>
</tr>
</tbody>
</table>

J. Slaunwhite 8 TeV

8 TeV
The uncertainties that have the greatest effect on the analysis are the ones that affect the number of jets/tags:

- Jet energy Scale, btag SF, mistag SF, madgraph scale

The analysis is also sensitive to the amount of irreducible background:

- Overall rate uncertainties in our prediction

These are nuisance parameters in our fit.

<table>
<thead>
<tr>
<th>Uncertainty</th>
<th>Max Rate Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jet Energy Scale</td>
<td>60%</td>
</tr>
<tr>
<td>tt+bb ONLY (theory)</td>
<td>50%</td>
</tr>
<tr>
<td>Btag SF</td>
<td>34%</td>
</tr>
<tr>
<td>Mistag SF</td>
<td>24%</td>
</tr>
<tr>
<td>Madgraph Scale</td>
<td>20%</td>
</tr>
<tr>
<td>Theory xsecs, Lumi, lepton efficiencies, etc</td>
<td>~15%</td>
</tr>
</tbody>
</table>

Signal size: ~4% of background
Yield Summary: 1 Lepton Events

Yields agree overall
Majority of background is $t\bar{t}$+light 65% - 90% of all events
Signal Extraction Strategy

- Yield in $\geq 6$ jets $\geq 4$ tags
  - 2.5 Signal on background of 63 $\pm$ 21
  - Counting experiment will not be very sensitive
- Improve sensitivity by simultaneously fitting discriminating distributions in all categories
  - Treat uncertainties as nuisance parameters in the fit
- Start by establishing a baseline using one kinematic variable in each category
- Then measure impact of combining multiple variables with an MVA technique
Initially expect the Higgs mass resonance to provide distinguishing power

- This is where discovery modes $H \rightarrow ZZ$ and $H \rightarrow \gamma\gamma$ get their power
- For $ttH$, mass is not so powerful
  - Helps somewhat in 6 jets 4 tags, but it is not the most sensitive

Reasons:
- $b$-jet energy resolution worse than photon/e/\(\mu\) energy resolution
- Combinatorics of $b$-jets in final state can wash out resonance

**Figure**

This figure shows the breakdown of jet1to1parton assignments for the two jets with the minimum $R$ separation in the event for events with lepton+$\geq$4jets+$\geq$4tags. The top left plot shows events with $b$1tagged jets, the top right plot shows events with $b$1tagged jets, and the bottom middle plot shows events with $b$1tagged jets.

**Data and MC Samples**

2.1 Data Samples

The results presented here are based on the full 400 CMS dataset. Table lists the datasets used for this analysis based on the triggers used to collect the data. Luminosities are quoted from the pixel luminosity calculation, including the effects of any trigger prescales, and have a 2.5% uncertainty.

2.2 Signal Samples

The $t\bar{t}H$ signal is modeled using the PYTHIA generator, generated privately using the same conditions and configuration as the "Fall00" MC campaign. The samples and associated cross sections used are listed in Table.

2.3 Background Samples

To model the backgrounds, this analysis primarily uses Monte Carlo (MC) samples from the "Fall00" MC campaign, except where noted in the table below. Most of the samples are generated either with the MADGRAPH tree-level matrix element generator matched to PYTHIA for the parton shower, or with the NLO generator POWHEG combined with PYTHIA. These samples are reconstructed with the same CMSSW version as the data samples listed above. The pileup distribution in all MC samples is reweighted, using the procedure listed below so that the MC pileup distribution matches the one expected for data. Table lists the background MC samples and associated cross sections.

<table>
<thead>
<tr>
<th>Lepton</th>
<th>Jets</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥4</td>
<td>≥4</td>
<td></td>
</tr>
</tbody>
</table>

**S/B ~ 3/100**

No mass peak visible on top of combinatoric background
Performance with Best Variable

- For most categories, the average CSV value for tagged jets is the best discriminant
  - Helps reject largest background: tt + light flavor
- Fit best single variable in each category and extract upper limit on xsec
  - 6.6x SM expectation
  - “If cross section was more than 6.6 times what we expect, then we would have seen it with this measurement”
We use Artificial Neural Networks (ANN) to combine multiple variables into a single discriminant.

- Multi-layer perceptron as implemented in ROOT and TMVA

Create one ANN per category with own set of input variables.

Structure: N inputs, 2 hidden layers, one output

- Hidden layer 1: N nodes
- Hidden layer 2: N-1 nodes

Training

- 50% Signal = ttH, M(H)=120
- 50% Background = tt
- Reserved testing sample for overtraining check

Categories of variables

- Kinematics of objects, single and composite
- Kinematics of jet pairs
- Event shape
- Btag CSV discriminant
**Example ANN: One lepton 6 jets and 4 tags**

11 input variables in total

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass (lep, MET, Jets)</td>
<td>Kin. of composite obj</td>
</tr>
<tr>
<td>Mass (j,j) closest jets</td>
<td>Jet pairs</td>
</tr>
<tr>
<td>Mass (j,j) best</td>
<td>Jet pairs</td>
</tr>
<tr>
<td>Average ΔR(tag, tag)</td>
<td>Jet pairs</td>
</tr>
<tr>
<td>Minimum ΔR(lep, jet)</td>
<td>Shape</td>
</tr>
<tr>
<td>Sphericity</td>
<td>Shape</td>
</tr>
<tr>
<td>H2</td>
<td>Shape</td>
</tr>
<tr>
<td>H3</td>
<td>Shape</td>
</tr>
<tr>
<td><strong>Average CSV</strong></td>
<td><strong>Btag</strong></td>
</tr>
<tr>
<td>2nd-highest CSV</td>
<td>Btag</td>
</tr>
<tr>
<td>lowest CSV</td>
<td>Btag</td>
</tr>
</tbody>
</table>

J. Slaunwhite
Limit Results

- Fit NN output distribution simultaneously in all 9 categories to extract overall limit
- **5.2xSM expectation at M=125**
- 27% improvement over single variable
- Equivalent to increasing data collected by 60%
  - Effectively 3/fb additional in this dataset
  - Effectively 12/fb on full dataset
  - Worth half a year of data taking

Expected @ 125: 5.2xSM
Observed @ 125: 5.8xSM
**Summary**

- Mass hierarchy is a compelling problem that can be explored through $ttH$
- Challenging: $ttH$ cross section is small compared to the backgrounds, the uncertainties are large, and the mass resonance is not especially powerful
- Multivariate techniques help us overcome some of these challenges by optimizing:
  - Object identification (b-tags, electrons)
  - Signal discrimination
- The optimizations help us get more performance out of the data we collected
BACKUPS
Figure 6: Performance curves obtained from simulation for the algorithms described in the text. (a) light-parton- and (b) c-jet misidentification probabilities as a function of the b-jet efficiency.

4.5 Impact of running conditions on b-jet identification

All tagging algorithms rely on a high track identification efficiency and a reliable estimation of the track parameters and their uncertainties. These are both potentially sensitive to changes in the running conditions of the experiment. The robustness of the algorithms with respect to the misalignment of the tracking system and an increase in the density of tracks due to pile up, which are the most important of the changes in conditions, has been studied.

The alignment of the CMS tracker is performed using a mixture of tracks from cosmic rays and minimum bias collisions [35, 36], and is regularly monitored. During the 2011 data taking, the most significant movements were between the two halves of the pixel barrel detector, where discrete changes in the relative $z$ position of up to 30 $\mu$m were observed. The sensitivity of b-jet identification to misalignment was studied on simulated $t\bar{t}$ samples. With the current estimated accuracy of the positions of the active elements, no significant deterioration is observed with respect to a perfectly aligned detector. The effect of displacements between the two parts of the pixel barrel detector was studied by introducing artificial separations of 40, 80, 120, and 160 $\mu$m in the detector simulation. The movements observed in 2011 were not found to cause any significant degradation of the performance.

Because of the luminosity profile of the 2011 data, the number of proton collisions taking place simultaneously in one bunch crossing was of the order of 5 to 20 depending on the time period. Although these additional collisions increase the total number of tracks in the event, the track selection is able to reject tracks from nearby primary vertices. The multiplicity distribution of selected tracks is almost independent of the number of primary vertices, as shown in Fig. 7 (a). There is an indication of a slightly lower tracking efficiency in events with high pileup. The rejection of the additional tracks is mainly due to the requirement on the distance of the tracks with respect to the jet axis. This selection criterion is very efficient for the rejection of tracks from pileup. The reconstruction of track parameters is hardly affected. The distribution of the second-highest IP significance is stable, as shown in Fig. 7 (b). The impact of high pileup on the b-jet tagging performance is illustrated in Fig. 8. This shows the light-parton misidentification probability versus the b-jet tagging efficiency for the TCHP and SSVHP algorithms.
Electron MVA

From DP-13-003

Cut based Medium ID

MVA ID

From DP-13-003
Table 4: The ANN inputs for the nine jet-tag categories in the 8 TeV \( t\bar{t}H \) analysis in the lepton+jets and dilepton channels. The choice of inputs is optimized for each category. Definitions of the variables are given in the text. The best input variable for each jet-tag category is denoted by ⭐.

<table>
<thead>
<tr>
<th></th>
<th>Lepton+Jets</th>
<th>Dilepton</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Jets</strong></td>
<td>≥6 4 5 ≥6 4 ≥6</td>
<td>2 ≥3</td>
</tr>
<tr>
<td><strong>Tags</strong></td>
<td>2 3 3 4 4 4</td>
<td>≥4</td>
</tr>
<tr>
<td>Jet 1 ( p_T )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Jet 2 ( p_T )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Jet 3 ( p_T )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Jet 4 ( p_T )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( N_{jets} )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( p_T(\ell, E_{T}^{miss}, \text{jets}) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( M(\ell, E_{T}^{miss}, \text{jets}) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Average ( M((j_{m}^{untag}, j_{n}^{untag})) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( M((j_{m}^{tag}, j_{n}^{tag})_{closest}) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( M((j_{m}^{tag}, j_{n}^{tag})_{best}) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Average ( \Delta R(j_{m}^{tag}, j_{n}^{tag}) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Minimum ( \Delta R(j_{m}^{tag}, j_{n}^{tag}) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( \Delta R(\ell, j_{closest}) )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Sphericity</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Aplanarity</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( H_0 )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( H_1 )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( H_2 )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( H_3 )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( \mu^{CSV} )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>( (\sigma^{CSV}_n)^2 )</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Highest CSV value</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>2( ^{nd} )-highest CSV value</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
<tr>
<td>Lowest CSV value</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
<td>✔ ✔ ✔ ✔ ✔ ✔</td>
</tr>
</tbody>
</table>
Significance

\[ \langle S^2 \rangle = \frac{1}{2} \int \frac{(\hat{y}_S(y) - \hat{y}_B(y))^2}{\hat{y}_S(y) + \hat{y}_B(y)} dy. \]