Evidence for $WW/WZ \rightarrow l\nu + \text{Heavy Flavor Jets}$
at CDF

Federico Sforza
on behalf of the CDF Collaboration

INFN & Università di Pisa

DPF 2011 – 12 August 2011
Evidence for $WW/WZ \rightarrow l\nu + HF$: Outline

1. Analysis Motivation
2. Event Selection
3. Background Composition
4. Results

Diboson/Higgs production in the $l\nu +$ Heavy Flavor (HF) Jets Channel.
Dibosons and Higgs $\rightarrow b\bar{b}$

$H \rightarrow b\bar{b}$:
- dominates low mass Higgs exclusion limits;
- decay mode still difficult at LHC;
- test of the SM couplings;
- most important production mode: $WH$

$WZ \rightarrow l\nu + b\bar{b}$ has same final state as $WH$:
- $\Rightarrow$ larger yield (but still rare, $\sigma_{WZ} \approx 3\text{ pb}$)
- $\Rightarrow$ different resonance mass.

SM processes with Heavy Flavor to be observed before a low mass Higgs.
Analysis Strategy: $WW/WZ \rightarrow l\nu HF$

**Goal:** Observe a rare SM resonance in the $l\nu + HF$ Channel:

- collect $W \rightarrow l\nu$ candidate with max efficiency:
  - several trigger strategies;
  - release lepton identification requirements.

- Di-jet Heavy Flavor enriched sample selection:
  - require two high energy jets;
  - $b-$hadron identification: secondary decay vertices.

- Signal discrimination based on *di-jet mass.*
The CDF II Experiment

Multipurpose detector:

- **Integrated tracking:**
  - silicon detector;
  - wire chamber (COT).

- **Calorimeter system:**
  - ElectroMagnetic; (EM)
  - HADronic (HAD)

- **Muon detection:**
  - outer muon chambers.

- First data was in 2002.

- This analysis uses: $\int L = 7.5 \text{ fb}^{-1}$ (March 2011).
$W \rightarrow l\nu$: Multiple Trigger Strategies

- Need to collect *all* the interesting events:
  ⇒ more trigger strategies to increase the acceptance.

**Strategy I:**
- single object triggers
- high energy lepton: electron or muon.

**Strategy II:**
- multiple object triggers
- Missing Energy + Jet.
$W \rightarrow l\nu$: Offline Selection

**Neutrino:** imbalance in the total transverse energy: $E_T > 20$ GeV.

**Tight Leptons:** standard set of CDF leptons ($e, \mu$) cuts:

**Electrons:**
- Energy cluster in EM calorimeter ($E_T > 20$ GeV)
- Energy deposit isolated in the calorimeter.
- Good quality track pointing to the cluster.

**Muons:**
- Hits in the muon chambers.
- Track pointing to the hits ($P_T > 20$ GeV/c).
- MIP energy deposit isolated in the calorimeter.

**Extended $\mu$ Category (EMC):** 30% acceptance gain:

**Loose Muons:**
- Few or missing hits in the muon chambers.
- MIP energy deposit isolated in the calorimeter.
- Track pointing to the MIP ($P_T > 20$ GeV/c).

**Isolated Tracks:**
- Good quality track ($P_T > 20$ GeV/c).
- No other track within $\Delta R = 0.4$. 
Heavy Flavor Jets Identification

High $P_T$ di-jet sample selection:
- CDF *standard* jet identification: cone algorithm ($\text{JETCLU04}$, $\Delta R < 0.4$);
- two central energy clusters: $|\eta| < 2.0$, $E_T > 20$ GeV (corrected for detector effects).

SecVtx algorithm

Heavy Flavor hadrons identification:
- "b-tag": secondary decay vertex.
- $b$-tag efficiency: $\sim 40\%$;
- $c$-tag: efficiency $\sim 6\%$;
- fake tags $\sim 1\%$ (*background*).

3 selection regions

1. **Pretag**: no SecVtx request, control region.
2. **Single Tag**: exactly 1 SecVtx tagged jet, signal region.
3. **Double Tag**: exactly 2 SecVtx tagged jet, signal region.
**WW vs WZ Selection**

Interested in $WZ \rightarrow l\nu b\bar{b}$ but not sensitive yet...

- Must use more advanced techniques (some already in use in CDF WH analysis):
  - improve $b-$tag efficiency
  - use multivariate discriminants;
  - ....

- Ongoing effort to obtain $3\sigma$ significance together with D0 for WZ alone.

Di-jet invariant mass resolution does not allow to distinguish WW from WZ.

- $WW \rightarrow l\nu + cs$ contributes to Heavy Flavor signal:
  - $c-$quark produces long living hadrons (6% tagging efficiency);
  - higher cross section for $WW$ associate production than $WZ$;
  - $W \rightarrow cs$ branching ratio $\approx 30\%$;
  - Large acceptance w.r.t $WZ$: both $W$ can decaying to $c-$quarks or to leptons.

We can probe high–$P_T$ behavior of $c-$quarks too!
Background Composition

Signal topology: exclusive 1 lepton + $\not{E}_T$ + 2jets (1 or 2 tags)

background estimate: Monte Carlo simulation & data control samples

⇒ 4 background components:

1. **MC based**: $t\bar{t}$, single top, $Z$ + jets ⇒ estimated from MC.
2. **QCD**: lepton and $\not{E}_T$ are faked by mis-reconstructed jets. ⇒ measured in data.
3. **W+HF**: Heavy Flavors ⇒ major background, large uncertainty.
   - Total $W$ + jets Normalization obtained from data (pretag control region);
   - Heavy Flavor Fractions $= \frac{W+HF}{W+jets}$ estimated from MC.
4. **W+LF**: Light Flavors, $W$ + fake tags ⇒ parametrized on jet data.
QCD Modeling and Rejection

- Fake $W$ models obtained reversing some of the lepton identification cuts.
- Multivariate tool (based Support Vector Machine) drastically removes QCD.
- Normalization extracted from data: $QCD$, $W + jets$ two component fit on $E_T$.

Tight electron QCD estimate before and after QCD rejection:

EMC leptons QCD estimate before and after QCD rejection:
Background Composition

\section*{W + Heavy Flavor Background}

\begin{itemize}
  \item **W + b\bar{b}, W + c\bar{c}, W + c** estimate
    \begin{itemize}
      \item Large theoretical uncertainty on $\sigma_{W+\text{jets}}$.
      \item Normalization ($N_{W+\text{jets}}$) from the \textit{pretag} data sample:
        \[ N_{W+\text{jets}} = N_{\text{Data pretag}} (1 - F_{\text{QCD}}) - N_{\text{MC}} \]
      \item Heavy Flavor Fractions: $f_{\text{HF}} = \frac{W+\text{HF}}{W+\text{jets}}$ derived from MC (Alpgen, LO)
        \Rightarrow \text{Technically difficult} \approx 90 \text{ exclusive MCs used.}
      \item Final normalization: $N_{\text{HF}} = N_{W+\text{jets}} \times f_{\text{HF}} \times K$
      \item $K = 1.4 \pm 0.4$, correction factor to HF production rate in MC:
        \Rightarrow \text{derived from } W + 1 \text{ jet control sample.}
    \end{itemize}
\end{itemize}
Understanding the Sample: Pretag

- Control region;
- Very large statistic;
- Dominated by $W + LF$

**Background Composition**

12 August 2011 13 / 21
Understanding the Sample: Single Tag

- Signal region;
- Large statistic;
- $W + c$, $W + b\bar{b}$, $W + LF$, top

**Background Composition**

F. Sforza (INFN & Università di Pisa)

12 August 2011
Understanding the Sample: Double Tag

- Signal region;
- Low statistic;
- Dominated by $W + b\bar{b}$ and top

- **TIGHT SVT**
  - CDF Run II Preliminary ($7.5fb^{-1}$)

  - **W Transverse Mass**
  - Number of events

  - **Δ R(jet1-jet2)**
  - Number of events

- **EMC SVT**
  - CDF Run II Preliminary ($7.5fb^{-1}$)

  - **W Transverse Mass**
  - Number of events

  - **Δ R(jet1-jet2)**
  - Number of events
RESULTS
Di-jet Mass Spectrum: 4 Channels

**Single Tag (Tight + EMC):** sizable signal excess over $W$+jets mass spectrum.

**Double Tag (Tight + EMC):** $WZ$ main signal contribution: still statistically limited.
Systematics Inclusion and Fitting Procedure

- Shape and rate systematics are considered in the measurement:

<table>
<thead>
<tr>
<th>(some of the) systematics</th>
<th>rate (priors)</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigger</td>
<td>1-3%</td>
<td>no</td>
</tr>
<tr>
<td>Lepton ID</td>
<td>1-4%</td>
<td>no</td>
</tr>
<tr>
<td>Luminosity</td>
<td>6%</td>
<td>no</td>
</tr>
<tr>
<td>$b$–tag SF</td>
<td>5-10%</td>
<td>no</td>
</tr>
<tr>
<td>QCD Normalization</td>
<td>40%</td>
<td>no</td>
</tr>
<tr>
<td>JES</td>
<td>5-10%</td>
<td>yes</td>
</tr>
<tr>
<td>Alpgen $Q^2$</td>
<td>3-6%</td>
<td>yes</td>
</tr>
<tr>
<td>HF K factor</td>
<td>30%</td>
<td>no</td>
</tr>
</tbody>
</table>

- Some systematics are large due to extrapolation from control regions:
  - K factor: 30%, QCD: 40%;

- Shape systematics can shift background and signal peak regions.

- A maximum binned likelihood fit is performed to enhance the sensitivity.
  - Prior uncertainties constrained by data.
Significance of the Observation

Significance:

1. Define a test statistics (Likelihood ratio):
   
   $-2\ln Q = -2\ln \frac{p(data|H_{\text{Test}}, \hat{\theta})}{p(data|H_{\text{Null}}, \hat{\theta})}$

   $\theta$: best fit nuisance parameters

2. Throw background only ($H_{\text{Null}}$ distribution) pseudo-experiments.

3. Compare data $-2\ln Q$ against pseudo-experiments $-2\ln Q$.

   - Expected $-2\ln Q$: $-8.85^{+6.2}_{-6.4}$ (Test hypothesis)
     
     $p-value = 0.00125 \Rightarrow 3.02\sigma$

   - Observed $-2\ln Q$: -9.0446.
     
     $p-value = 0.00120 \Rightarrow 3.03\sigma$
Cross section measurement:

- Central value ⇒ maximum of the Bayesian posterior.
- Error ⇒ minimum interval covering 68% of the Bayesian posterior.

CDF Run II Preliminary (7.5 fb⁻¹)

$\sigma_{(WW/WZ \rightarrow l_1\nu+HF)} = 1.08^{+0.26}_{-0.40} \times SM$
Conclusions

Analysis conclusions:

- obtained $3\sigma$ evidence of $WW/WZ$ production in the tagged, heavy flavor enriched channel using di-jet mass as signal discriminant;
- measured cross section in good agreement with SM and other diboson channels measurements:

$$\sigma(\text{WW}/WZ \rightarrow l\nu bb/l\nu cs) = 1.085^{+0.26}_{-0.40} \times SM;$$

- strong support to the CDF $WH$ analysis strategy and results;

- public material available at:

  http://www-cdf.fnal.gov/physics/new/hdg/Results_files/results/wzlnubb_071911/
Back Up Slides
Looking for the Rarest Processes

SM diboson production cross sections at the TeVatron:

Extremely rare processes:

\[ \sigma_{p\bar{p}\rightarrow W^+ W^-} = 11.34 \pm 0.7 \text{ pb} \]
\[ \sigma_{p\bar{p}\rightarrow W^\pm Z^0} = 3.22 \pm 0.2 \text{ pb} \]
\[ \sigma_{p\bar{p}\rightarrow Z^0 Z^0} = 1.2 \pm 0.07 \text{ pb} \]

(arXiv:1107.5518)

TeVatron experiments already observed all the diboson production modes!

F. Sforza (INFN & Università di Pisa)
Diboson Observations at CDF

Observation in both leptonic and hadronic mode:

- **ZZ → llll**
  - four leptons final state;
  - clean signature;
  - CDF Public Note 9910 (2009).

- **WW/WZ/ZZ → νν + 2Jets**
  - invisible energy + hadronic final state
  - PRL 103, 091803 (2009)

- **WW/WZ → lν + 2Jets**
  - semileptonic decay + hadronic final state
  - PRL 104, 101801 (2010)
Next Challenge: Lepton + Neutrino + Heavy Flavor

Dibosons have a variety of final states:

**WW**

**WZ**

Experimental challenges:

- Selection of long living heavy hadrons in jets:
  - identification of *secondary* decay vertex $\Rightarrow$ low efficiency.

- Reconstruct mass resonances in Heavy Flavor enriched final states.
  - hadronic environment $\Rightarrow$ low energy resolution;
  - background dominated by large irreducible components.
Triple Gauge Couplings

t-channel:

s-channel:

- s-channel sensible to Triple Gauge Couplings (TGC);
- Cross section can be enhanced by New Physics.
Selection Overview

- High energy lepton ($E_T(P_T) > 20$ GeV(Gev/c)). *Maximize acceptance:*
  - **TIGHT:** CEM, CMUP, CMX.
  - **LOOSE:** BMU, CMU, CMP, CMXNT, CMIO, SCMIO.
  - **ISOTRK:** $\Delta R > 0.1$ w.r.t other categories.
  - **LOOSE+ISOTRK** classified in Extended Muon Category (**EMC**)
- **LOOSE+TIGHT** Dilepton Veto (TopTools definition);
- **LOOSE+TIGHT+TRACK Z Veto** (TopTools definition);
- $E_T$ corrected for $\mu$, track MIP, L5 jet corrections, Primary Vtx:
  - $E_T > 20$ GeV for CEM and EMC (standard cut).
  - $E_T > 10$ GeV for CMUP and CMX (**relaxed cut**).
- QCD rejection based on Support Vector Machine Algorithm.
- 2 **TIGHT** Jets ($E_T^{L5Cor} > 20$ GeV, $|\eta| < 2.0$).
- 1 or 2 SecVtx **TIGHT** Heavy Flavor Tags.
QCD Background (Multi-jet Events)

Lepton and $\not{E_T}$ faked by mis-reconstructed jets: hard to model instrumental effects.

- Fake $W$ models obtained reversing some of the lepton identification cuts.
- Normalization measured from data using a two component fit on $\not{E_T}$ (QCD, $W$+Jet).
- Key element for total $W$+Jet estimate:
  \[
  N_{W+\text{jets}} = N_{\text{Data}}^{\text{Pretag}}(1 - F_{QCD}) - N_{MC}^{\text{MC}}
  \]

QCD normalization and modelling have large uncertainties:

- important to reduce the QCD contribution.
- New multivariate tool (based Support Vector Machine) drastically removes QCD.
QCD and Multivariate Techniques

- **Electrons sample**: large multi-jet contamination.
- **Modeling fake** $W$: “anti-electron” sample $\Rightarrow$ reverse $\geq 2$ out of 5 cuts for the shower-id;

  Is it possible to use multivariate techniques to solve this problem?

- Support Vector Machines algorithm supposed to be optimal in this case.
  - sample statistically limited ($\approx 12k$ events);
  - partial bias due to inverted selection.
- **Results**: optimal sample of 6 kinematic variables:
  - $W$ related: $P_T^{\text{ele}}, E_T^{\text{cor}}, \Delta \phi(\text{ele}, E_T^{\text{raw}})$
  - Jets related: Jet2 $E_T^{\text{raw}},$ Jet2 $E_T^{\text{cor}}, E_T$ Significance.
- $F_{\text{QCD}} \lesssim 10\%$;
- signal efficiency: $\varepsilon_{W(e, \nu)+2\text{jets}} \approx 95\%, \varepsilon_{WZ} \approx 97.5\%.$

More details in CDF NOTE 10197

- Software and results presented at CHEP2010 (22 October, Taipei)
SVM Discriminant

Concept: best hyper-plane dividing two classes of vectors.

The Support Vector Machines (or SVM) are optimal in low statistics separable samples.

- Minimization of $|\vec{w}|^2$ ($\vec{w}$ ≡ normal to the plane) with constrain:
  \[
  y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \begin{cases} 
    y_i = +1; & i \in \text{signal} \\
    y_i = -1; & i \in \text{bkg}
  \end{cases}
  \]

- Equivalent to maximize:
  \[
  L = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j
  \]

- Non-linear separation obtained with a transformation on the scalar products:
  \[
  K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad \text{with } \phi : \mathbb{R}^n \mapsto \mathcal{H} \quad K = \text{Kernel function}
  \]

Work performed in collaboration with V. Lippi, PhD student of the Scuola Superiore S. Anna.

Software used: LibSVM C++ (by C.-C. Chang, C.-J. Lin).
(4) W+LF (Mistags) Background

Wrongly reconstructed tracks and long living LF hadrons can produce fake $b$-tags:

- symmetric distribution around primary vertices:  
  $\Rightarrow$ negative vertices: opposite to the jet direction.

- per-jet fake probability parametrized on a jet control sample.

- correction factor due to long living Light Flavor hadrons.
**Full Background Estimate**

Four independent channels:

- **2 Lepton categories, **\textbf{Tight}** and **\textbf{EMC}**:  
  ⇒ homogeneous kinematics.

- **2 Tag categories, **\textbf{Single}** and **\textbf{Double}** tagged jets:  
  ⇒ homogeneous background composition.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Single Tag</th>
<th></th>
<th></th>
<th>Double Tags</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tight</td>
<td>EMC</td>
<td>Tight</td>
<td>EMC</td>
</tr>
<tr>
<td>(tt)</td>
<td></td>
<td>366 ± 37</td>
<td>172 ± 17</td>
<td>75 ± 11</td>
<td>34 ± 5</td>
</tr>
<tr>
<td>Single Top (s + t)</td>
<td></td>
<td>224 ± 25</td>
<td>87 ± 9</td>
<td>33 ± 5</td>
<td>13 ± 2</td>
</tr>
<tr>
<td>(Z + \text{jets})</td>
<td></td>
<td>98 ± 13</td>
<td>65 ± 8</td>
<td>4.1 ± 0.6</td>
<td>3.1 ± 0.4</td>
</tr>
<tr>
<td>QCD</td>
<td></td>
<td>223 ± 89</td>
<td>101 ± 40</td>
<td>12.8 ± 5.6</td>
<td>0.0 ± 0.5</td>
</tr>
<tr>
<td>(W + b\bar{b})</td>
<td></td>
<td>1135 ± 455</td>
<td>309 ± 124</td>
<td>148 ± 60</td>
<td>45 ± 18</td>
</tr>
<tr>
<td>(W + c\bar{c})</td>
<td></td>
<td>582 ± 235</td>
<td>164 ± 66</td>
<td>8 ± 3</td>
<td>3 ± 1</td>
</tr>
<tr>
<td>(W + \text{cj})</td>
<td></td>
<td>463 ± 187</td>
<td>106 ± 43</td>
<td>6.6 ± 2.7</td>
<td>1.8 ± 0.7</td>
</tr>
<tr>
<td>Mistag</td>
<td></td>
<td>1070 ± 124</td>
<td>346 ± 39</td>
<td>5.7 ± 1.2</td>
<td>2.2 ± 0.4</td>
</tr>
<tr>
<td>(ZZ)</td>
<td></td>
<td>1.51 ± 0.15</td>
<td>0.86 ± 0.08</td>
<td>0.27 ± 0.04</td>
<td>0.16 ± 0.02</td>
</tr>
<tr>
<td>(WW)</td>
<td></td>
<td>123 ± 17</td>
<td>38 ± 5</td>
<td>1.0 ± 0.2</td>
<td>0.33 ± 0.08</td>
</tr>
<tr>
<td>(WZ)</td>
<td></td>
<td>41 ± 4.5</td>
<td>14 ± 2</td>
<td>7.3 ± 1.1</td>
<td>2.4 ± 0.4</td>
</tr>
<tr>
<td>Prediction</td>
<td></td>
<td>4325 ± 895</td>
<td>1405 ± 243</td>
<td>301.6 ± 68.5</td>
<td>105.3 ± 21.5</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td>4168</td>
<td>1318</td>
<td>260</td>
<td>106</td>
</tr>
</tbody>
</table>
Likelihood Definition

Bayesian Posterior Probability

\[ p(R|\bar{n}) = \frac{\int \int d\delta d\bar{\beta} L(R, \bar{s}, \bar{b}|\bar{n})\pi(R, \bar{s}, \bar{b})}{\int \int \int dR d\delta d\bar{\beta} L(R, \bar{s}, \bar{b}|\bar{n})\pi(R, \bar{s}, \bar{b})} \Rightarrow \int_0^{R_{0.95}} p(R|\bar{n})dR = 0.95 \]

\[ R = (\sigma \times BR)/(\sigma_{SM} \times BR_{SM}), \quad R_{0.95} : 95\% \text{ Credible Level Upper Limit} \]

\[ \bar{s}, \bar{b}, \bar{n} = s_{ij}, b_{ij}, n_{ij} (\text{# of signal, background and observed events in } j\text{-th bin for } i\text{-th channel}) \]

\[ \pi : \text{Bayes' prior density} \]

Combined Binned Poisson Likelihood

\[ L(R, \bar{s}, \bar{b}|\bar{n}) = \prod_{i=1}^{N_{\text{channel}}} \prod_{j=1}^{N_{\text{bin}}} \frac{\mu_{ij}^{n_{ij}} e^{-\mu_{ij}}}{n_{ij}!} \]

Principle of ignorance

- for the number of higgs events (instead of higgs Xsec)

\[ \pi(R, \bar{s}, \bar{b}) = \pi(R)\pi(\bar{s})\pi(\bar{b}) = s_{tot}\theta(Rs_{tot})\pi(\bar{s})\pi(\bar{b}) \]

\[ s_{tot} = \sum_{i,j} s_{ij} : \text{Total number of signal prediction} \]

\[ \pi(x) = G(x|\hat{x}, \sigma_x) \quad (x = s, b) \quad \hat{x} : \text{expected mean, } \sigma_x : \text{total uncertainty} \]