Evidence for Diboson Production in the $p\bar{p} \rightarrow l\nu +$ Heavy Flavor Channel at CDF

Federico Sforza
on behalf of the CDF Collaboration

INFN & Università di Pisa

Higgs Hunting Workshop – 28 July 2011
Outline

1. Analysis Overview
2. Selection Details
3. Results

WW/WZ → lν + HF

CDF-II detector at the Fermilab TeVatron p̅p collider ($\sqrt{s} = 1.96$ TeV).
Analysis Overview

**WW/WZ → ℓν + Heavy Flavor Jets: Basics**

\[ p\bar{p} \rightarrow WW/WZ \rightarrow ℓν + b/c \text{ Jets} \]

**Event selection:**
- high \( E_T \) lepton (> 20 GeV) and large \( E_T \).
- \( Z \)-veto (orthogonal selection to \( Z+\)jets).
- 2 central jets (\( E_T > 20 \) GeV, \(|\eta| < 2.0\)).
- At least one **Heavy Flavor (HF)** Tag:
  \( \Rightarrow \) presence of a secondary vertex identifies \( b/c \) jets.

**Motivations:**
- Same final states of the \( WH \) golden channel (\( M_H \lesssim 135 \text{ GeV/}c^2 \)).
- Discrimination based on \( M_{\text{inv}}(j_1, j_2) \): understanding the HF sample on a SM resonance.
- Final \( S/B \sim O(10^{-3}) \) \( \Rightarrow \) challenging measurement.
Lepton Selection

- Composite structure of the CDF detector: unify lepton reconstruction algorithms;
- Extend the acceptance: relax cuts, keep multi-jet (QCD) background under control;

**Tight Leptons:**
- **electrons**: track + EM deposit, calorimetric isolation;
- **muons**: track + signal in the muon chambers + calorimeter MIP, calorimetric isolation;
- 3 trigger paths: high $P_T$ leptons.

**Extended $\mu$ Category (EMC)**
- **Loose Muons**: track + loose or absent signal in the muon chambers + calorimeter MIP, calorimetric isolation.
- **Isolated Track**: track + track isolation.
- 3 trigger paths: MET+jets.

**Multivariate QCD rejection** based on the Support Vector Machine algorithm:
- minimizes uncertainties associated to multi-jet events.
QCD Rejection

SVM Algorithm applied to $W \rightarrow e\nu$ (Tight Lepton example):

Same algorithm applied to the EMC leptons:
Understanding the Sample

2 Jets, Pretag
- Control sample;
- Very large statistics.
- Dominated by $W^+ \text{ Light Flavor (LF)}$ jets;

2 Jets, Single $b$-tag
- Signal sample;
- Large statistic.
- $W + c$, $W + b\bar{b}$, $W + \text{LF}$, top;

2 Jets, Double $b$-tag
- Signal sample;
- Low statistic.
- Main contribution from $W + b\bar{b}$, top;
$M_{\text{inv}}(\text{Jet1}, \text{Jet2})$ Spectrum

4 Channels: 2 lepton categories $\times$ 2 tag types
Cross Section Measurement

Channels combination

- binned likelihood fit:
  ⇒ systematics as nuisance parameters;
  ⇒ constraints from correlations across channels;

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

\[ \sigma(\text{WW}/\text{WZ} \rightarrow \ell \nu bb/\ell \nu cs) = 1.085^{+0.26}_{-0.40} \times \text{SM} \]
Significance of the Observation

- **Test statistics (TS):** Test/Null hypothesis Likelihood ratio:

\[ TS = -2\ln Q = -2\ln \frac{p(data|H_{Test}, \hat{\theta})}{p(data|H_{Null}, \hat{\theta})} \]

\( \theta \): nuisance parameters; \( \hat{\theta} \) best values under \( H_{Test} \); \( \hat{\theta} \) best values under \( H_{Null} \)

- **Significance:** compare data TS against TS generated from pseudo-data drawn from \( H_{Null} \) distribution.

- **Expected TS:** \(-8.85^{+6.2}_{-6.4}\) (Test hypothesis)

  \[ p-value = 0.00125 \Rightarrow 3.02\sigma \]

- **Observed TS:** -9.0446.

  \[ p-value = 0.00120 \Rightarrow 3.03\sigma \]
Conclusions... What’s Next

Conclusions:

- We obtained a $3\sigma$ evidence of $WW/WZ$ production in the tagged, heavy flavor enriched channel using $M_{inv}(jet1, jet2)$ as signal discriminant.
- We obtained a good agreement with SM and the measurements in other diboson channels: this supports the CDF $WH$ analysis strategy and results.

Future Plans:

- $WZ \rightarrow b\bar{b}$ alone is close at hand!!
- Ready to test-bench many other tools on a SM resonance.

Thanks for you attention!
Back Up Slides
The CDF II Detector

1. 3 silicon sub-detectors (L00, SVX II, ISL)
   - $r_{max} \simeq 30$ cm $\rightarrow$ high track density
   - coverage: $|\eta| \lesssim 2$

2. Wire chamber (COT):
   - $r_{max} \simeq 130$ cm
   - coverage: $|\eta| \lesssim 1$

3. Calorimeter system:
   - 2 sub-detectors: central e forward
   - electromagnetic (EM) and hadronic (HAD) sections.

4. Muon chambers:
   - Many sub-detectors: CMU, CMP, CMX, BMU
   - coverage: $|\eta| \lesssim 1.5$

$r, \phi, \eta \equiv -\ln[\tan(\theta/2)]$

$\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2}$

$E_T = E \sin \theta$
Selection Overview

- High energy lepton ($E_T(P_T) > 20 \text{ GeV(Gev/c)}$). 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:
  - $E_T > 20 \text{ GeV}$ for CEM and EMC (standard cut).
  - $E_T > 10 \text{ GeV}$ for CMUP and CMX (**relaxed cut**).
- QCD rejection based on Support Vector Machine Algorithm.
- **2 TIGHT** Jets ($E_T^{L5Cor} > 20 \text{ GeV}, |\eta| < 2.0$).
- 1 or 2 SecVtx TIGHT Heavy Flavor Tags.
- Signal discrimination based on $M^{inv}(\text{Jet1, Jet2})$. 

F. Sforza (INFN & Università di Pisa)
28 July 2011 13 / 10
Lepton Categories and Detector Coverage

Identified Electrons Scatter Plot

Identified Trigger Muons Scatter Plot

Lepton EMC, jet bin 2: fQCD=6.8%

Identified Tracks Scatter Plot

Identified Leptons Scatter Plot

- CEM
- PHX
- CMUP
- CMX
- LOOSE $\mu$
- ISOTRK

F. Sforza (INFN & Università di Pisa)
Jet ≡ final state of quark hadronization
- reconstruction algorithm JETCLU04
- energy corrected for detector effects (JES).

Quarks $b \Rightarrow$ Heavy Flavor hadrons (HF) long lifetime: $c\tau \approx 450 \, \mu \text{m} \Rightarrow$ secondary decay vertexes.

SecVtx algorithm:
- a jet is “tagged” if the tracks within the cone form a secondary vertex.
- $b$-tag efficiency $\sim 40\%$
- $c$-tag efficiency $\sim 6\%$
- mistags (fake tags) $\sim 1\%$ (background)
Background Composition

Signal topology: $\text{lepton} + E_T + 2\text{jets}(1 \text{ or } 2 \text{ tags})$

background estimate based both on MC and data control samples

$\Rightarrow$ 4 background components:

- **EWK**: estimate from MC ($t\bar{t}$, single top, $Z$+jets).
- **$W + LF$**: Light Flavors, $W$+fake tags $\Rightarrow$ parametrized on jet data.
- **QCD**: multi-jet events: lepton and $E_T$ are faked by mis-reconstructed jets.
  $\Rightarrow$ measured from data using a fit on $E_T$.
- **$W + HF$**: Heavy Flavors $\Rightarrow$ major background with large uncertainty.
  - Normalization obtained from data;
  - Heavy Flavor Fraction: $f_j^{HF} = \frac{W+HF}{W+\text{jets}}$ estimated from MC.
$W + b\bar{b}$, $W + c\bar{c}$, $W + c$ estimate

- Large theoretical uncertainty on $\sigma_{W + \text{jets}}$.
- Ratio $W + HF / W + \text{jets}$ derived from MC (Alpgen, LO).
- Normalization ($N^W_j$) from the pretag data sample ($N^{\text{data}}_j$):

$$N^W_j = N^{\text{data}}_j (1 - F_{j}^{\text{nonW}}) - N^{\text{EWK}}_j$$

- $(1 - F_{j}^{\text{nonW}})$: free parameter in a maximum Likelihood fit.
- $\approx 90$ MCs used:
QCD Background (Multi-jet Events)

\( (1 - F_{j}^{\text{nonW}}) \) estimated in the pretag sample:

- fake \( W \) models by reversing lepton identification cuts:
  - 1. isolation;
  - 2. EM fraction;
  - 3. shower-id.

- kinematic characteristics identical to the lepton under examination;

- maximum likelihood fit on \( E_T \);

- systematic error of 30% on \( F_{j}^{\text{nonW}} \) (conservative approach);

- important to reduce the QCD contribution in the pretag sample.
QCD and Multivariate Techniques

- **Electrons**: sample with larger multi-jet contamination.

Modeling fake $W$:

“anti-electron” sample, reverse $\geq 2$ out of 5 cuts for the shower-id;

Main issue:

- *sample statistically limited* ($\approx 12k$ events)

Is it possible to use multivariate techniques in this problem?

- Support Vector Machines algorithm supposed to be optimal in this case.
- SVM is a recent (1995) “machine learning” technique
  $\Rightarrow$ interesting field of research, never used in high energy physics.

\[
\begin{align*}
\text{Had/Em} & \leq 0.0055 + 0.00045 \times E \\
\text{Strip } \chi^2 & \neq 10 \\
L_{shr} & \leq 0.2 \\
|dz_{CES}| & \leq 3.0 \text{ cm} \\
-3.0 \text{ cm} & \geq Q_e \cdot dx_{CES} \leq 1.5 \text{ cm}
\end{align*}
\]
**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 $|\vec{w}|^2$ (where $\vec{w}$ is normal to the plane) with constrain:

$$y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0$$

\[
\begin{align*}
    y_i = +1; & \quad i \in \text{signal} \\
    y_i = -1; & \quad i \in \text{bkg}
\end{align*}
\]

- 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).
QCD Veto Based on the SVM

Training procedure and parameter selection:
- 4 training samples, 21 kinematic variables;
- *thousands of input combinations*: brute force approach using grid computing;
- SVM optimized on the training samples: maximal efficiency;
- bi-component fit on $E_T$ data distribution: consistency check of the trained discriminant.

Results:
- optimal minimal sample of variables: $P_T^{ele}$, $E_T^{cor}$, $\Delta\phi(\text{ele}, E_T^{\text{raw}})$
- Jet2 $E_T^{\text{raw}}$, Jet2 $E_T^{cor}$, $E_T$ Significance.
- QCD contamination $\approx 10\%$
- signal efficiency:
  - $\varepsilon_W(e, \nu) + 2\text{jets} \approx 95\%$, $\varepsilon_{WZ} \approx 97.5\%$.

Software and results presented to the CHEP2010 conference (22 October, Taipei)
Likelihood Definition

**Bayesian Posterior Probability**

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

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

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

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

**Combined Binned Poisson Likelihood**

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

**Principle of ignorance**

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

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

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

\[\pi(x) = G(x|\hat{x}, \sigma_x) \quad (x = s, b) \quad \hat{x}: \text{expected mean, } \sigma_x: \text{total uncertainty}\]
Minimum \( \chi^2 \) fit over the systematics nuisance parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>XS_ttbar</td>
<td>0.212 ± 0.830</td>
</tr>
<tr>
<td>CDFLUMI</td>
<td>0.112 ± 0.930</td>
</tr>
<tr>
<td>CDFLEPTONID_EMCC</td>
<td>-0.116 ± 0.954</td>
</tr>
<tr>
<td>CDFTRIGEFF_EMCC</td>
<td>-0.0769 ± 0.980</td>
</tr>
<tr>
<td>CDFVHF</td>
<td>-0.465 ± 0.223</td>
</tr>
<tr>
<td>CDFBTAGSF</td>
<td>-0.121 ± 0.894</td>
</tr>
<tr>
<td>CDFMISTAG</td>
<td>0.525 ± 0.861</td>
</tr>
<tr>
<td>ZJETS</td>
<td>-0.252 ± 0.954</td>
</tr>
<tr>
<td>CDFQCD_EMCC</td>
<td>-0.370 ± 0.803</td>
</tr>
<tr>
<td>ISRFSRPDF</td>
<td>-0.018 ± 0.999</td>
</tr>
<tr>
<td>CDFJES</td>
<td>0.427 ± 0.342</td>
</tr>
<tr>
<td>CDFQ2</td>
<td>-0.318 ± 0.662</td>
</tr>
<tr>
<td>CDFLEPTONID</td>
<td>0.089 ± 0.991</td>
</tr>
<tr>
<td>CDFTRIGEFF</td>
<td>0.044 ± 0.997</td>
</tr>
<tr>
<td>CDFQCD</td>
<td>0.244 ± 0.902</td>
</tr>
</tbody>
</table>