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

Federico Sforza
on behalf of the CDF Collaboration

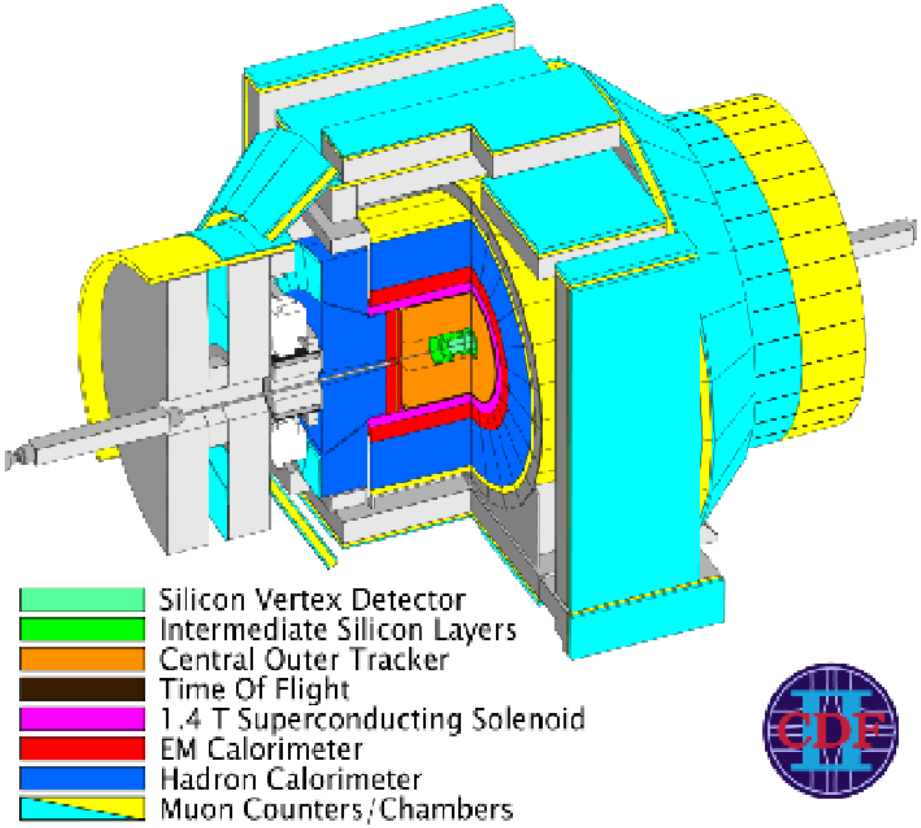
INFN & Università di Pisa

Higgs Hunting Workshop – 28 July 2011

Outline

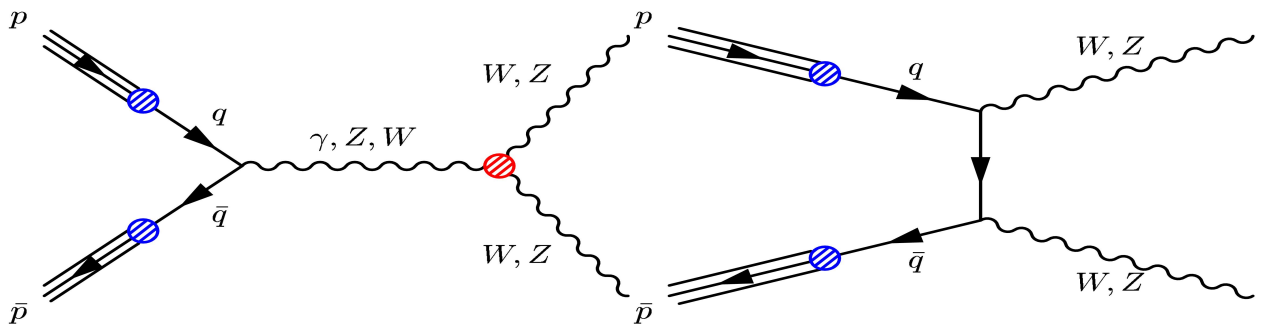
$WW/WZ \to l\nu + HF$

1. Analysis Overview
2. Selection Details
3. Results



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

$WW/WZ \to l\nu$+Heavy Flavor Jets: Basics



$p\bar{p} \to WW/WZ \to l\nu + b/c$ Jets

Event selection:

- high E_T lepton ($>$ 20 GeV) and large $\not{E}_T$.
- 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 ($M_H \lesssim 135\ \mathrm{GeV}/c^2$).
- Discrimination based on $M_{lnv}(j1, j2)$: 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.

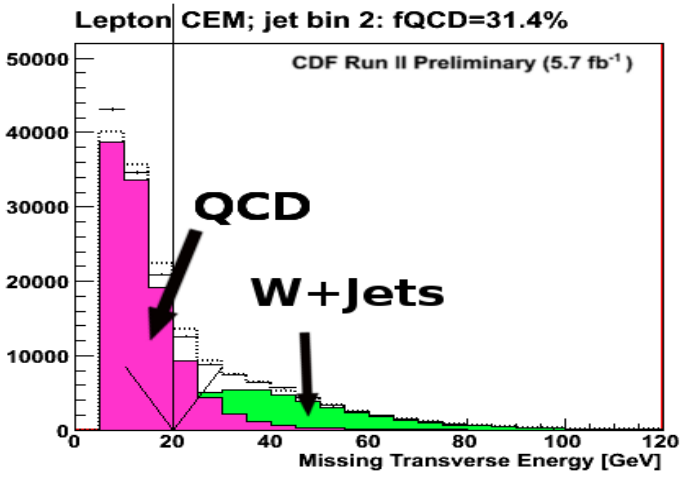
Extended μ Category (EMC)
- **Loose Muons**: track + loose or absent signal in the muon chambers + calorimenter MIP, calorimetric isolation.
- **Isolated Track**: track + track isolation.
- 3 trigger paths: MET+jets.

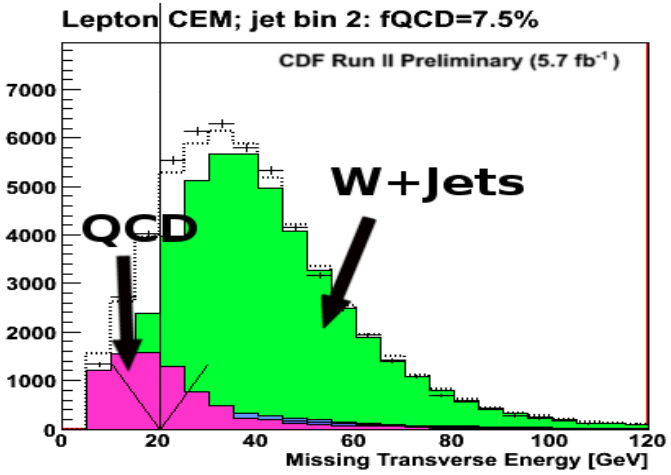
Multivariate **QCD rejection** based on the Support Vector Machine algorithm:

- minimizes uncertainties associated to multi-jet events.

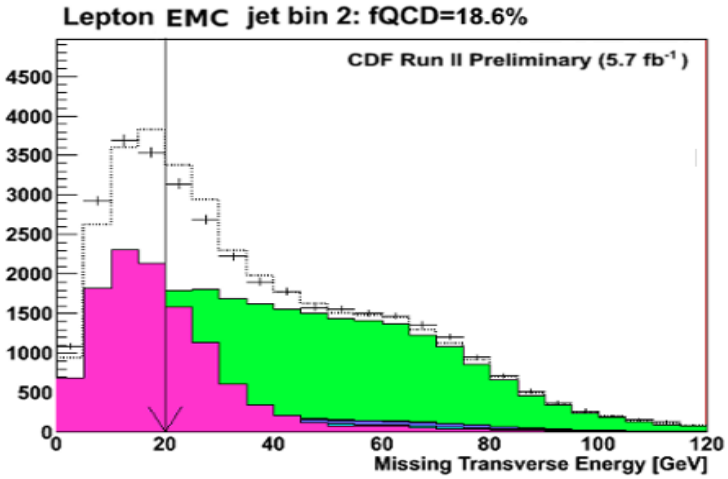
QCD Rejection

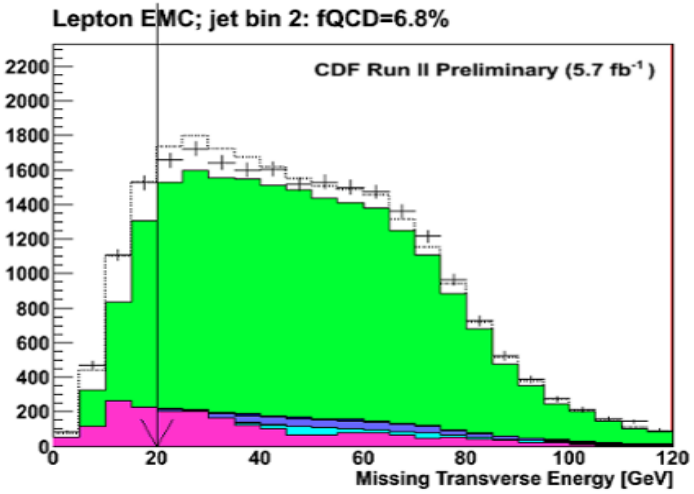
SVM Algorithm applied to $W \to e\nu$ (Tight Lepton example):





Same algorithm applied to the EMC leptons:

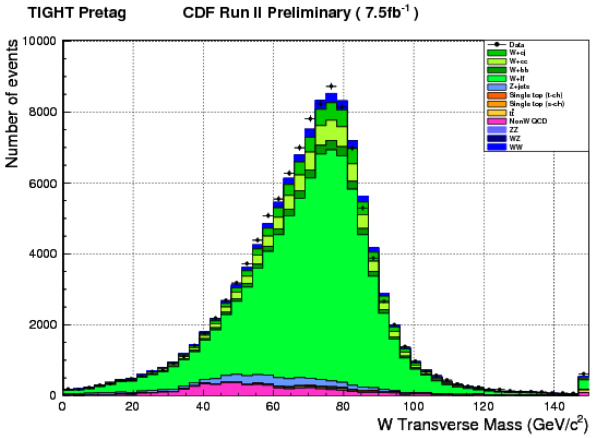




Understanding the Sample

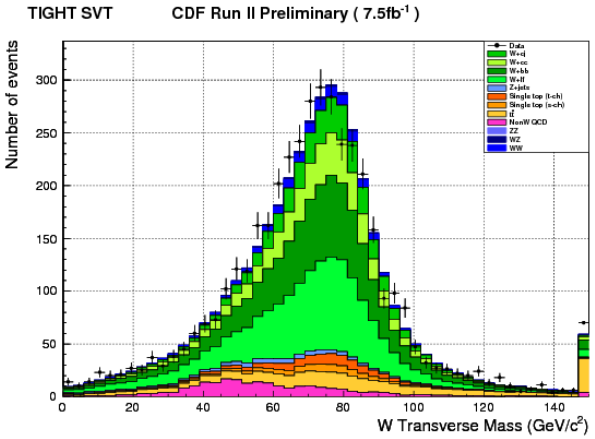
2 Jets, Pretag

- Control sample;
- Very large statistics.
- Dominated by W+ Light Flavor (LF) jets;



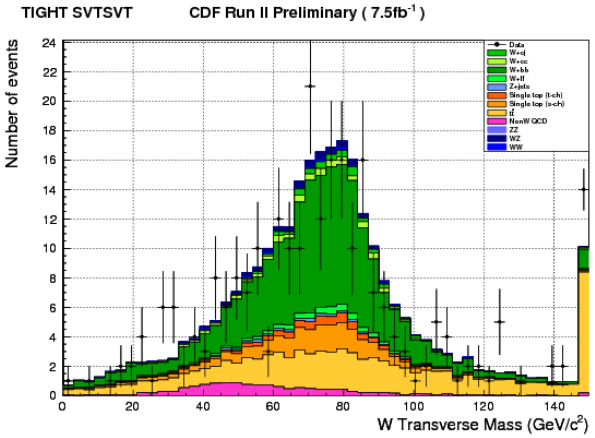
2 Jets, Single b-tag

- Signal sample;
- large statistic.
- $W + c$, $W + b\bar{b}$, $W + LF$, top.;



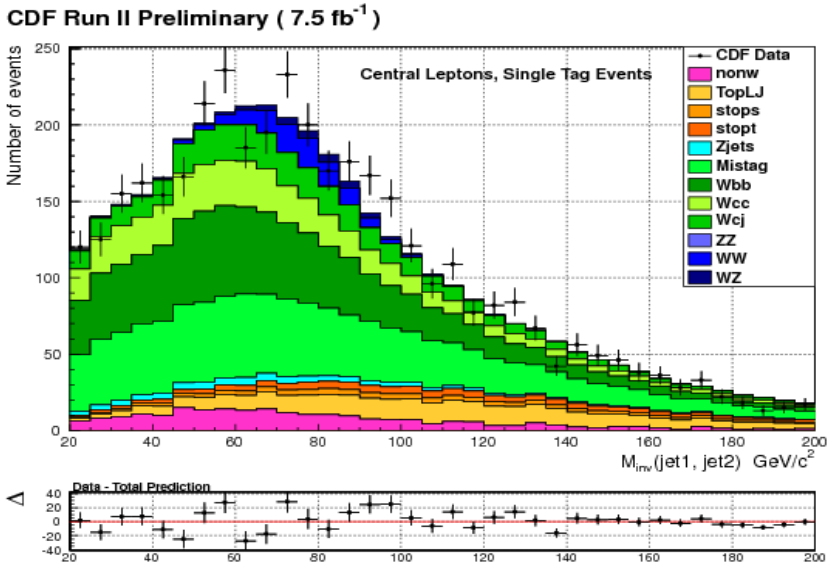
2 Jets, Double b-tag

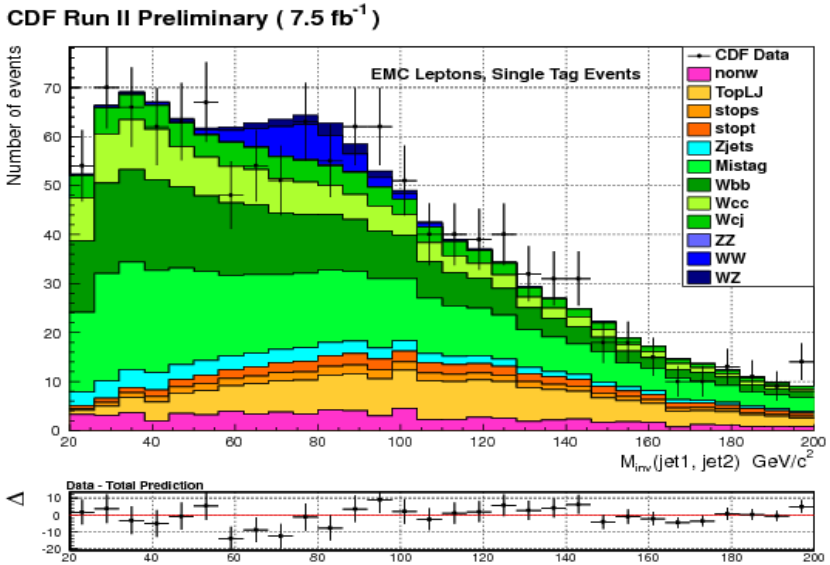
- Signal sample;
- Low statistic.
- Main contribution from $W + b\bar{b}$, top;

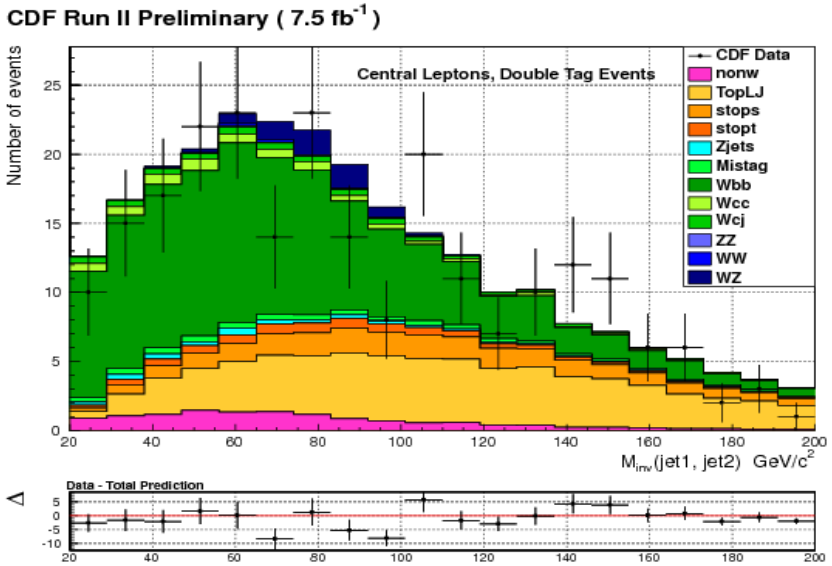


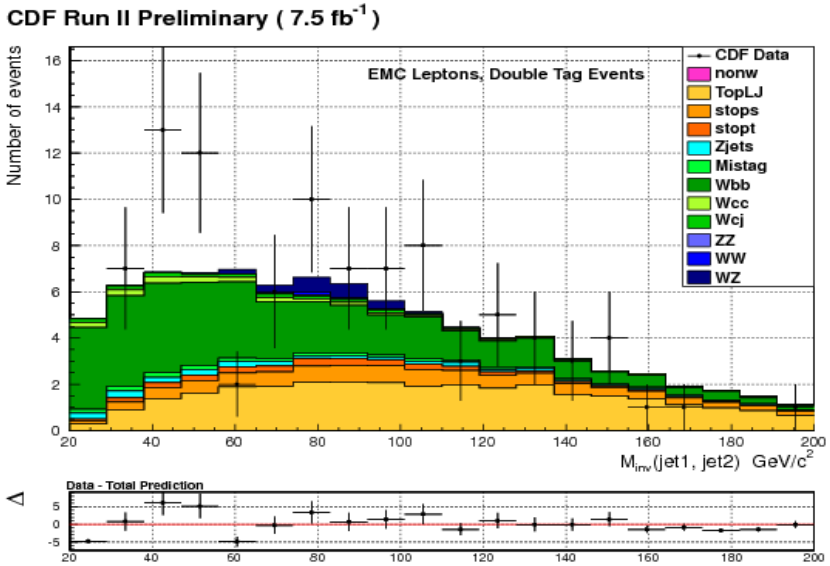
$M_{Inv}(Jet1, Jet2)$ Spectrum

4 Channels: 2 lepton categories $\times$ 2 tag types





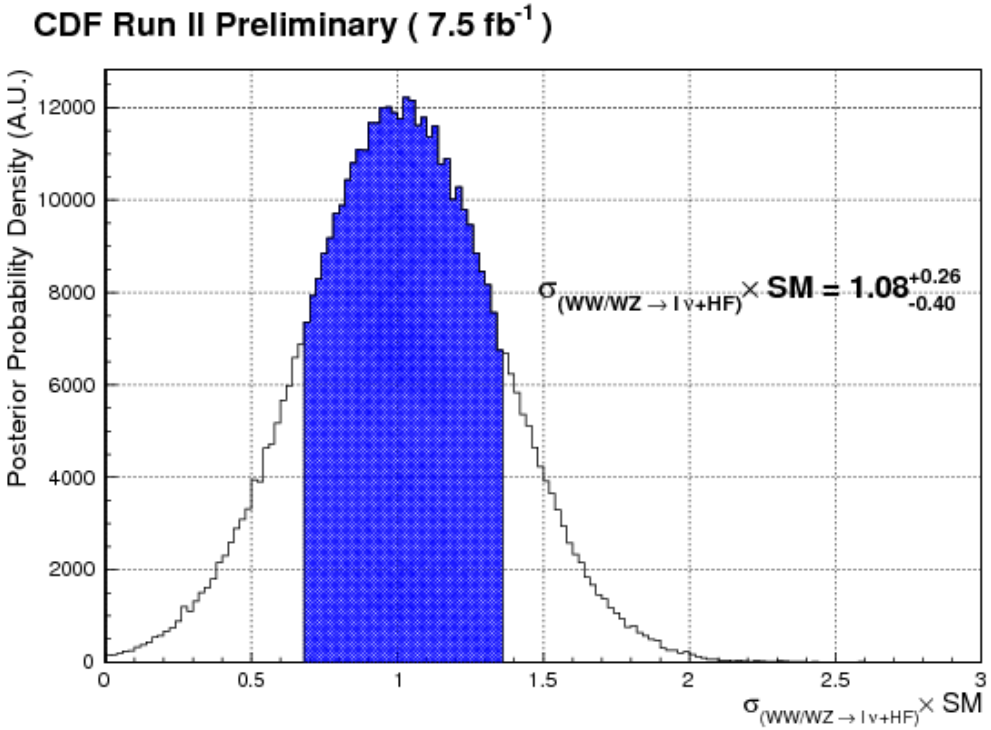




Cross Section Measurement

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.



$$\sigma_{(WW/WZ \to l\nu bb/l\nu cs)} = 1.085^{+0.26}_{-0.40} \times SM$$

Significance of the Observation

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

$$TS = -2lnQ = -2ln\frac{p(data|H_{Test}, \hat{\theta})}{p(data|H_{NUll}, \hat{\hat{\theta}})}$$

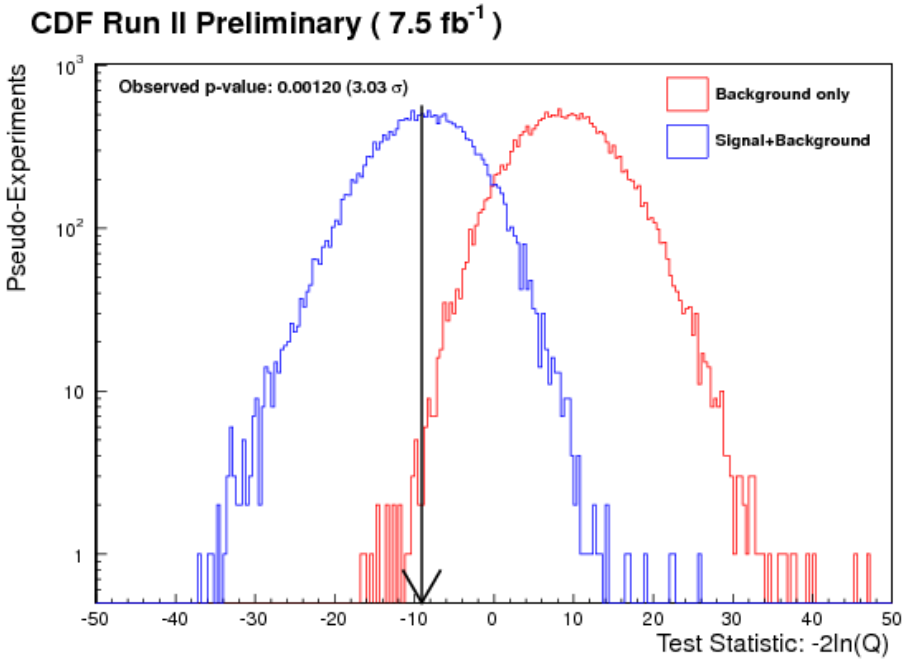
  θ: nuisance parameters; $\hat{\theta}$ best values under H_{Test}; $\hat{\hat{\theta}}$ best values under H_{Null}
- Significance: compare data TS against TS generated from *pseudo-data* drawn from H_{Null} distribution.

- 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 3σ 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:

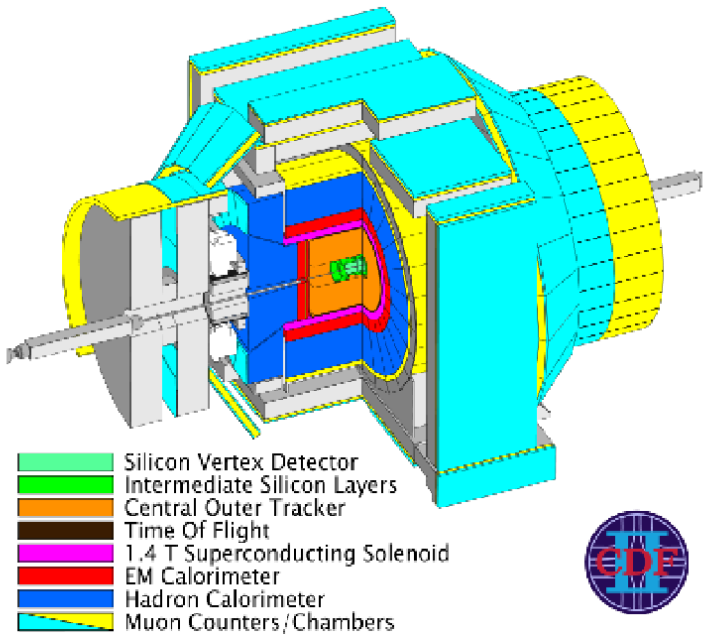
- $WZ \to b\bar{b}$ alone is close at hand!!
- Ready to test-bench many other tools on a SM resonance.

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: $|\eta| \lesssim 1.5$



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

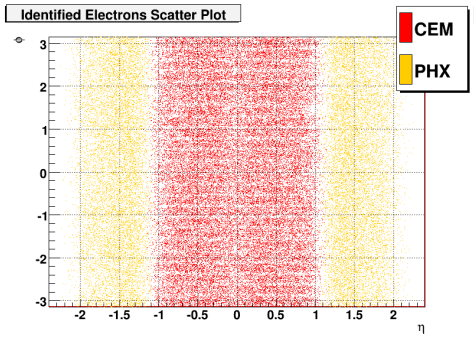
$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$

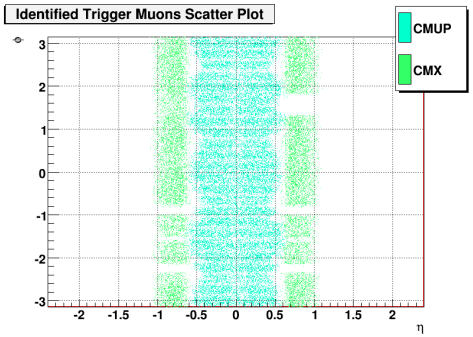
$E_T = E \sin\theta$

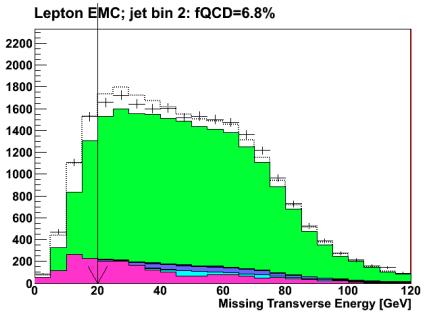
Selection Overview

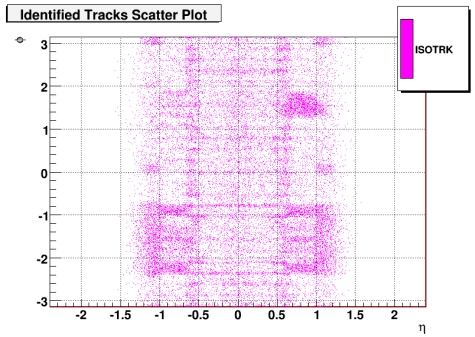
- High energy lepton ($E_T(P_T) > 20$ GeV(Gev$/c$)). *Maximize acceptance:*
  - TIGHT: CEM, CMUP, CMX.
  - LOOSE: BMU, CMU, CMP, CMXNT, CMIO, SCMIO.
  - ISOTRK: $\Delta R > 0.1$ w.r.t other categories.
  - LOOSE+ISOTRK classified in Extended Muon Category (**EMC**)
- LOOSE+TIGHT Dilepton Veto (TopTools definition);
- LOOSE+TIGHT+TRACK Z Veto (TopTools definition);
- $\not\!\!E_T$ corrected for μ, track MIP, $L5$ jet corrections, Primary Vtx:
  - $\not\!\!E_T > 20$ GeV for CEM and EMC (standard cut).
  - $\not\!\!E_T > 10$ 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.
- Signal discrimination based on $M^{inv}(Jet1, Jet2)$.

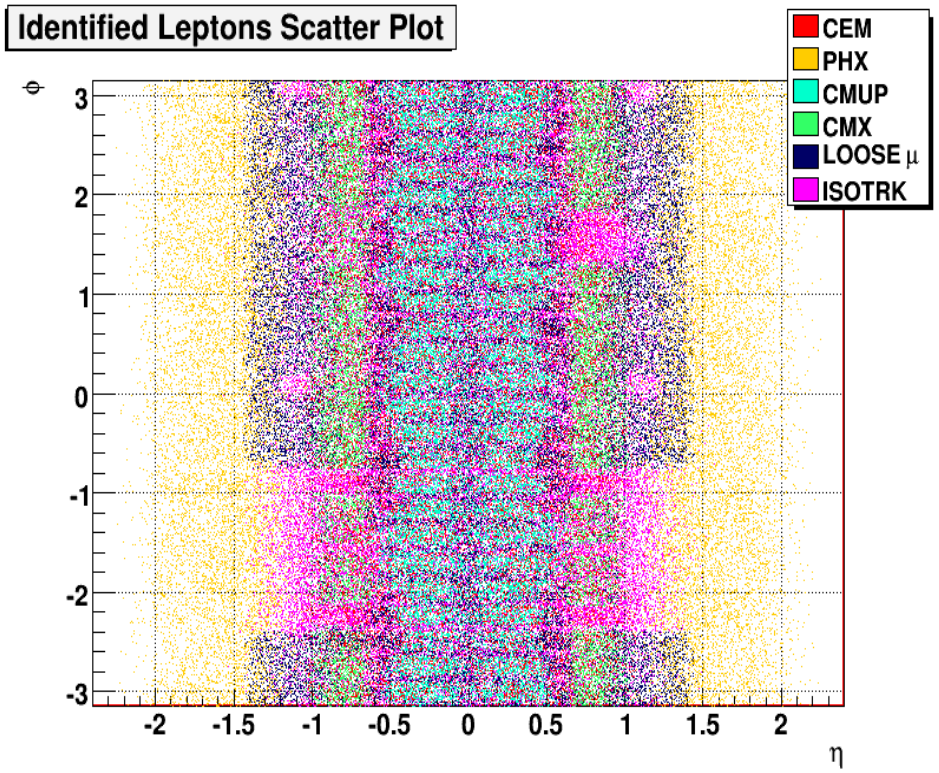
Lepton Categories and Detector Coverage









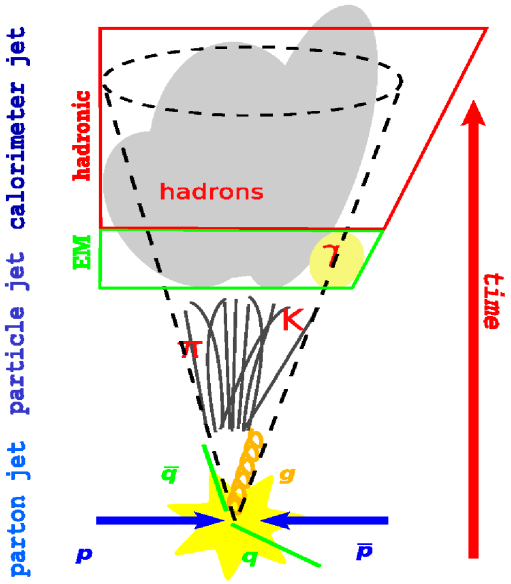


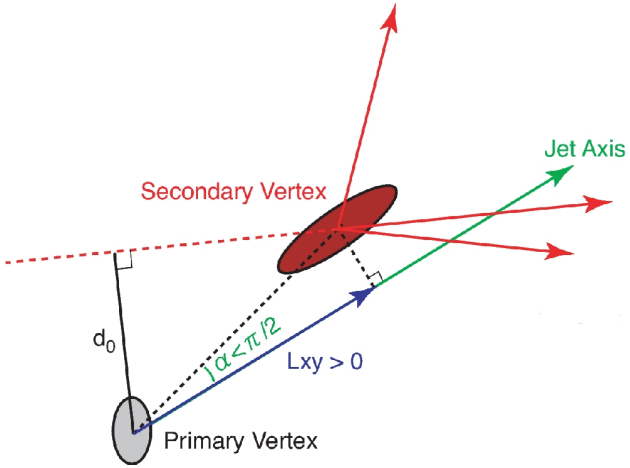
Jet Selection

Jet≡ final state of quark hadronization

- reconstruction algorithm `JETCLU04`
- energy corrected for detector effects (JES).

Quarks $b \Rightarrow$ Heavy Flavor hadrons(HF) long lifetime: $c\tau \simeq 450\ \mu$ 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: *lepton*$+\not{E}_T + 2$*jets(1 or 2 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 $\not{E}_T$ are faked by mis-reconstructed jets.
$\Rightarrow$ measured from data using a fit on $\not{E}_T$.

$W + HF$: Heavy Flavors $\Rightarrow$ major background with large uncertainty.

- Normalization obtained from data;
- Heavy Flavor Fraction: $f_j^{HF} = \frac{W+HF}{W+jets}$ estimated from MC.

W+Heavy Flavor Background

$W + b\bar{b}$, $W + c\bar{c}$, $W + c$ estimate

- Large theoretical uncertainty on σ_{W+jets}.
- Ratio $W + HF/W + jets$ derived from MC (Alpgen, LO).
- Normalization (N_j^W) from the *pretag* data sample (N_j^{data}):

$$N_j^W = N_j^{data}(1 - F_j^{\mathrm{nonW}}) - N_j^{EWK}$$

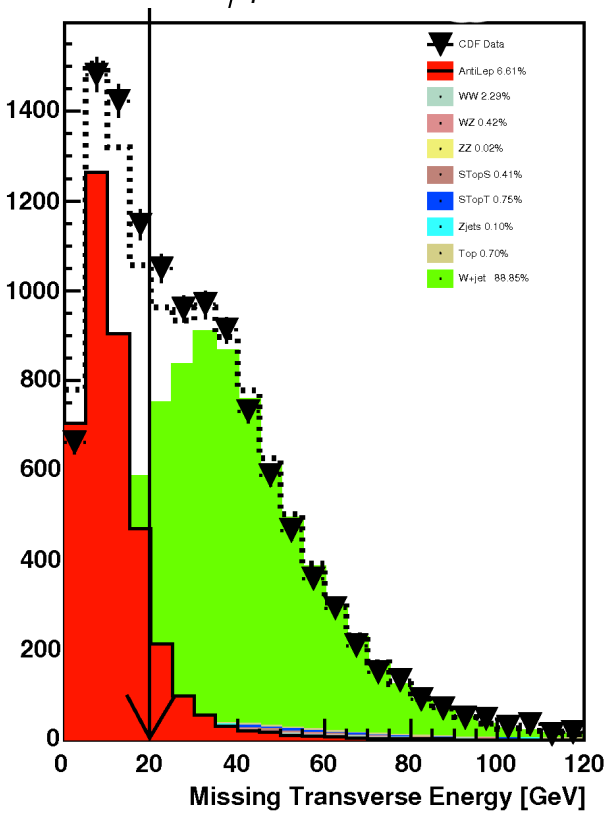
- $(1 - F_j^{\mathrm{nonW}})$: free parameter in a maximum Likelihood fit.
- $\approx$ 90 MCs used:

QCD Background (Multi-jet Events)

$(1 - F_j^{\mathrm{nonW}})$ 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 $\not\!\!E_T$;

Example of a $\not\!\!E_T$ fit on a muon sample:



- systematic error of 30% on F_j^{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 ≥ 2 out of 5 cuts for the shower-id;

$$Had/Em \leq 0.0055 + 0.00045 \times E$$
$$\text{Strip } \chi^2 \neq 10$$
$$L_{shr} \leq 0.2$$
$$|dz_{CES}| \leq 3.0 \text{ cm}$$
$$-3.0 \text{ cm} \geq Q_e \cdot dx_{CES} \leq 1.5 \text{ cm}$$

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.

SVM Discriminant

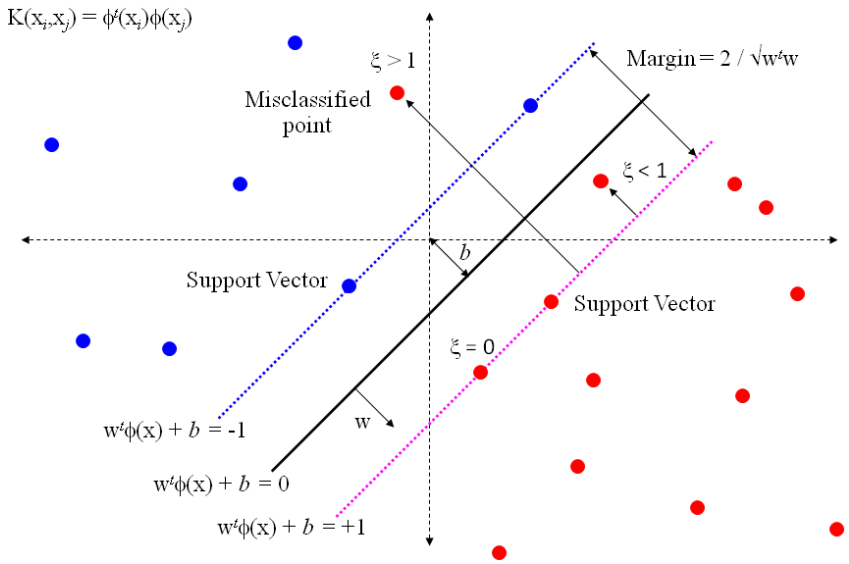
Concept: best hyper-plane dividing two classes of vectors.

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

- Minimization of $|\vec{w}|^2$ ($\vec{w} \equiv$ 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:

$$\mathbf{K}(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad \text{with } \phi : \Re^n \mapsto \mathcal{H} \;\; \text{K = Kernel function}$$

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

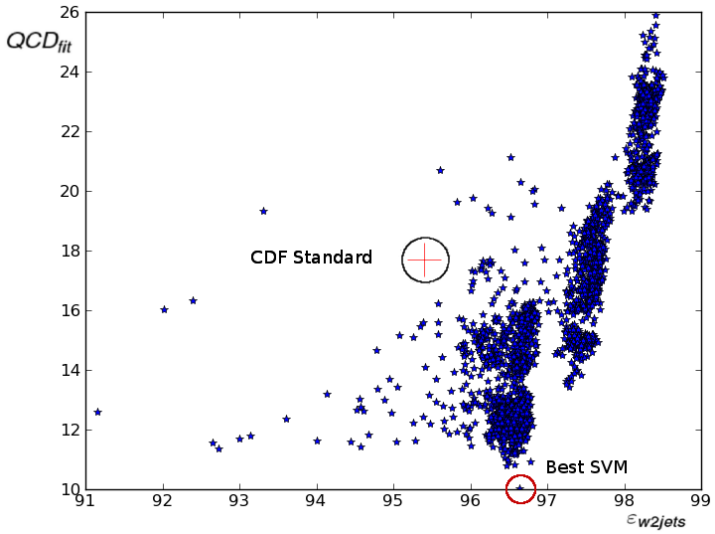
Training procedure and parameter selection:

- 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 $\not{E}_T$ data distribution: consistency check of the trained discriminant.

Results:

- optimal minimal sample of variables:
  P_T^{ele}, $\not{E}_T^{cor}$, $\Delta\phi(\mathrm{ele}, \not{E}_T^{raw})$
  Jet2 E_T^{raw}, Jet2 E_T^{cor}, $\not{E}_T$Significance.
- QCD contamination $\approx$ 10%
- signal efficiency:
  $\varepsilon_{W(e,\nu)+2\mathrm{jets}} \approx 95\%$, $\varepsilon_{WZ} \approx 97.5\%$.

SVM in the f_{QCD}^{fit} vs $\varepsilon_{W+2\mathrm{jets}}$ plane:



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

Likelihood Definition

Bayesian Posterior Probability

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

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

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

π : Bayes' prior density

Combined Binned Poisson Likelihood

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

Principle of ignorance

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

$$\pi(R,\vec{s},\vec{b}) = \pi(R)\pi(\vec{s})\pi(\vec{b}) = s_{tot}\theta(Rs_{tot})\pi(\vec{s})\pi(\vec{b})$$

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

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

Systematics

Minimum χ^2 fit over the systematics nuisance parameters.

XS_ttbar	0.212 ± 0.830
CDFLUMI	0.112 ± 0.930
CDFLEPTONID_EMC	-0.116 ± 0.954
CDFTRIGEFF_EMC	-0.0769 ± 0.980
CDFVHF	-0.465 ± 0.223
CDFBTAGSF	-0.121 ± 0.894
CDFMISTAG	0.525 ± 0.861
ZJETS	-0.252 ± 0.954
CDFQCD_EMC	-0.370 ± 0.803
ISRFSRPDF	-0.018 ± 0.999
CDFJES	0.427 ± 0.342
CDFQ2	-0.318 ± 0.662
CDFLEPTONID	0.089 ± 0.991
CDFTRIGEFF	0.044 ± 0.997
CDFQCD	0.244 ± 0.902