Diboson Search and Multivariate Tools in the \( l\nu + b/c \) Jets Channel at CDF

Federico Sforza

INFN & Università di Pisa
Examples of improvements thanks to machine learning techniques:

1. Diboson Search: Basics and Problems
2. Support Vector Machines
3. Neural Networks
**Diboson Search: Basics and Problems**

**WW/WZ → lν + b/c Jets Search Basics**

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

**Event selection:**
- high \( E_T \) lepton (> 20 GeV) and \( \not{E}_T \) (> 20 GeV).
- 2 central jets \( (E_T > 20 \text{ GeV}, |\eta| < 2.0) \).
- At least one *Heavy Flavor (HF) Tag*:
  \[ \Rightarrow \text{ presence of a secondary vertex identifies } b/c \text{ jets.} \]

**Motivations:**
- Rare process (e.g. \( \sigma_{WZ} = 3.22 \pm 0.23 \text{ pb} \)) never observed in this decay channel.
- Same final states of the \( WH \) golden channel \( (M_H \lesssim 145 \text{ GeV}/c^2) \).
- \( S/B \sim O(10^{-3}) \), large \( \Delta B \): counting experiments are not possible.
Challenges and Solutions

Machine learning techniques are used in several areas:

1. Remove hard to model backgrounds:
   - need to: reduce multi-jet ($QCD$) background.
   - solution: QCD Veto based on Support Vector Machines (SVM) algorithm.

2. Distinguish quark flavor:
   - need to: understand underlying structure of $HF$ tagged jets.
   - solution: Neural Network Flavor Separator.

3. Improve invariant mass resolution:
   - need to: exploit maximum information to refine jet energy measure.
   - solution: Neural Network b-quark specific jet energy correction.
Central Electrons (CEM): sample with \( \approx 25 - 30\% \) events from multi-jet (QCD) contamination.

*data driven* QCD model of the fake \( W \) obtained reversing \( \geq 2 \) out of 5 electron-id requirements:

- not reproducible with MC;
- sample statistically limited (\( \approx 20k \) events).

Is it possible to use multivariate techniques in this problem?

- **Support Vector Machines** algorithm supposed to be optimal in this case.
- “recently” (1995) developed machine learning technique.
- interesting tool rarely exploited in high energy physics.
Concept: best hyper-plane dividing two classes of vectors.

- Minimization of $|\vec{w}|^2$ ($\vec{w} \equiv$ normal to the plane) with constrain:

\[
y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \left\{ \begin{array}{l} y_i = +1; \quad i \in \text{signal} \\ y_i = -1; \quad i \in \text{bkg} \end{array} \right.
\]

- 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:

\[
K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad \text{with} \quad \phi : \mathbb{R}^n \mapsto \mathcal{H} \quad K = \text{Kernel function}
\]
QCD Veto Based on the SVM Algorithm

Training procedure and parameter selection:
- discrimination based on combinations of 21 test variables.
- *thousands of input combinations*: grid computing for brute force approach;
- Criteria: maximal efficiency on the training samples.
- Consistency check on data control region.

Results:
- optimal “machine” obtained with minimal set of 6 variables:
  - 3 related to the $W$ kinematic;
  - 2 related to the 2$^{nd}$ Jet energy;
  - 1 relating all jet correction to the $E_T$.
- QCD contamination $\lesssim 10\%$
- signal efficiency:
  - $\varepsilon_{W(e,\nu)+2$jets} \approx 95\%$, $\varepsilon_{WZ} \approx 97.5\%$.

Software and results presented to the CHEP2010 conference and accepted in proceedings.
QCD Rejection

Algorithm applied to $W \rightarrow e\nu$:

![Graph showing QCD and W+Jets distributions](image1)

Same algorithm applied to $W \rightarrow \mu\nu$:

![Graph showing QCD and W+Jets distributions](image2)
Neural Network Based Improvements:

- Heavy Flavor Separator.
- b-jet NN-energy corrections.

These improvements play a fundamental role in several CDF analysis (Single-top, WH, etc.)
Neural Network Generalities

**Concept:** non-linear model of input distributions based on a sigmoid functions serie.

Optimal algorithm to trace hidden correlations, prefers large training sets.

- Obtain the best weights \((\omega_i)\) for \(o_k\):

\[
o_k = S \left( \sum_{j=0}^{M} \omega_{jk} \cdot S \left( \sum_{i=0}^{d} \omega_{ij} x_j + \mu_{0j} \right) \right)
\]

- \(d\) input nodes, \(M\) hidden nodes, \(k\) output nodes
- Sigmoid or activation function: \(S = \pm 1\) if node report signal or bkg

Training:

- function minimization in a \(M \times d\)-dim space.
**KIT* Flavor Separator**

**Aim:** retain most of the signal (b-jets) pulling apart c and light flavor jets contribution.
- Played fundamental role in Single-top discovery.

- Secondary vertex identification already tags the jet.
- Weaker correlations:
  - \( b \) production \( \leftrightarrow \) jet structure:
    - per track variables, tracks multiplicity, vertex mass...
- NN with 26 input nodes, 10 hidden nodes, 1 output node.
- 2 side distribution: \( \Rightarrow \) \( b-c \) quarks separation.

* developed by the Karlsruhe Institute of Technology.
**b-Jet Energy Corrections**

Standard CDF jet energy corrections can be improved assuming $b$-quark as the initial parton:

- $b$-hadrons fragmentation, presence of semileptonic decays, color flow effects: difficult to disentangle the correlations.
- NN with 9 input variables: from tracking, calorimeter, secondary vertex

- Optimization on WH MonteCarlo.
- Improvement also in WZ invariant mass resolution.
- $Z$ peak resolution \((mean/sigma)\):
  
  \[0.154 \pm 0.003 \Rightarrow 0.116 \pm 0.002\]
Conclusions

Cut based analysis can be not enough in complicated environments when looking for rare signal.

- Machine learning techniques are a powerful tool but can not be used thoughtlessly:
  - over-fitting, training set choosing, test on control samples, etc.
- Understanding the involved physical processes is crucial to obtain good results.

Thanks for you attention!
Pretag: CEM (top), CMUP (bottom)

**Met**

**Lepton $P_T$**
Pretag: CMX (top), EMC (bottom)

Met

Lepton $P_T$
Pretag: CEM (top), CMUP (bottom)

\( \Delta R(\text{lep}, \text{jet}1) \)  \quad \Delta \phi(\text{met}, \text{jet}1) \quad WM_T

\begin{align*}
\text{CEM Pretag} & \quad \text{CDF Run II Preliminary (5.7 fb}^{-1}) \\
\text{CMUP Pretag} & \quad \text{CDF Run II Preliminary (5.7 fb}^{-1}) \\
\end{align*}

F. Sforza (INFN & Università di Pisa)
Pretag: CMX (top), EMC (bottom)

$\Delta R(\text{lep, jet1})$  $\Delta \phi(\text{met, jet1})$  $W_{M_T}$
The CDF II Detector

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

2. Wire chamber (COT):
   - \( r_{\text{max}} \simeq 130 \text{ cm} \)
   - 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: \(|\eta| \lesssim 1.5\)

\( r, \phi, \eta \equiv -\ln[\tan(\theta/2)] \)
\[ \Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2} \]
\[ E_T = E \sin \theta \]
Lepton Selection

The detector has a composite structure:

effort to unify lepton reconstruction algorithms (9) and trigger paths (7):

**CEM, PHX:**
- **electrons**: EM deposit + track, calorimetric isolation;
- 2 dedicated trigger paths.

**CMUP, CMX:**
- **muons**: signal in the muon chambers + track + MIP in the calorimeter, calorimetric isolation;
- 2 dedicated trigger paths.

**BMU, CMU, CMP, CMXNT**
- **muons (LOOSE)**: signal in the muon chambers + tracks, calorimetric isolation;
- 3 generic trigger paths: MET+jets.

**ISOTRK:**
- **tracks (mainly muons)**: high quality track, track isolation;
- 3 generic trigger paths: MET+jets.
Jet $\equiv$ final state of quark hadronization

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

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

SecVtx algorithm:

- a jet is "tagged" if the tracks within the cone form a secondary vertex.
- $b$-tag efficiency $\approx 40\%$
- $c$-tag efficiency $\approx 6\%$
- mistags (fake tags) $\approx 1\%$ (background)
Signal topology: \textit{lepton} + $E_T$ + 2jets (1 or 2 tags)

⇒ \textit{4 background components:}

\textbf{EWK:} estimate from MC ($t\bar{t}$, s-top, $Z$+jets, $WW$, $ZZ$).

\textbf{Mistag:} $W$+fake tags ⇒ parametrized on jet data.

\textbf{QCD:} multi-jet events: lepton and $E_T$ are faked by mis-reconstructed jets.

⇒ measured from data using a fit on $E_T$.

\textbf{$W$+HF:} Heavy Flavors ⇒ major background with large uncertainty.

- Normalization obtained from data;
- $f_j^{HF} = \frac{W+HF}{W+\text{jets}}$ estimated from MC.
$W + b\bar{b}, W + c\bar{c}, W + c$ estimate

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

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

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

\[ (1 - F_{\text{nonW}}^j) \] 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 \( \not{E}_T \);

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

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

**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
  \[\Rightarrow \text{interesting field of research, never used in high energy physics.}\]