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

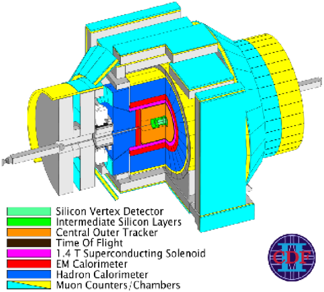
Federico Sforza

INFN & Università di Pisa

Outline

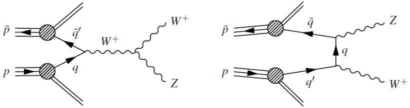
Examples of improvements thanks to machine learning techniques:

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



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

$WW/WZ \rightarrow l\nu + b/c$ Jets Search Basics



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

Event selection:

- high E_T lepton (> 20 GeV) and $\not{E}_T$ (> 20 GeV).
- 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:

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

Challenges and Solutions

Machine learning techniques are used in several areas:

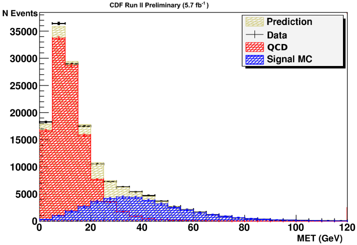
1. Remove hard to model backgrounds:
   - need to: reduce multi-jet (*QCD*) background.
   - solution: QCD Veto based on Support Vector Machines (SVM) algorithm.
2. Distinguish quark flavor:
   - need to: understand underlying structure of *HF* tagged jets.
   - solution: Neural Network Flavor Separator.
3. Improve invariant mass resolution:
   - need to: exploit maximum information to refine jet energy measure.
   - solution: Neural Network b-quark specific jet energy correction.

QCD and Multivariate Techniques

- Central Electrons (CEM): sample with $\simeq 25 - 30\%$ events from multi-jet (*QCD*) contamination.

data driven QCD model of the fake W obtained reversing ≥ 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.

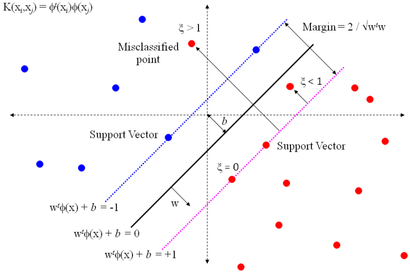
SVM Discriminant

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

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

$$y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \begin{cases} y_i = +1; & i \in \text{signal} \\ y_i = -1; & i \in \text{bkg} \end{cases}$$

- 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:

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

QCD Veto Based on the SVM Algorithm

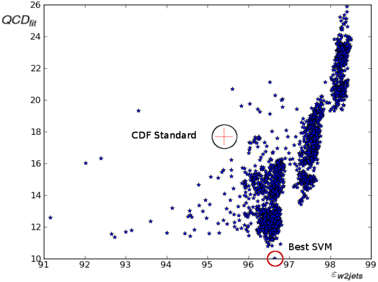
Training procedure and parameter selection:

- 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:
  $\Rightarrow$ **3** related to the W kinematic;
  $\Rightarrow$ **2** related to the 2^{nd} Jet energy;
  $\Rightarrow$ **1** relating all jet correction to the $\not\!\!E_T$.
- QCD contamination $\lesssim 10\%$
- signal efficiency:
  $\varepsilon_{W(e,\nu)+2\text{jets}} \approx 95\%$, $\varepsilon_{WZ} \approx 97.5\%$.

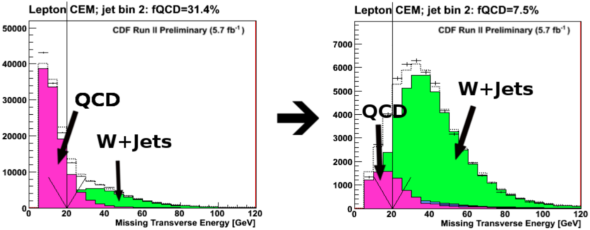
SVM in the f^{fit}_{QCD} vs $\varepsilon_{W+2\text{jets}}$ plane:



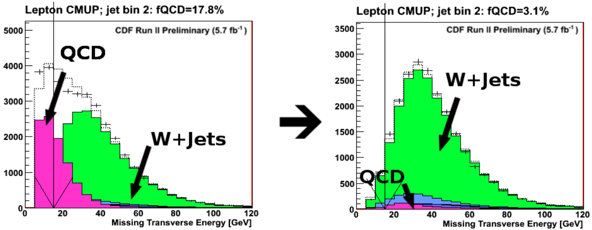
Software and results presented to the CHEP2010 conference and accepted in proceedings.

QCD Rejection

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



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



Neural Network Based Improvements:

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

These improvements play a foundamental role in several CDF analysis (Single-top, WH, etc.)

Neural Network Generalities

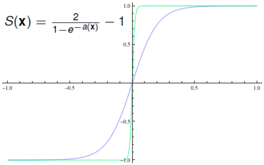
Concept: non-linear model of input distributions based on a sigmoid functions serie.

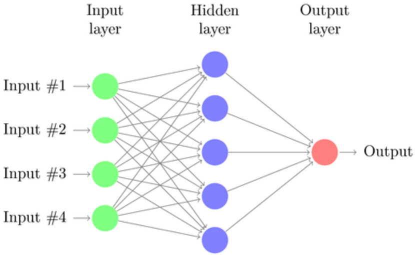
Optimal algorithm to trace hidden correlations, prefers large training sets.

- Obtain the best weights (ω_i) for o_k:

$$o_k = S\Big(\sum_{j=0}^{M} \omega_{jk} \cdot S(\sum_{i=0}^{d} \omega_{ij} x_j + \mu_{0j})\Big)$$

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





Training:

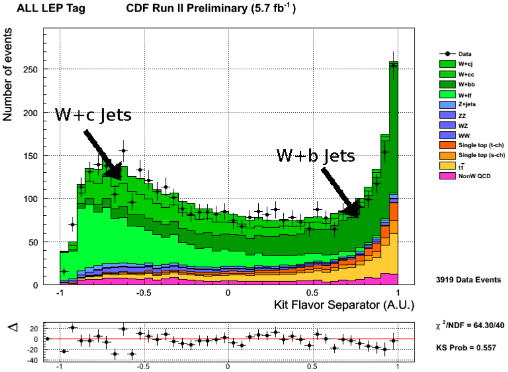
- function minimization in a $M \times d$-dim space.

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.

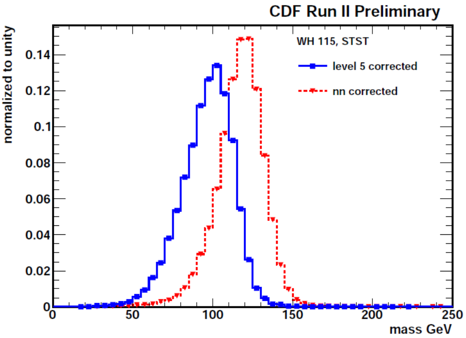
b-Jet Energy Corrections

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 thoughtlessy:
  - over-fitting, training set choosing, test on control samples, etc.
- Understanding the involved physical processes is crucial to obtain good results.

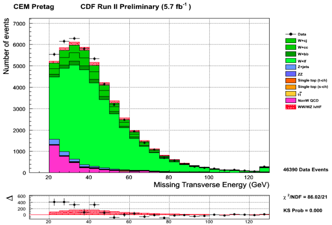
Thanks for you attention!

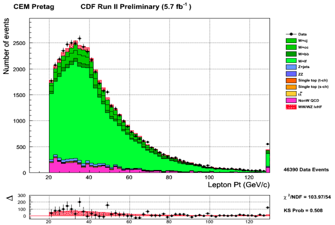
Back Up Slides

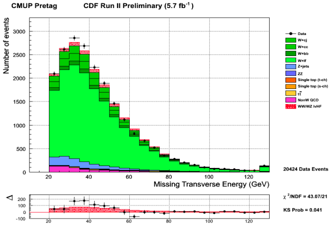
Pretag: CEM (top), CMUP (bottom)

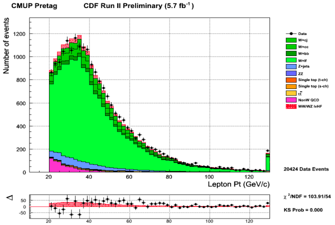
Met

Lepton P_T



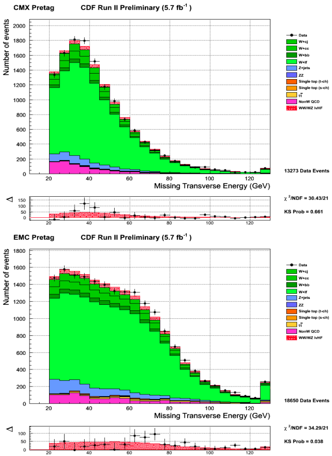




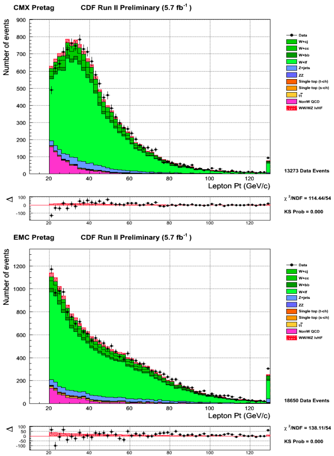


Pretag: CMX (top), EMC (bottom)

Met

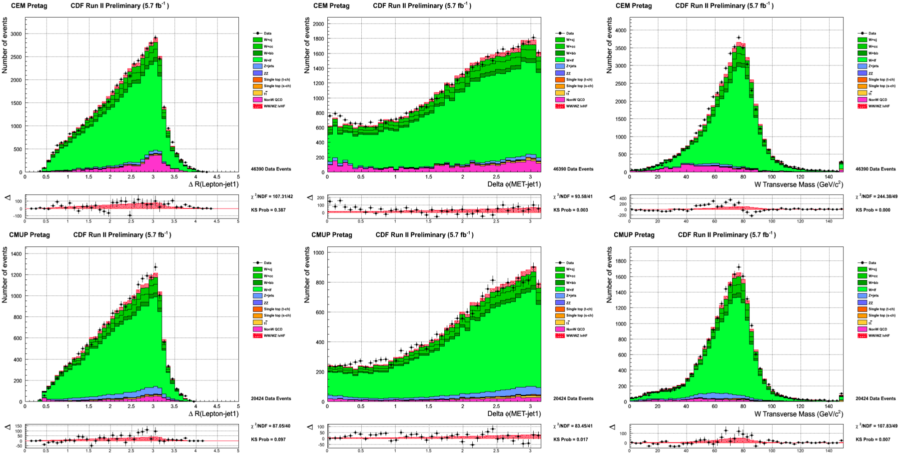


Lepton P_T



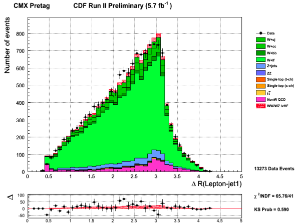
Pretag: CEM (top), CMUP (bottom)

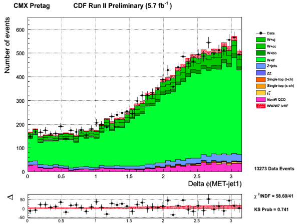
$\Delta R(lep, jet1)$ $\Delta\phi(met, jet1)$ WM_T

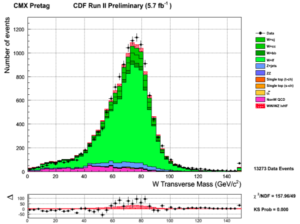


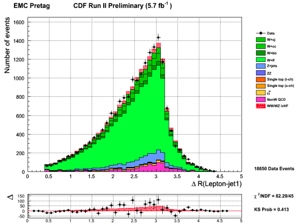
Pretag: CMX (top), EMC (bottom)

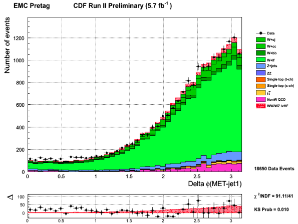
$\Delta R(lep, jet1)$ $\Delta\phi(met, jet1)$ WM_T

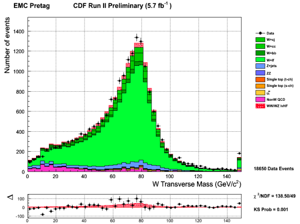






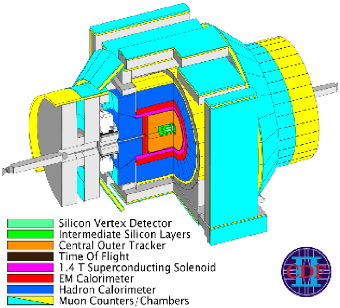






The CDF II Detector

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: $|\eta| \lesssim 1.5$



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

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

$E_T = E \sin\theta$

Lepton Selection

The detector has a composite structure:

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

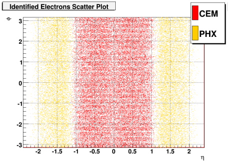
CEM, PHX:
- **electrons**: EM deposit + track, calorimetric isolation;
- 2 dedicated trigger paths.

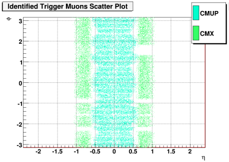
CMUP, CMX:
- **muons**: signal in the muon chambers + track + MIP in the calorimeter, calorimetric isolation;
- 2 dedicated trigger paths.

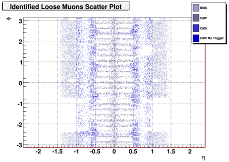
BMU, CMU, CMP, CMXNT
- **muons (LOOSE)**: signal in the muon chambers + tracks, calorimetric isolation;
- 3 generic trigger paths: MET+jets.

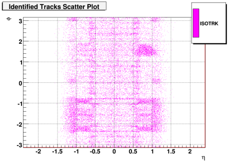
ISOTRK:
- **tracks (mainly muons)**: high quality track, track isolation;
- 3 generic trigger paths: MET+jets.

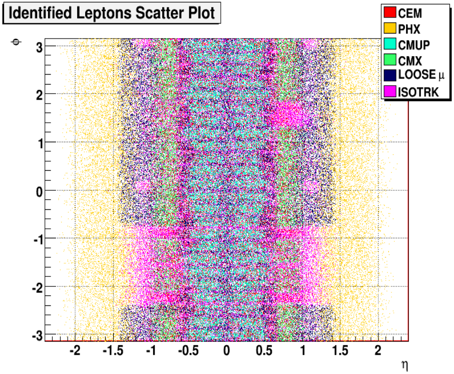
Lepton Categories and Detector Coverage









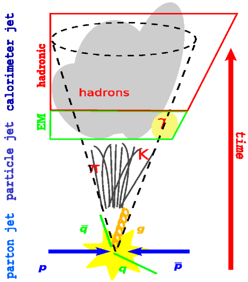


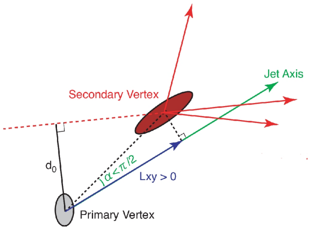
Jet Selection

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 o 2 tags)*

$\Rightarrow$ **4 background components:**

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

Mistag: $W+$fake tags $\Rightarrow$ parametrized on jet data.

QCD: multi-jet events: lepton and $\not{E}_T$ are faked by mis-reconstruted 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;
- $f_j^{HF} = \frac{W+HF}{W+jets}$ 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}):

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

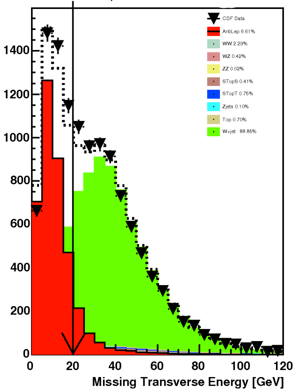
- $(1 - F_j^{\mathrm{nonW}})$: free parameter in a maximum Likelihood fit.
- $\approx$ 90 MCs used:

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;

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

Main issue:

- *sample statistically limited (≈ 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
  ⇒interesting field of research, never used in high energy physics.