IL NUOVO CIMENTO Vol. 34 C, N. 6 Novembre-Dicembre 2011
DOI 10.1393/ncc/i2011-11025-6

Colloquia: IFAE 2011

Diboson search and multivariate tools in the $p\bar{p} \to l\nu$ + heavy flavor channel at CDF

F. Sforza
INFN and Università di Pisa - Pisa, Italy

(ricevuto il 29 Luglio 2011; pubblicato online il 25 Ottobre 2011)

Summary. — This paper describes the application of machine learning techniques to the diboson search in the lepton plus neutrino plus heavy flavor jets channel at CDF. Three different aspects of this challenging search are analyzed: multi-jet background rejection with the use of a support vector machine discriminant, light/heavy flavor jets separation with a 26 input variable neural network and b-jet specific energy corrections, where a resolution improvement is obtained feeding a neural network with both calorimeter and tracking information.

PACS 14.70.-e – Gauge bosons.

1. – Introduction

Diboson associate production (WW/WZ) has been observed at the CDF detector in the lepton plus jets decay channel but no distinction was made about the flavor composition of the jets. The tagging of Heavy Flavor (HF) jets allows the identification of the process $WZ \to l\nu + b\bar{b}$, a fundamental step along the Higgs search.

Multivariate algorithms can provide help but a good understanding of the physics problem and of the technique itself is required. This paper reviews three different algorithms used in diboson searches at CDF. The first algorithm, based on a new Support Vector Machine (SVM) discriminant [1], reduces the multi-jet background contamination: the algorithm is optimal as we deal with a data-driven, not perfect background model. The second and the third algorithms [2] provide, respectively, b-quark separation and improved jet energy resolution; they are based on Neural Networks (NN) and on their ability to model the non-linear correlation in a set of input variables.

2. – Event selection and backgrounds description

The analyzed candidate events are required to have an electron (tranverse energy, E_T, greater than 20 GeV) or a muon (tranverse momentum, P_T, greater than 20 GeV/c), missing transverse energy ($\not\!\!E_T > 20$ GeV) and two central jets ($E_T > 20$ GeV, absolute value of the pseudorapidity, $|\eta|$, less than 2.0), at least one of the jets must contain a

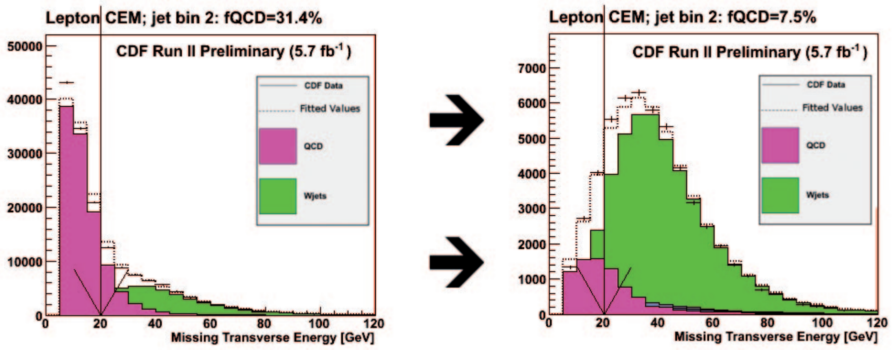

Fig. 1. – Selected electron+jets events before (left) and after (right) the application of the SVM QCD veto.

reconstructed secondary vertex that *tags* the event as HF. Multi-jet events (QCD background) can pass the selection when the lepton is faked by a jet with high electromagnetic energy fraction or which is not fully reconstructed, this can produce also an incorrect evaluation of the $\not{E}_T$. QCD background is the main limiting factor when trying to relax the selection. A second background is produced by Light Flavor (LF) jets with an identified secondary vertex, which can mimic the signal when long-living LF are present or because of resolution effects. The last background is composed by real $l\nu + HF$ non-resonant events and can be separated from the signal only on a kinematic basis, for this reason is extremely important a good jet energy resolution.

3. – QCD veto based on the SVM algorithm

The QCD modeling is based on a data-driven approach: we derive a background enriched sample reversing 2 out of 5 electron ID cuts. This produces a low-statistics ($\sim$ 10000 events) sample with a partial bias due to the inverted selection. The SVM algorithm was found to perform well in this problem: it looks for the maximum margin hyperplane between the classes of the elements (x_i) of the training set. This is derived minimizing $|w|^2$ ($w \equiv$ vector normal to the plane) with the constraint

$$(1) \quad y_i(x_i \cdot w + b) - 1 \geq 0 \quad (y_i = +1,\ i \in \text{signal},\ y_i = -1;\ i \in \text{background},\ b = \text{offset}).$$

Or, introducing the Lagrange multipliers α: $L = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$.

Non-linear separation can be obtained with a modification of the scalar products:

$$(2) \quad \mathbf{K}(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \qquad \text{with } \phi : \Re^n \mapsto \mathcal{H}, \quad \mathbf{K} : \Re^n \mapsto \Re,\ K \equiv \text{Kernel}.$$

where ϕ is a non-linear transformation that remains unknown.

The work in [1] describes the development of a QCD rejection SVM discriminant that exploits the information of the W related kinematic, the energy of the second most energetic jet and the global calorimeter activity. The obtained discriminant is very efficient (QCD contamination $F_{\text{QCD}} \lesssim 8\%$; signal efficiency $\varepsilon_{W(e,\nu)+2\text{jets}} \approx 95\%$, $\varepsilon_{WZ} \approx 97.5\%$) as can be see from fig. 1.

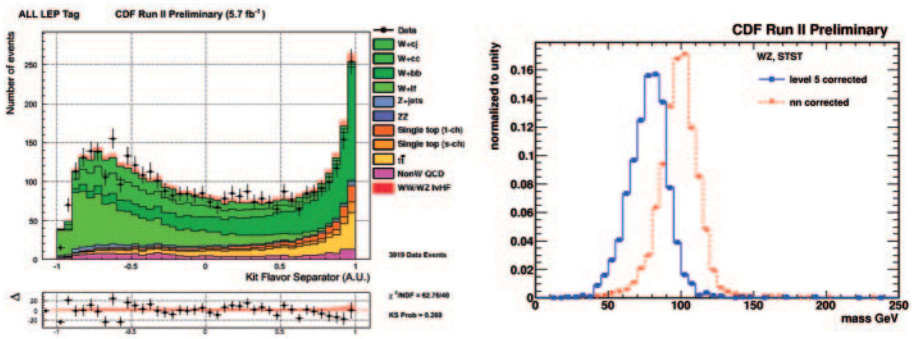

Fig. 2. – Application of NN based tools to HF/LF separation in the $W+$jet b-tagged channel (left) and to jet energy correction ($Z \rightarrow b\bar{b}$ mass resolution on the right).

4. – Neural-network–based improvements

Neural-network–based tools are used to improve HF/LF jets separation [2] and b-jet energy resolution, at the moment they play a fundamental role in several CDF analysis (*e.g.*, Single-top and WH analysis). NN are designed to fit a multi-variate input distribution with a composition of series of sigmoid functions ($S(x)$). Given an input vector x_i, the output of a NN is given by

$$o_k = S\left(\sum_{j=0}^{M} \omega_{jk} \cdot S\left(\sum_{i=0}^{d} \omega_{ij} x_i + \mu_{0j}\right)\right), \tag{3}$$

where d are the input nodes, M the hidden nodes, k the output nodes. The weights $(\omega_{ij}, \omega_{jk})$ and the offset term (μ_{0j}) are optimized on the training set, that should be large. The NN algorithm is optimal to trace hidden correlations between variables.

The first NN (KIT Flavor Separator [2]) discriminates between LF and HF jets when a secondary vertex is already reconstructed. The NN exploits the weaker correlations between b-quarks and the jet sub-structure: 26 input variables are used, *i.e.* per-track variables, second vertex characteristics and global variables of the jet. Figure 2 (left) shows the good separation obtained. The last considered tool produces a b-jet specific energy correction that improves the invariant mass resolution by $\simeq 10\%$ (fig. 2, right). This is possible combining the calorimeter, tracking and secondary vertex information.

5. – Conclusions

We reviewed three different multivariate tools used in the $WZ \rightarrow b\bar{b}$ search: the first is based on a SVM and reduces the multi-jet contamination to $\lesssim 8\%$, the second and the third algorithms are based on a NN and provide, respectively, a good LF/HF separation and a better jet energy resolution (10% improvement on di-jet mass resolution).

REFERENCES

[1] Sforza F., Lippi V., Chiarelli G. and Leone S., CDF Public Note 10197 (2010).
[2] Richter S., Ph.D. thesis, Universitat Karlsruhe (2007).