IL NUOVO CIMENTO Vol. 33 C, N. 4 Luglio-Agosto 2010
DOI 10.1393/ncc/i2010-10689-6

Colloquia: TOP2010

# Experimental methods in top physics

F. Canelli

*Enrico Fermi Institute and Department of Physics, University of Chicago - Chicago, IL, USA*
*Fermi National Accelerator Laboratory - Batavia, IL, USA*

(ricevuto l' 8 Novembre 2010; approvato il 10 Novembre 2010; pubblicato online il 19 Novembre 2010)

**Summary.** — Top quark physics at the Tevatron has inspired particle physicists to develop new methods of understanding hadron collider data. From sophisticated one-dimensional fitting techniques to multidimensional neural networks, the top quark physics program has innovated many techniques which at this time are used extensively in the high-$p_T$ program at the Tevatron and LHC.

PACS 14.65.Ha – Top quarks.

## 1. – Introduction

This proceeding describes methods and techniques used in top quark physics at the Tevatron. Many of these techniques were developed for the observation of the top quark in pair production. Some of them were developed with the goal of improving the precision of the measurement of the top quark mass. New techniques were introduced during Run II with the purpose of extracting single top quarks from immense background events.

As the experimental challenges increased, most methods implemented in analyses became multivariate, meaning more than one variable per event is used to extract information about the signal. These techniques are not only more complicated than analysis of one variable, but they demand more understanding of the correlations among inputs.

In this paper I will describe top quark measurement techniques from CDF and D0 with an emphasis on multivariate techniques and their comparison. Due to the large length of this topic and to avoid missing some references, the reader can find all the information about these analyses in the D0 and CDF experiments top group webages [1, 2].

## 2. – Methods in top quark physics

Methods can be categorized into four different types: counting experiment, templates, matrix element and machine learning. Counting experiment is probably the simplest of all and is used in cases where we can find regions of space where the amount of signal over background is considered large. Template techniques refer to the use of the shape

 

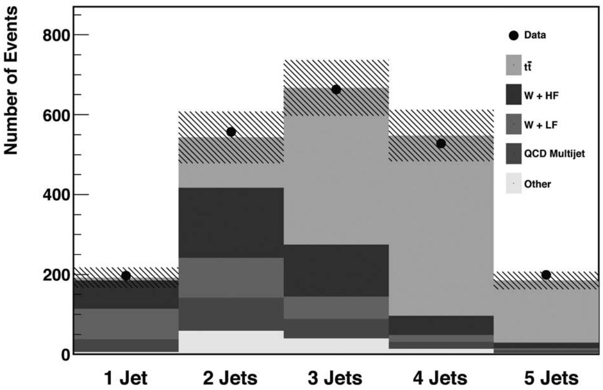


Fig. 1. – CDF measurement of the top cross section in lepton + jets events presented in winter 2010. Histograms of expected and observed number of events as a function of jet multiplicity for events with at least one b-tagged jet.

of one variable for extracting information about the signal. It relies on having a good signal-to-background ratio in some region of that particular variable. Matrix element techniques calculate probability densities for the signal and background hypothesis and evaluate these for each event. The way that this information is used varies depending on the type of analysis, measurement or search, that is performed. Machine learning techniques use the signal and background Monte Carlo samples to create a discriminating function between them. It uses many variables to feed in the machine learning algorithm.

**2**˙1. *Counting experiment*. – A counting experiment is the simplest way to establish a cross section. Once the selection criteria is established the main component of this method is to estimate the expected background. The final result is obtained from subtracting the expected background events from the events found in the data. The selection criteria is chosen to minimize uncertainty by having large acceptance for top quarks and small background and systematic uncertainties. The uncertainty on the background must be many times smaller than the expected signal in order to obtain a decent result. The requirement of identifying b-hadrons (b-tagging) reduces background rates to a level such that top quarks can be counted above background. This method establishes the background expectations used for all other types of top quark measurements. Figure 1 shows the numbers of events as a function of the number of jets for one of the latest measurements of the top cross section in the lepton+jets channel.

**2**˙2. *Templates*. – Template methods fit the shape of the data to one or two variables chosen because the signal and background have noticeably different shapes. A likelihood fit is done between the data and the different templates obtained from Monte Carlo. Many of the variables used require reconstruction of the top quarks momenta. The longitudinal momentum of the neutrino and the different choice of jet permutations are chosen by minimizing a $\chi^2$, which is a different function for each of the dilepton, lepton+jets, and all-jets channel final states. Rate uncertainties as well as ones which change the shape of the signal and background templates are taken as systematic uncertainties. One of the main uses of this technique is for the measurement of the top quark mass and for $t\bar{t}$ resonance searches. The reconstructed top quark mass is the most used variable. During

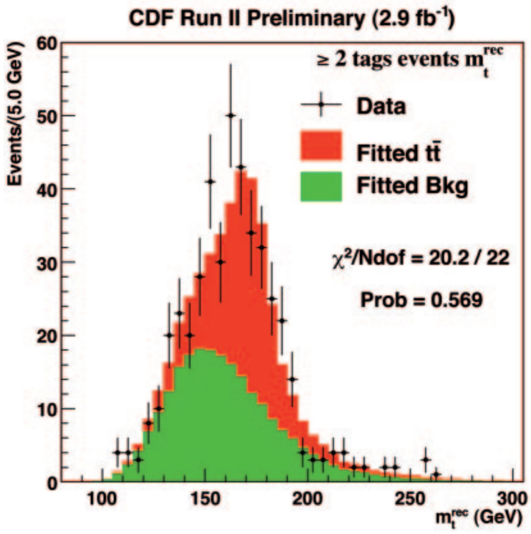


Fig. 2. – CDF measurement of the top quark mass measurement in all + jets events. Reconstructed mass for background and $t\bar{t}$ templates and data.

Run II of the Tevatron these analyses included a second variable, the invariant mass of the two light jets constrained to be that of a $W$ candidate, in order to minimize the effect of the jet energy scale systematic uncertainty. In the last few years, additional variables such as the lepton $p_T$ have been introduced. Although these variables are simpler to measure and reconstruct, they are more susceptible to shape systematic uncertainties. The discriminant outputs of neural networks and matrix element methods are also used as templates. Figure 2 shows the fit of the top quark mass measurement in the all-hadronic channel.

**2**·3. *Matrix element*. – The matrix element method defines a likelihood for an event to be due to a given production process based on the differential cross section of the process. The probability density for a given process is defined by normalizing the differential cross section to the total cross section: $P \sim \frac{\mathrm{d}\sigma}{\sigma}$. $P$ is not a true probability, as various approximations are used in the calculation of the differential cross section. Since the initial state momenta are unknown, the differential cross section is weighted with parton distribution functions (PDFs) for the proton. Because knowledge of the final state is limited by detector resolution, an integration over transfer functions encoding the relationship between the measured quantities $x$ and the parton-level quantities $y$ is performed. The input measured quantities are the four momenta of the jets and lepton in the event. The probability density is then given by

$$P(x) = \frac{1}{\sigma} \int \mathrm{d}\sigma(y) \mathrm{d}q_1 \, \mathrm{d}q_2 f(q_1) f(q_2) W(y, x), \tag{1}$$

where $f(q_1)$ and $f(q_2)$ are the PDFs in terms of the fraction of the proton momentum ($q_i = E_{qi}/E_{\mathrm{beam}}$), and $W(y, x)$ is the transfer function. The squared matrix element, $|\mathcal{M}|^2$, is calculated at tree level using Monte Carlo generator packages like MADGRAPH for

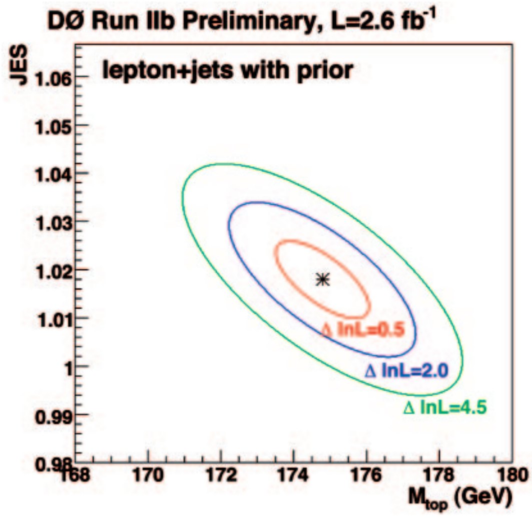


Fig. 3. – D0 measurement of the top quark mass using a matrix element technique in the lepton+jets channel.

signal and background hypothesis. In $W(y,x)$, the lepton energy and angle are assumed to be measured exactly. The jet energy transfer function is derived by parametrizing a comparison between parton energies to the fully simulated jet response in Monte Carlo events. Recent iterations of the top quark mass measurements include the jet angle resolutions in addition to energies. The measured missing transverse energy from all calorimeter towers is not used to calculate the event probability. Instead, conservation of momentum of the measured high-$p_T$ objects determines the momentum of the neutrino.

Depending on the channel and type of backgrounds, the integral to derive the event probability can be cumbersome. Usually the integration is carried out numerically using an adaptation of the CERNLIB routine or the faster DIVONNE integration algorithm implemented in the CUBA library. The integration time has been one of the disadvantages of this method. Although this strongly depends on the technology implemented to do the integration.

**2**˙3.1. Measurements. The matrix element method was created with the purpose of measuring the top quark mass precisely. Signal and background probability densities are combined in an unbinned likelihood function. This likelihood is properly normalized using the detector acceptance which includes the effects of trigger, event selections, etc. Events enter into the likelihood with different weights, different permutation of jets contributing differently, and no assumptions of the initial configuration. Background events tend to have a low signal probability value and higher background probability value. Figure 3 shows the two-dimensional measurement of the top quark mass and the jet energy scale uncertainty.

**2**˙3.2. Searches. The probability densities from matrix element formulation can be used to do searches for rare processes by building a discriminant of the form $P_{\text{signal}}/(P_{\text{background}} + P_{\text{signal}})$ so that background-like events will have values close to zero and signal-like events will have values close to unity. The $P_{\text{signal}}$ and $P_{\text{background}}$ are just the sum of individual probabilities for signal and background processes, the event

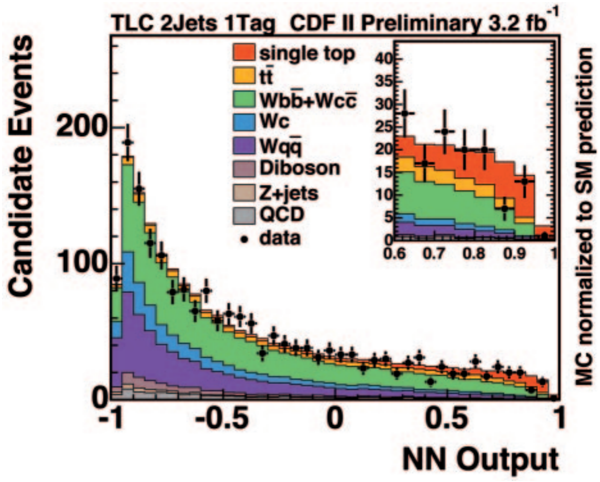


Fig. 4. – Neural network discriminant for the CDF single top quark cross section measurement.

probability discriminant, or EPD. This method was first used by the single top search. It is now a standard way to search for Higgs or other processes.

**2·4.** *Machine learning.* – Machine learning techniques are a mathematical model or a computational model which using many types of input (signal and background events) qualifies an event as more signal-like or background-like event. The most frequently used algorithms in top quark physics are artificial neural networks and boosted decision trees. Monte Carlo is used to select good discriminating variables which are input to the machine learning algorithms. In contrast with matrix element methods, the inputs can be more convoluted variables or even functions of other variables. Good variables are defined by how different they appear for signal and background and also by how well they are modeled in the data in control regions deprived of signal. The second requirement is usually satisfied by $\chi^2$ testing each variable in one or more control regions where background is abundant. The first requirement is estimated by the multivariate algorithm. A significance of each variable is evaluated, incorporates only those that include relevant information that is not already incorporated by other variables.

Neural networks are widely used in top quark physics. They are part of the selection criteria in analyses like the D0 top cross section and CDF all-hadronic top quark cross section and mass measurements. In other analyses the final discriminant output is used as a template for extracting a cross section measurement. Figure 4 shows a neural network analysis output for the extraction of the single top quark cross section. Neural networks are used also in flavor-tagging algorithms.

## 3. – Comparison of methods

The top quark physics groups have taken advantage of many different methods for measurements and searches, in order to provide redundant cross-checks of results and to improve precision by combining them.

All these methods rely on Monte Carlo at different levels. A good Monte Carlo simulation is essential to achieve precise and accurate results. In general the main variables, the high-$p_T$ 4 vectors, are well simulated. Softer $p_T$ quantities, like $\Delta R$ among jets, forward jets, small jet $p_T$, etc. are not modeled as accurately.

The current most precise $t\bar{t}$ cross section uses a neural network fit with 7 input variables. It takes advantage of the good top quark statistics in the lepton+jets before b-tagging (pre-tag). The main systematic uncertainty is due to the jet energy scale. Although the top cross section using b-tagging has a larger uncertainty, its major systematic uncertainty is due to b-tagging. Since the overlap among the pre-tag and b-tag samples is small and the systematic uncertainties complementary, the best result is obtained combining these two results.

In top mass measurements it has been found that the matrix element is more precise than the template method by about 20%.

In the single top cross section many methods have been developed. The neural network, boosted decision trees, and matrix element methods have comparable sensitivities. Neural networks provided the most sensitive method at CDF, while D0 found boosted decision trees achieved the best result. The matrix element method has been found in both experiments to have a very similar sensitivity to them.

New methods were also incorporated to handle systematic uncertainties. In general, a more natural way to deal with systematic uncertainties is to include them in the fitting method. Analyses like the measurement of the top quark mass and the single top cross section have included the major uncertainties as part of the fitting method.

## 4. – Conclusions

The top quark group at the Tevatron developed many methods and techniques. The race for the observation of the top quark in Run I, the precision top quark mass measurement in Run I and Run II, and the observation of single top quark production has caused the top quark groups in both collaborations to explore and compare many methods of analyses. Good Monte Carlo modeling and Tevatron data has allowed these techniques to mature. More importantly, these techniques have helped us get a deeper knowledge of the top quark.

REFERENCES

[1] `http://www-d0.fnal.gov/Run2Physics/top/`
[2] `http://www-cdf.fnal.gov/physics/new/top/top.html`