PROCEEDINGS OF SCIENCE

Uses of multivariate analysis methods

Yann Coadou[*†]
Simon Fraser University (Canada)
E-mail: yann@sfu.ca

The top quark study groups at the Tevatron accelerator use several analysis techniques to extract precision measurements of top quark properties and search for single top quark production. Cut-based analyses and advanced multivariate techniques are used to extract the signal from large backgrounds and to improve the sensitivity of measurements. As an example of the performance gain obtained with multivariate methods, the DØ single top quark search is presented. Likelihood discriminants, neural networks and decision trees have similar sensitivity, much improved over the cut-based analysis. Boosted decision trees may improve results even more.

International Workshop on Top Quark Physics
January 12-15, 2006
Coimbra, Portugal

[*]Speaker.
[†]on behalf of the DØ and CDF collaborations

PoS(TOP2006)016

1. Motivation

The top quark was discovered in 1995 at the Fermilab Tevatron collider in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV [1]. By far the heaviest elementary particle found to date ($m_t = 172.7 \pm 2.9$ GeV [2]), the top quark plays a central role in the standard model of particle physics (SM).

At the Tevatron, top quarks are mainly produced in top-antitop pairs through the strong interaction. This is the mode in which the top quark was discovered and the only one observed until now. The SM also predicts the production of single top quarks through the electroweak interaction. Given its large mass the top quark decays before hadronization, predominantly into a W boson and a b quark.

Measuring precisely some of the properties of the top quark is one of the main goals of the CDF and DØ collaborations. A top quark could decay into exotic particles such as a charged Higgs boson [3], which would lead to a $t\bar{t}$ production cross section measurement different from the SM prediction. The mass of the top quark has a significant impact on the predicted mass of the Higgs boson [4] and its precise knowledge constrains the SM as well as other models beyond the SM.

If observed (and both the DØ and CDF collaborations are actively searching for this production mode) single top quark production would allow the first direct measurement of the electroweak coupling strength of the top quark as well as constraints on the SM and its extensions.

The potential consequences of precise measurements of top quark properties have nurtured many different approaches to the various facets of top physics [5], some of which are presented in these Proceedings [6].

Some analyses use regular cut-based techniques, in which all events have to satisfy a list of selection criteria (e.g., a jet transverse momentum above a certain threshold). The final analysis result (top mass, production cross section, etc.) may then be derived by simply counting how many data events satisfy all criteria and comparing it to the expected signal and background contributions. It may also be obtained by comparing a discriminating variable distribution to Monte Carlo templates generated for different values of the parameter to be measured. The template giving the best match to data provides the parameter measurement.

In many cases, however, cut-based analyses are limited in their reach. First, such techniques reject many events on the sole basis of one particular variable, hence limiting their statistical power when such events may indeed look very much signal-like except for this one variable. Using a multivariate approach would give a better estimation of whether such events are signal-like or not.

Data is in essence multivariate and combining all relevant information about an event is bound to give at least the same level of knowledge as a single criterion. It can help separating signal from background when event characteristics are very similar. It also allows to make use of all available measurements to extract more information about the events that are selected, to increase the statistical power of the measurement or to increase the signal acceptance (in particular for a search analysis).

Several multivariate techniques are used by the CDF and DØ collaboration top quark groups. They include matrix element calculations, kernel density estimation, dynamic likelihood and neural networks. Their use is documented elsewhere [5, 6].

In order to illustrate how multivariate techniques improve the reach of top quark analyses at the Tevatron, the DØ collaboration single top quark searches are described in the following

sections. Section 2 is a reminder that whatever technique is to be used, its power relies first of all on the quality of its inputs. The benchmark for all analyses is an optimised cut-based analysis presented in Section 3. Commonly used likelihood discriminants and neural networks are described in Sections 4 and 5. Another powerful technique not yet very popular in high energy physics, decision trees, is introduced in Section 6 as well as a way to improve its performance with boosting. A qualitative comparison of the different techniques and prospects for future results are presented in Section 7.

2. Dataset preparation

No matter how sophisticated an analysis scheme is, it will only perform as well as its inputs permit. The emphasis therefore has to be first on the quality of data and Monte Carlo events. After events are reconstructed, identification algorithms are evaluated and their performance modelled. Good object identification efficiency and rejection against fake objects are necessary in order to be able to perform a good selection of data based on the characteristics of signal events compared to background events. Monte Carlo events require an accurate representation of the detector material and of the readout chain, as well as reliable models for the different physics processes under study.

Once good data samples and realistic Monte Carlo events are available, analyses rely on discriminating variables and their correlations to separate signal from background. Examples of such variables for the single top quark search are given in Fig. 1. The histograms show how the background model and its different components describe very well the data after event selection and before applying *b*-tagging. At this stage the single top quark analysis is dominated by background (the signal curves in Fig. 1 are scaled up by a factor ten). These figures confirm that the model agrees with data both in overall normalisation and in variable shapes.

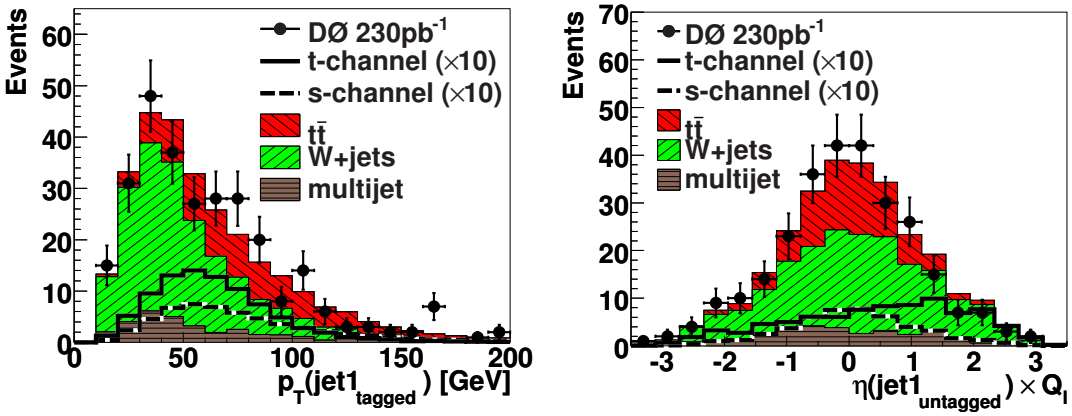


Figure 1: Comparison of signals, backgrounds and data after selection for the DØ single top quark search. The transverse momentum for the leading *b*-tagged jet is shown to the left and the pseudorapidity of the leading untagged jet multiplied by the lepton charge to the right.

3. Cut-based analysis

The technique used is a random grid search. The first step is to find the best cut point for each discriminating variable separately. Once an ordered list of variables is found (ordered by

the expected limit they yield), sets of variables are formed, starting with the most powerful one and adding one by one the other variables. For each set the optimal cut point of each variable is reoptimised. The set of variables that gives the lowest expected limit is chosen.

Limits obtained using this technique are shown in Table 1 [7]. They improve on the limits derived after the initial selection but are still far from the estimated SM production cross sections.

	s-channel	t-channel		
SM prediction	$0.88^{+0.07}_{-0.06}$ pb		$1.98^{+0.23}_{-0.18}$ pb	
	Expected limits		Observed limits	
	s-channel	t-channel	s-channel	t-channel
Initial selection	14.5	16.5	13.0	13.6
Cut-based	9.8	12.4	10.6	11.3

Table 1: Expected SM production cross sections. Expected and observed upper limits (in picobarns) at the 95% confidence level on the production cross sections of single top quarks after event selection and with the cut-based analysis. Results correspond to 230 pb^{-1} of analysed data collected with the DØ detector.

4. Likelihood discriminants

After the event selection, instead of using criteria on discriminating variables, a final discriminating variable is constructed from these variables to characterise events, using the shape of the different signal and background variable distributions. The variables used are essentially uncorrelated.

Consider a vector of measurements $\vec{x} = \{x_i\}$ for the different discriminating variables x_i. The likelihood of the event is given by:

$$\mathscr{L}(\vec{x}) = \frac{\mathscr{P}_{\text{signal}}(\vec{x})}{\mathscr{P}_{\text{signal}}(\vec{x}) + \mathscr{P}_{\text{background}}(\vec{x})},$$

where $\mathscr{P}_{\text{signal}}(\vec{x})$ and $\mathscr{P}_{\text{background}}(\vec{x})$ are the probability density functions for signal and background events, respectively. Signal events tend to have a value of $\mathscr{L}$ close to 1 while it is close to 0 for background events.

The probability density functions $\mathscr{P}_{\text{signal}}(\vec{x})$ and $\mathscr{P}_{\text{background}}(\vec{x})$ are determined from the product of Monte Carlo one-dimensional distributions of the input variables (therefore potential correlations between variables are not taken into account):

$$\mathscr{P}_{\text{signal}}(\vec{x}) = \prod_{i}^{N_{\text{variables}}} P_{\text{signal}}(x_i), \quad \mathscr{P}_{\text{background}}(\vec{x}) = \sum_{j}^{N_{\text{backgrounds}}} f_j \prod_{i}^{N_{\text{variables}}} P_{j\ \text{background}}(x_i),$$

where f_j are constant normalisation factors for the various background contributions.

Different such likelihood discriminants were built for two kinds of main backgrounds ($t\bar{t}$ and W+jets), each channel (s and t), each lepton channel (electron and muon) and on events with exactly one b-tagged jet or with two b-tagged jets, leading to a total of 16 likelihood discriminant variables. Each discriminant uses between seven and ten input variables. The output of such a discriminant is shown in Fig. 2.

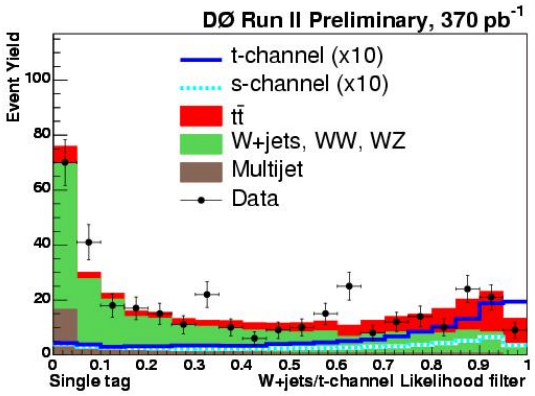


Figure 2: Comparison of signals, backgrounds and data for one of the likelihood discriminant distributions in the DØ single top quark search. The discriminant shown is for the t-channel/W+jets filter with a single b-tagged jet.

Production cross section limits derived from the combination of these discriminants are given in Table 2 [8]. They are the world's best limits to date. They represent a big improvement over the cut-based analysis results of Table 1, although part of the improvement is due to the increased dataset size. An interesting feature of this technique is that it doesn't require any training. Only templates for signal and background discriminating variables are necessary.

	Expected limits		Observed limits	
	s-channel	t-channel	s-channel	t-channel
Likelihood	3.3	4.3	5.0	4.4

Table 2: Expected and observed upper limits (in picobarns) at the 95% confidence level on the production cross sections of single top quarks with the likelihood discriminants analysis. Results correspond to 370 pb^{-1} of analysed data collected with the DØ detector.

5. Neural networks

Neural networks are a widely used technique in particle physics. The same software used by DØ in the Run I single top quark search is used in this analysis: the Multi-Layer Perceptron fit (MLPfit) package [9].

The structure of the network consists of a layer of input nodes, a single layer of hidden nodes and one output node. The input layer is made of one node for each discriminating variable x_i. Each of the input nodes is connected to every hidden node. A hidden node n_k is described by a sigmoid of the input nodes:

$$n_k = \frac{1}{1+\exp^{-\sum w_{ik}x_i}},$$

where w_{ik} is the importance of the contribution of variable x_i to node n_k. The output node is in turn the linear combination of the hidden nodes, $O = \sum w_k n_k$.

The network is described by the sets of weights w_{ik} and w_k, which are determined through a training process. During the first cycle (epoch) of training all weights are initialised randomly.

Running signal and background training events through the network, the difference between the network output O^{observed} and the desired output O^{desired} (1 for signal, 0 for background) is calculated (taking into account the event weight W_j):

$$\text{Error} = \sum_j^{N_{\text{events}}} W_j^2 (O_j^{\text{desired}} - O_j^{\text{observed}})^2.$$

Network parameters are adjusted in order to reduce the value of the error function, and the next epoch begins. After each epoch, the network is run on a set of independent signal and background events and the testing error is computed. The procedure is repeated until the testing error reaches a minimum (see Fig. 3). This early stopping technique avoids overtraining, which happens when the training error improves while the testing error starts to go up again. At this point the network loses part of its generalisation power as it has learnt too much about specific events in the training sample.

For the DØ single top quark search 60% of events were used for training and the remaining 40% for testing. Several networks were trained, using MLPfit hybrid method 7, for different combinations of signal and background channels, like in the likelihood discriminant analysis in Section 4. In order to regularise certain distributions, the logarithm of all nonangular variables were used as inputs. Several rounds of optimisation were necessary to choose the optimal list of variables (11 variables were chosen for each network), the best number of hidden nodes (found to be close to 30 in all cases) and the number of training epochs (between 150 and 250 for each network).

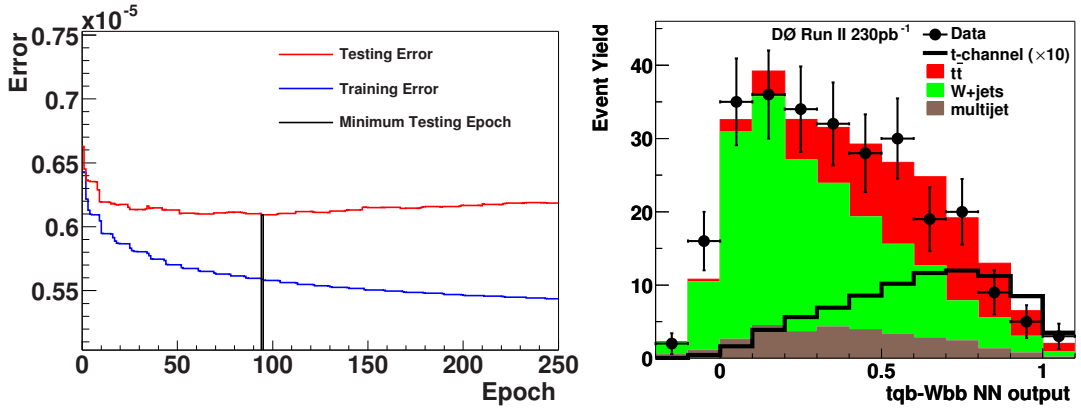


Figure 3: Left: Neural network training and testing errors as a function of training epoch in the DØ single top quark search. The vertical black line shows the position of the minimum of the testing curve, defining the network to be used in the analysis. Right: Comparison of signal, backgrounds and data for one of the neural network outputs in the DØ single top quark search. The network shown is for the t-channel/Wbb filter.

An example of network output is given in Fig. 3. The limits derived from these outputs are shown in Table 3 [10]. This result represents a factor 2 improvement over the cut-based analysis on the same dataset.

Neural networks are a powerful method to separate signal from background. Some of the drawbacks are that they are relatively slow to train and that the set of weights is sensitive to the training events (different training samples may lead to different sets of weights, although they may

	Expected limits		Observed limits	
	s-channel	*t*-channel	*s*-channel	*t*-channel
Neural network	4.5	5.8	6.4	5.0
Decision tree	4.5	6.4	8.3	8.1

Table 3: Expected and observed upper limits (in picobarns) at the 95% confidence level on the production cross sections of single top quarks with the neural network and decision tree analyses. Results correspond to 230 pb^{-1} of analysed data collected with the DØ detector.

have similar performance). Neural networks are also sensitive to the input variables: using too many variables can degrade the performance of the network.

6. Decision trees

Decision trees are a machine learning technique not (yet) commonly used in high energy physics, although it has been widely used in the social sciences [11]. The goal is to extend a simple cut-based analysis into a multivariate technique by continuing to analyse events that fail a particular criterion. The decision tree building algorithm is described in Section 6.1, followed in Section 6.2 by a look at the different parameters influencing the tree construction. Decision tree performance is discussed in Section 6.3. A novel technique to improve the performance of decision trees, boosting, is introduced in Section 6.4.

6.1 Algorithm

Mathematically, decision trees are rooted binary trees. An example is shown in Fig. 4. Consider a training sample made of known signal and background events: they form the root node of the tree. Given a list of variables $\{x_i\}$, all events are sorted in turn according to each variable. For each x_i the splitting value that gives the best separation of the events into two child nodes — one with mostly signal events, the other with mostly background events — is found (see Section 6.2 for details). The variable and split value giving the best separation are selected and two new nodes are created, one corresponding to events satisfying the split criterion (labelled (P)assed in Fig. 4), the other containing events that failed it (labelled F).

The algorithm is then applied recursively to the two child nodes. When the splitting stops, the terminal node is called a leaf, with an associated purity (weighted signal fraction of the training sample in this node).

When a new event is passed through the tree, its properties are compared to the criterion at each node until it reaches a leaf. For instance on the example in Fig. 4 the event will go right after the root node if $H_T > 212$ GeV, and right again if $p_T < 31.6$ GeV. It would then have reached a leaf and the output of the tree for this event is the leaf purity.

6.2 Tree parameters and node splitting

Several internal parameters can influence the development of a decision tree. Normalisation of the total signal and background training event weights was taken to be the same, equal to 1. Varying it had little effect on the separation power.

A terminal node has to be labelled signal or background during training. A leaf was chosen to be signal if its purity was greater than 0.5, and background otherwise.

Criteria to decide when to stop the splitting procedure were chosen so as to guarantee statistical significance. A node is not split if the best split would create a child with less than 100 events, or if the improvement in separation is negligible.

The most important part in the decision tree building process is how to take the decision to split a node. This requires a good list of candidates, that is, discriminating variables, and an automated way to decide what the best split is.

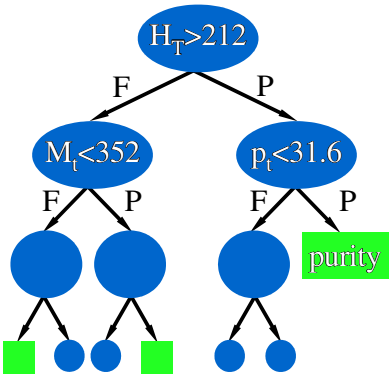
Consider an impurity measure $i(t)$ for node t. Desirable features of such a function are that it should be maximal for an equal mix of signal and background (no separation), minimal for nodes with either only signal or only background events (perfect separation), symmetric in signal and background purity, and strictly concave in order to reward purer nodes.

Several such functions exist, like entropy or Gini:

$$\text{entropy} = -\sum_{i=s,b} p_i \log p_i, \quad \text{Gini} = 1 - \sum_{i=s,b} p_i^2 = \frac{2sb}{(s+b)^2},$$

where s (b) is the sum of signal (background) weights, $p_s = s/(s+b)$ (purity) and $p_b = b/(s+b)$. In the DØ analysis Gini was chosen, but entropy gives very comparable results.

One can now define the decrease of impurity (goodness of split) associated to a split S of node t into children t_P and t_F, $\Delta i(S,t) = i(t) - p_P \cdot i(t_P) - p_F \cdot i(t_F)$, where p_P (p_F) is the fraction of events that passed (failed) split S. The goal is to find the split S^* that maximises the decrease of impurity, which corresponds to finding the split that minimises the overall tree impurity.



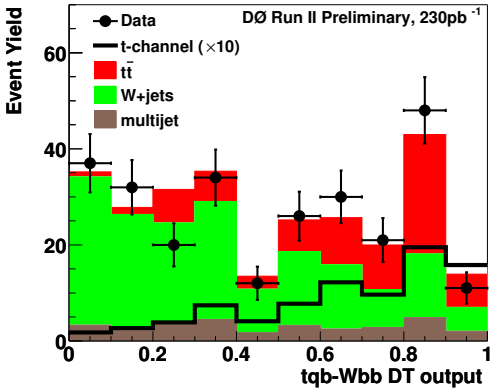


Figure 4: Left: Graphical representation of a decision tree. Nodes are in blue, with their associated splitting test; terminal nodes (leaves) are in green. Right: Comparison of signal, backgrounds and data for one of the decision tree outputs in the DØ single top quark search. The decision tree shown is for the t-channel/Wbb filter.

6.3 Results

For this analysis [7] the same strategy as for the neural network described in Section 5 was used. An example of a decision tree output is shown in Fig. 4 and limits derived from such outputs are reported in Table 3. The sensitivity is similar to the neural network analysis.

Some limitations of decision trees are the instability of the tree structure with respect to the training sample composition and the piecewise nature of the output. Training on different samples may produce very different trees with similar separation power. The discrete output is due to the fact that the only possible values are the purities of all leaves.

Decision tree techniques also have very interesting features. A decision tree is not a "black box": it has a human-readable structure, making it possible to know why a particular event was labelled signal or background. Training is very fast compared to neural networks. It can deal with discrete variables directly and no preprocessing of input variables is necessary: any monotonic transformation of the discriminating variables would yield exactly the same result. Finally it is relatively insensitive to extra variables: unlike neural networks, adding variables that are not very powerful does not degrade the performance of the decision tree.

6.4 Boosting

A very powerful technique to improve the performance of any weak classifier (anything that does better than random guess) was introduced a decade ago: boosting [12]. It was recently used in high energy physics with decision trees by the MiniBooNe experiment [13].

The basic principal of boosted decision trees is to train a tree T_n, minimise some error function and create a tree T_{n+1} as a modification of tree T_n. The specific algorithm used in the DØ single top quark search is AdaBoost [12], or adaptive boosting.

Once a tree T_n is built, its associated error is computed as proportional to the weight of misclassified events: $\text{err}_n = \left(\sum_i w_i \times \text{isMisclassified}_n(i)\right) / \sum_i w_i$ and the tree weight is $\alpha_n = \beta \times \ln((1-\text{err}_n)/\text{err}_n)$ where β is the boosting parameter. For each misclassified event, its weight w_i is multiplied by e^{α_n}, hence giving more weight to misclassified events, on which the next tree will have to work harder to classify them properly. The boosted decision tree result for event i is $T(i) = \sum_{n=1}^{N_{\text{tree}}} \alpha_n T_n(i)$.

Boosting usually improves performance and preliminary observations in the DØ single top quark search confirm this. Another advantage of boosted decision trees is that the piecewise nature of decision tree outputs is diluted by the averaging of the boosted result.

7. Summary and outlook

In order to extract as much information as possible from the Run II dataset currently collected by the CDF and DØ experiments at the Fermilab Tevatron collider, both top quark study groups use many different analysis techniques. Their goals include precision measurements of the $t\bar{t}$ production cross section and of the top quark mass, measurements of other top quark properties and, if it exists, observation of the electroweak production of single top quarks.

All techniques rely on the availability of quality data and realistic Monte Carlo simulations of the different physics processes and detector effects. Only then can one make optimal use of powerful analysis methods.

To illustrate the gain achievable thanks to advanced multivariate techniques, the DØ single top quark searches were presented. Cut-based analyses are the benchmark of all studies and a random grid search was used in this case.

Likelihood discriminants, neural networks and decision trees were also used and have comparable sensitivity, yielding upper limits on the single top quark production cross section a factor two better than the cut-based result. Boosted decision trees were also introduced. Results should become available soon; they are expected to increase the measurement sensitivity. More results are expected in the near future with more data analysed and refined optimisation of the different analysis techniques.

References

[1] S. Abachi *et al.* [DØ Collaboration], *Observation of the top quark*, Phys. Rev. Lett. **74**, 2632 (1995), [hep-ex/9503003];
F. Abe *et al.* [CDF Collaboration], *Observation of top quark production in $\bar{p}p$ collisions*, Phys. Rev. Lett. **74**, 2626 (1995), [hep-ex/9503002].

[2] CDF and DØ Collaborations, TEVEWWG, *Combination of CDF and DØ results on the top-quark mass*, hep-ex/0507091.

[3] J.F. Gunion, H.E. Haber, G.L. Kane and S. Dawson, *The Higgs Hunter's Guide*, Perseus Publishing (2000)

[4] V.M. Abazov *et al.* [DØ Collaboration], *A precision measurement of the mass of the top quark*, Nature **429**, 638 (2004) [hep-ex/0406031].

[5] See http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/top_public.html for all top quark analysis results from the DØ collaboration and
http://www-cdf.fnal.gov/physics/new/top/top.html for results from the CDF collaboration.

[6] See the following contributions to these proceedings:
R. Rossin, *Measurements of top quark pair production cross section & search for resonances*
E. Varnes, *Measurements of top quark decay properties*
J. Cammin, *Precision measurement of top quark mass in lepton+jets channel*
B. Jayatilaka, *Precision measurement of top quark mass in dilepton channel*
M. Begel, *Search for single top quark production at the Tevatron*.

[7] DØ Collaboration, DØ Note 4722-CONF, http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/conference_notes/spring05_singletop_svt_confnote.pdf

[8] DØ Collaboration, DØ Note 4871-CONF, http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/conference_notes/summer05_singletop_jlip_confnote.pdf

[9] MLPfit package, J. Schwindling, http://schwind.home.cern.ch/schwind/MLPfit.html.

[10] V.M. Abazov *et al.* [DØ Collaboration], *Search for single top quark production in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV*, Phys. Lett. B **622**, 265 (2005), [hep-ex/0505063].

[11] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone , *Classification and Regression Trees*, Wadsworth (1984).

[12] Y. Freund and R.E. Schapire, *Experiments with a new boosting algorithm*, in *Machine Learning: Proceedings of the Thirteenth International Conference*, pp 148-156 (1996).

[13] B. P. Roe, H. J. Yang, J. Zhu, Y. Liu, I. Stancu and G. McGregor, *Boosted decision trees as an alternative to artificial neural networks for particle identification*, Nucl. Instrum. Meth. A **543**, 577 (2005), [physics/0408124].