Fermi National Accelerator Laboratory

Search for the Top Quark at D0 Using Multivariate Methods

Pushpalatha C. Bhat

For the D0 Collaboration

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

July 1995

Disclaimer

Search for the Top Quark at DØ using Multivariate Methods

Pushpalatha C. Bhat
for the DØ Collaboration

Fermi National Accelerator Laboratory,
Batavia, IL 60510, U.S.A.

We report on the search for the top quark in $p\bar{p}$ collisions at the Fermilab Tevatron ($\sqrt{s} = 1.8\ TeV$) in the di-lepton and lepton+jets channels using multivariate methods. An H-matrix analysis of the $e\mu$ data corresponding to an integrated luminosity of $13.5 \pm 1.6\ pb^{-1}$ yields one event whose likelihood to be a top quark event, assuming $m_{top} = 180\ \mathrm{GeV/c}^2$, is ten times more than that of WW and eighteen times more than that of $Z \to \tau\tau$. A neural network analysis of the e+jets channel using a data sample corresponding to an integrated luminosity of $47.9 \pm 5.7\ pb^{-1}$ shows an excess of events in the signal region and yields a cross-section for $t\bar{t}$ production of 6.7 ± 2.3 (stat.) pb, assuming a top mass of $200\ \mathrm{GeV/c}^2$. An analysis of the e+jets data using the probability density estimation method yields a cross-section that is consistent with the above result.

INTRODUCTION

The top quark that remained elusive for over a decade and a half has finally been observed by both the CDF and DØ collaborations (1,2). The top quark events have been observed in the di-lepton and lepton+jets decay modes of $t\bar{t}$ pairs produced in $p\bar{p}$ collisions at $\sqrt{s} = 1.8\ TeV$ at the Fermilab Tevatron. The collaborations have used conventional analysis methods to optimize cuts on kinematic variables together with the tagging of the b quarks, to discriminate top quark events from background. The conventional analysis methods do not exploit correlations amongst the variables on which the cuts are applied and thus may suffer a loss in signal efficiency. The DØ collaboration has been applying multivariate methods such as the H-matrix, probability density estimation (PDE) and neural networks for identifying top quark events (3,4), to improve the signal efficiency. In this paper, we describe the multivariate methods used, we present an analysis of the channel $t\bar{t} \to e\mu$ and report on the measurement of $t\bar{t}$ production cross-section from a study of the channel $t\bar{t} \to e$+jets.

MULTIVARIATE CLASSIFIERS

A classifier is any procedure that assigns objects to classes. In the present context, a classifier would separate signal events from the background. The time-honored conventional classification methods of examining uni-variate (1-dimensional) and bi-variate (2-dimensional) distributions of variables to optimize cuts for separating signal and background events do not in general provide the maximum possible discrimination when correlations exist between variables. Multivariate classifiers which fully exploit the correlations that exist among several variables provide a discriminating boundary between signal and background in multi-dimensional space that can yield discrimination close to the theoretical maximum (Bayes' limit (5)).

In the multivariate approach, one encodes each event as a point in a multi-dimensional space, called *feature space*, corresponding to a vector x of feature variables such as electron E_T (E_T^e), neutrino E_T, ($\not{E}_T$), H_T ($\Sigma E_T(jets)$), *etc.* This feature space is then mapped into a one or a few-dimensional output space in such a way that the signal and background vectors are mapped onto different regions of the output space. The aim of the multivariate methods is to reduce the dimensionality of the problem without losing information in the process. The optimal way to partition the feature space into signal and background regions is to choose the mapping to be the Bayes discriminant function. Each cut on the value of the function corresponds to a discriminating boundary in feature space. The Bayes discriminant function is simply the ratio of the probability $P(s|x)$ that a given event is a signal event and the probability $P(b|x)$ that it is a background event. It is written as

$$R(x) = \frac{P(s|x)}{P(b|x)} = \frac{P(x|s)P(s)}{P(x|b)P(b)}. \tag{1}$$

The quantities $P(x|s), P(x|b)$ are the likelihood functions for the signal and background, respectively (hereafter denoted as $f(x)$ with or without appropriate subscript). The ratio of the prior probabilities $\frac{P(s)}{P(b)}$ is the ratio of the signal and background cross-sections. Some multivariate classifiers approximate the likelihood functions while the neural network classifier arrives at the Bayesian probability for the signal, $P(s|x)$, without calculating the likelihood functions for each class separately. The three classifiers being used by DØ are described in the following sections.

$H-$matrix Method

This is the familiar covariance matrix method which is also known as the Gaussian Classifier. It was introduced in the 1930s (6,7) as a tool for discriminating one class of feature vector x from another. The vector x is assumed to be distributed according to a multivariate Gaussian with covariance matrix M and mean $\bar{x}$. The likelihood function is therefore,

$$f(x) = A \cdot \exp\{-\frac{1}{2}\sum_{i,j}(x_i - \bar{x}_i)^T (M^{-1})_{ij}(x_j - \bar{x}_j)\} \equiv A \cdot \exp(-\chi^2) \qquad (2)$$

where $(M^{-1})_{ij}$ is the H-matrix. Fisher (6) showed that the optimal way to separate two overlapping multivariate Gaussian distributions with a common covariance matrix but with different means $\bar{x}_s$ and $\bar{x}_b$ is to cut on the function,

$$F = \frac{1}{2}(\chi_b^2 - \chi_s^2); \qquad (3)$$

F is called the Fisher linear discriminant function. If the two distributions have different correlation matrices, one can introduce a more general Gaussian classifier (8), where the χ^2 values are calculated using the corresponding H-matrices as well as mean values for the signal and background classes. We note that this method is useful even when the distribution of x is non-Gaussian.

The Bayes discriminant function $R(x)$ can be written in terms of the Fisher variable F as $R(x) = \exp(F)$ when $P(s) = P(b)$.

Probability Density Estimation (PDE) Method

In the PDE method (9), the likelihood functions or the probability density functions (pdf's) are approximated by summing over multivariate kernel functions with one kernel function centered at each data point for the two classes of events. The expression for the likelihood function is,

$$f(x) = \frac{1}{N_{events}h_1...h_d} \sum_{i=1}^{N_{events}} \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right), \qquad (4)$$

where the kernel function K is chosen to be a Gaussian. The jth variable is denoted by x_j and, x_{ij} denotes the jth variable of the ith event. By appropriate transformation, the variables x_j are rendered uncorrelated within the signal and background classes. The quantity h_j is the jth smoothing parameter. We use a single "global" smoothing parameter h defined by

$$h_{sj} = h\sigma_{sj}, \qquad h_{bj} = h\sigma_{bj}, \qquad (5)$$

where σ_{sj} and σ_{bj} are the estimated standard deviations of the jth variable for the signal and background classes, respectively. The value of the smoothing parameter h is set by maximizing the signal to background ratio (S/B) at the required signal efficiency.

The discriminant function we use in the PDE method is

$$D(x) = \frac{f_s(x)}{f_s(x) + f_b(x)} \qquad (6)$$

where $f_s(x)$ and $f_b(x)$ are the pdf's for signal and background classes of events, respectively. The function $D(x)$ approximates the Bayesian probability for the signal $P(s|x)$. When $P(s) = P(b)$, the Bayes discriminant function becomes $R(x) = D(x)/(1 - D(x))$.

Neural Networks

Artificial neural networks provide a powerful new paradigm for event classification. The most commonly used architecture in classification problems is the multi-layer perceptron or feed-forward neural network. In Fig. 1, we show the representation of a three layer feed-forward neural network with one hidden layer. The nodes in the input layer correspond to the components x_k of the feature vector x, and the output layer has a single node commonly used in binary classification problems. The network builds an internal representation of the mapping of the feature space into the output space. The output of the network is given by

$$O(x) = g(\sum_j w_j g(\sum_k w_{jk} x_k + \theta_j) + \theta),\tag{7}$$

where the "weights" w_{jk} and w_j and, the "thresholds" θ_j and θ are parameters that are adjusted during the "training" process. The quantity g is a non-linear "transfer" function of the form $g(y) = 1/(1 + e^{-2y})$. (Use of such transfer functions enables the mapping of any real function (10).) The parameters are determined by minimizing the mean square error between the actual output O^p and the desired output t^p

$$E = \frac{1}{2N_p}\sum_{p=1}^{N_p}(O^p - t^p)^2\tag{8}$$

with respect to the parameters. Here p denotes a feature vector or *pattern*. Once the parameters are determined using a large number of signal and background events the network can be used to classify events. It has been shown (11) that the feed-forward neural network when trained as a classifier using the back-propagation algorithm for updating the parameters, yields an output that approximates the Bayesian probability for the signal *i.e.*, $O(x) = P(s|x)$. (This assumes t^p is 1 for signal and 0 for background.) The Bayes discriminant in terms of the network output will be $R(x) = O(x)/(1 - O(x))$.

ANALYSIS OF THE DØ DATA

We have applied the multivariate methods to the analysis of the $t\bar{t} \to e\mu$ and $t\bar{t} \to e+$jets channels. The on-line trigger selection, off-line electron and muon identification criteria and description of variables used here can be found elsewhere (12,13).

The $e\mu$ data analysed here correspond to an integrated luminosity of 13.5 ± 1.6 pb^{-1}. The overall trigger efficiency is about $(90\pm7)\%$ for $m_{top}=$ 180 GeV/c^2 and varies slightly with top mass. In the off-line selection, before analysis with the H-matrix method, we apply loose electron and muon identification criteria and require $E_T^e >$11 GeV and P$_T^\mu >$11 GeV/c and at least two reconstructed jets ($E_T^{jet} > 8$ GeV). Using the H-matrix method we
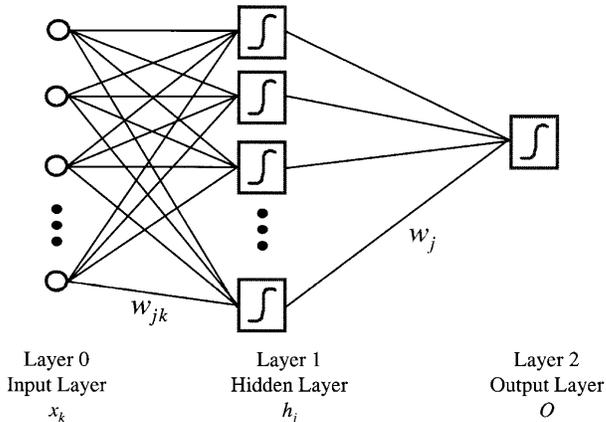
FIG. 1. A feed-forward neural network with one hidden layer.

have examined the signal to background ratio with respect to $WW \to e\mu$ and $Z \to \tau\tau \to e\mu$ (the dominant backgrounds).

We have used the PDE and neural network methods to analyze e+jets data corresponding to an integrated luminosity of 47.9±5.7 pb^{-1}. The dominant background to the $t\bar{t} \to e$+jets channel is from the QCD production of W+multi-jets where the signature is the same as that of the signal, *viz.* a high P_T electron, high $\not{E}_T$, arising from the leptonic decay of the W boson, and several jets. In addition, we have background from QCD multi-jet events where one of the jets is mis-identified as an electron and the event also has a high $\not{E}_T$ from mis-measurements as well as neutrinos from any heavy flavor decays. We refer to this background as QCD fakes. The two backgrounds in our data prior to the multivariate analyses are estimated directly from data. The QCD fake background is estimated by the joint probability of multi-jet events having $\not{E}_T$ larger than the cut applied and a jet being mis-identified as an electron. The W+jets background is estimated using Berends' scaling (14). The inclusive jet multiplicity data (after subtracting the QCD fakes background) is fitted for this "jet-scaling" allowing for contribution from top quark events. These background numbers are multiplied by the fraction of events surviving the multivariate cuts to get the final background estimates.

H-matrix Analysis of $e\mu$ Data

We have chosen the following feature vector $x=(E_T^e,\ P_T^\mu,\ E_T^{jet1}, E_T^{jet2},\ \not{E}_T^{cal},\ H_T,\ M_{e\mu},\ \Delta\phi_{e\mu})$ to build the signal and background H-matrices, where $\not{E}_T^{cal}$ is the $\not{E}_T$ in the calorimeter, $M_{e\mu}$ is the invariant mass of the two leptons and $\Delta\phi_{e\mu}$ is the azimuthal angle between the two leptons. We have built the signal H-matrix H_{Top} using 180 GeV $t\bar{t}$ ISAJET Monte Carlo events

($t\bar{t}180$) processed through the DØ detector simulation program. Since in the $e\mu$ data we expect very few $t\bar{t} \rightarrow e\mu$ events, we have chosen to use the data to model the background. The data consist largely of QCD $b\bar{b}$ events and $W \rightarrow \mu$+jets+e$'$ events (where e$'$ denotes a jet that fakes an electron). We have considered the $Z \rightarrow \tau\tau$ background separately to get better rejection against that background. We define two Fisher discriminant functions

$$F_1 = \frac{1}{2}(\chi^2_{Data} - \chi^2_{Top}), \qquad F_2 = \frac{1}{2}(\chi^2_Z - \chi^2_{Top}). \qquad (9)$$

where,

$$\chi^2_{Data} = \sum_{i,j}(x_i - \bar{x}_i)^T H_{Data}(x_j - \bar{x}_j), \quad \chi^2_Z = \sum_{i,j}(x_i - \bar{x}_i)^T H_Z(x_j - \bar{x}_j),$$

$$(10)$$

$$\chi^2_{Top} = \sum_{i,j}(x_i - \bar{x}_i)^T H_{Top}(x_j - \bar{x}_j), \qquad (11)$$

where H_{Data} and H_Z are the background H-matrices built using data and $Z \rightarrow \tau\tau$ Monte Carlo events, respectively. The χ^2 values, F_1 and F_2 are determined for signal, backgrounds and for data. In Fig. 2 we show the lego plots of F_1 *vs* F_2 for each of the samples. By applying the cuts $F_1 > 15$ and $F_2 > 3$ we have 16%, 22% and 25% efficiency for top events with top masses of 140, 160 and 180 GeV/c^2, respectively. The only event that survives the cuts is the same as that found in the conventional analysis (15). This event lies in a region of phase space where the the signal to background ratio (S/B) is about 18 with respect to $Z \rightarrow \tau\tau$ and 10 with respect to WW for a 180 GeV/c^2 top quark.

PDE Analysis of $e+$ jets Data

The PDE method has been applied to the $e+ \geq 3$jets data (4). The selection criteria used are $E_T^e > 20$ GeV, $\not{E}_T > 20$ GeV and at least 3 jets with $E_T > 15$ GeV. These five transverse energies define our feature vector in the analysis. The two backgrounds are combined in the ratio estimated as in the conventional analysis and are treated as a single background to build the pdf. Figure 3 shows the distributions of the discriminant function $D(x)$ for background events, 180 GeV/c^2 top quark events and for DØ data. Applying a cut of $D(x) > 0.8$ yields 21 data events with an estimated background of 14.0 ± 1.6 events in 47.9 pb^{-1}. The product of efficiency and branching ratio for $t\bar{t}180$ events is 3.1% (as compared to 1.8% for conventional analysis). The $t\bar{t}$ cross-section is calculated to be 4.7 ± 3.3 (stat.) pb, in agreement with the results from the conventional analysis .
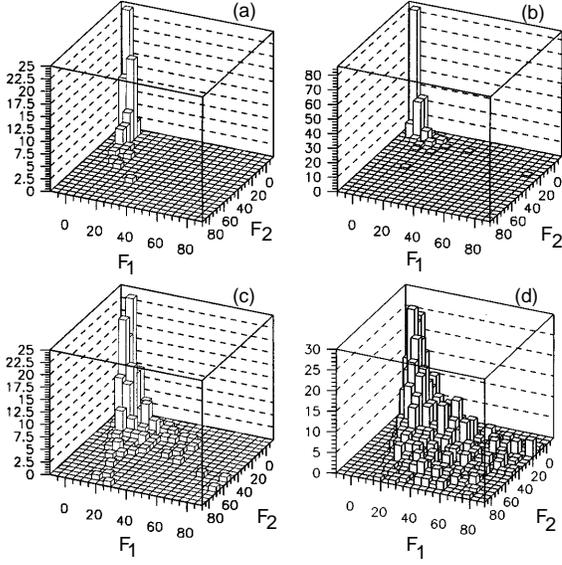
FIG. 2. F_1 vs F_2 from H-matrix analysis of $e\mu$ channel for (a) DØ data ($\int Ldt$=13.5 pb^{-1}), (b) $Z \to \tau\tau$ (($\int Ldt$=3.1 fb^{-1}), (c) WW ($\int Ldt$=22.3 fb^{-1}) and (d) $t\bar{t}$180 ($\int Ldt$=20.1 fb^{-1}) samples.

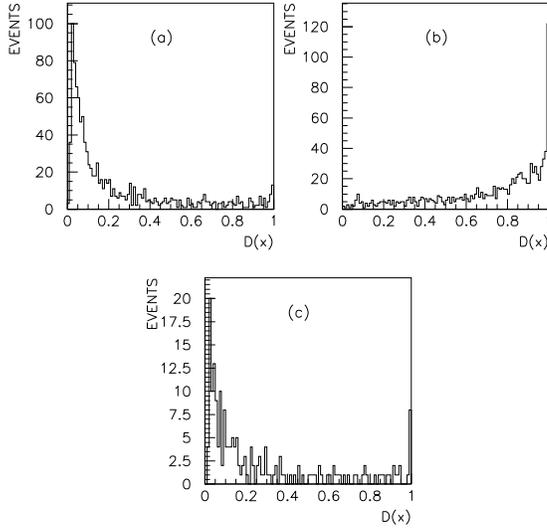**FIG. 3.** The PDE discriminant function for (a)background, (b)$t\bar{t}$180 and for (c)DØ data.

Neural Network Analysis of e+jets Data

A discussion of a two-variable and a six-variable analysis of the $e+ \geq 4$jets data using neural networks has been presented previously (3). Here we present results of the analyses including recent data. The neural network program used here is JETNET 3.0 (16). We use the E_T of the various measured objects in the event ($E_T^e, \not{E}_T, E_T$ of jets), the event shape variable aplanarity (A) and the total transverse energy H_T of central jets (pseudorapidity $|\eta| <2.0$) to discriminate the top signal events from the backgrounds. In our conventional analysis of $e+ \geq 4$jets using non-tagged data we have applied selection cuts of $E_T^e >20$ GeV, $\not{E}_T >25$ GeV, E_T(jet4) >15 GeV, $A >.05$ and $H_T >200$ GeV. (Jets are ordered in decreasing E_T; jet4 refers to the jet with fourth highest E_T.) For demonstration purposes, we compare in Fig. 4 the conventional cuts on A and H_T with the contour cut obtained by a simple network with 2 input nodes, 2 hidden nodes and one output node. The A and H_T are used as inputs and the network is trained on $t\bar{t}$180 and background events (a mixture of W+jets and QCD fakes combined in the proper ratio). The contour provides better signal efficiency than the conventional cuts for the same signal to background ratio.
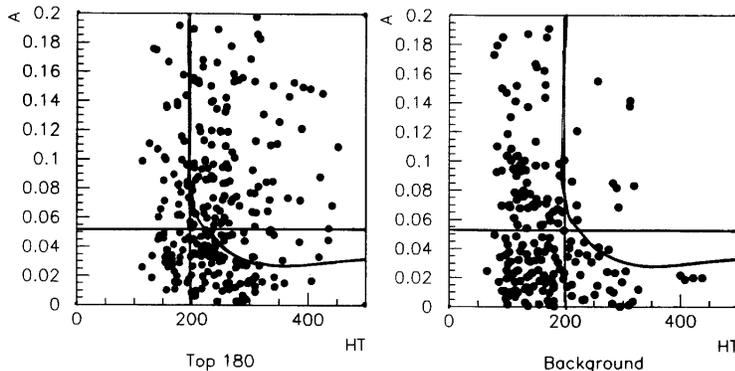

FIG. 4. A and H_T scatter plot for signal and background with conventional cuts (lines parallel to the co-ordinate axes) and neural network cut (the contour).

In order to achieve higher signal efficiency we have relaxed the number of jets required and have carried out an analysis of the $e+ \geq 3$jets data. Also, we use a five-dimensional feature vector $x=(E_T^e, \not{E}_T, H_T, A , E_T$(jet3)). For this analysis, we use two different networks to discriminate against W+jets and QCD fake background separately. That is, we train one network with $t\bar{t}$180 as signal and W+jets Monte Carlo events (using the VECBOS event generator) as background and a second network with $t\bar{t}$180 as signal and QCD fakes (data events that fail the electron ID cuts) as background. We use networks with 5 input nodes (corresponding to the 5-dimensional feature vector), 5 hidden nodes in one hidden layer and 1 output node. We use 1300 $t\bar{t}$ events, 1300 W+jets events and 590 QCD fake events for training. The testing is done on

2400 $t\bar{t}$ events (which include the 1300 events used for training), 1300 W+jets events and 590 QCD fake events that were used for training. Training and testing on the same set of events with the given sample size can give rise to an uncertainty ($\approx 10\%$) in the estimated background which is included in the systematic uncertainty. The target output of the network t^p during training is set to be 1 for the signal and 0 for the background.
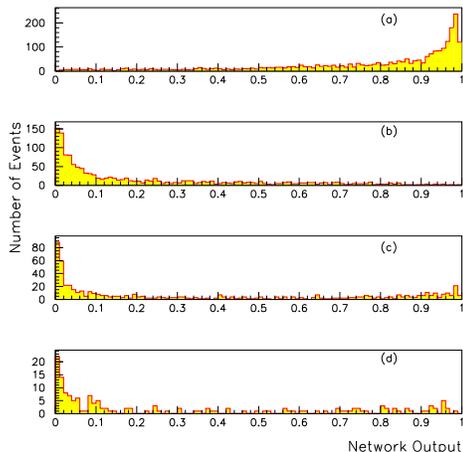

FIG. 5. Distributions of the output from the first network for (a)$t\bar{t}$180, (b)W+jets (VECBOS), (c)QCD fakes and for (d)DØ data.

Figure 5 shows the output of the first network (NN1) for $t\bar{t}$180, W+jets, QCD fakes and for DØ data. The distributions peak close to 1 for signal events and close to 0 for background events, as expected. In Fig. 6, we show the output distributions from the first network for data and for the background (W+jets and QCD fakes combined) normalized to the number of events expected in 47.9 pb^{-1}. We have estimated our background to be $(80\pm11)\%$ W+jets and $(20 \pm5)\%$ QCD fakes. (Errors are statistical only.) The distributions in Fig. 6 are statistically consistent with each other in the background region (NN1 close to 0) and we observe an excess of data events in the signal region.

For the most part, the kinematic distributions of QCD fakes are similar to W+jets and, therefore, all but a small part of QCD fakes can be rejected with the first network. However, to get better rejection of QCD fakes, we process all samples through the second network. In Fig. 7 we show the output of the second network (NN2) for signal, background and data events which satisfy the cut NN1 >0.7. Applying a cut NN2>0.5, we get a factor of three more reduction in the QCD fake background.
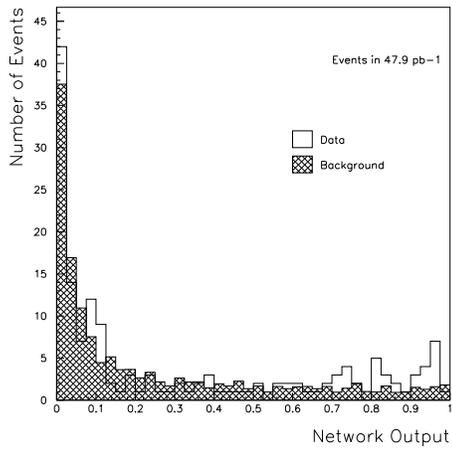
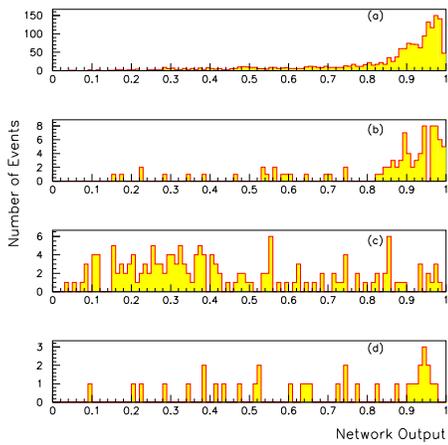**FIG. 6.** Comparison of outputs from the first network for DØ data and background

FIG. 7. Output of the second network (NN2) after requiring NN1>0.7 for (a)$t\bar{t}$180, (b)W+jets (VECBOS), (c)QCD fakes and for (d)DØ data.

We examine the distributions of the five input variables for data and background in the region NN1<0.4, NN2<0.4 noting that only about 5% of the $t\bar{t}180$ events lie in that region. Given that the events in the region are mostly background we can check if our background modeling is correct. The distributions for data and the combined background are compared in Fig. 8. There is good agreement between data and the background model.
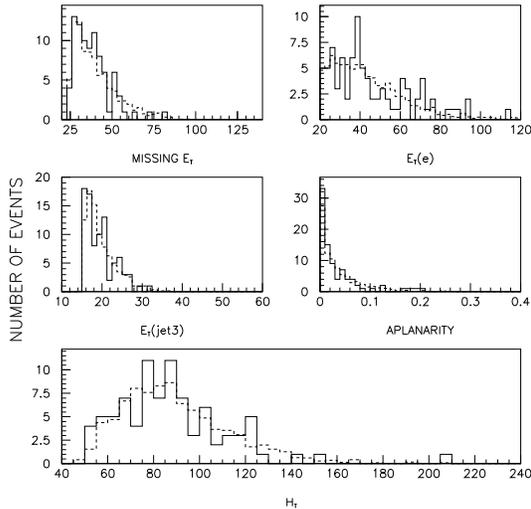

FIG. 8. Distributions of input variables compared for data(solid histograms) and combined backgrounds (dashed histograms) after applying cuts NN1<0.4 and NN2<0.4 (anti-Top cuts).

Applying the cuts NN1>0.7 and NN2>0.5 yields 25 candidate events with an estimated background of 10.1±1.5. This gives an excess over background of 14.9±5.2 events. The 25 candidate events found here include most of the non-tagged and μ-tagged candidate events found by the conventional analysis. The product of efficiency and branching ratios are 4.0% and 4.6% (compared to 1.8 % and 2.4% for conventional analysis) for $t\bar{t}180$ and $t\bar{t}200$, respectively. For a top quark mass of 200 GeV/c^2, we obtain a $t\bar{t}$ production cross-section of 6.7±2.3 pb. (For $t\bar{t}180$, the cross-section obtained is 7.8±2.7 pb.) The errors quoted are statistical only. A preliminary estimate of the systematic uncertainty which includes errors in background estimation and signal efficiency (prior to multivariate analyses) and neural network specific uncertainties is about 30%. This is dominated by the first two components and work is in progress to reduce these uncertainties.

SUMMARY

We have applied multivariate analysis methods to search for top quark events in the DØ data and we find a significant excess of events over background. An H-matrix analysis of $e\mu$ data ($\int Ldt = 13.5 \pm 1.7\ pb^{-1}$) yields one candidate event (same as found in the conventional analysis) which lies in a phase space region where S/B = 10 with respect to WW and S/B = 18 with respect to $Z \rightarrow \tau\tau$. Preliminary results from a PDE analysis of the $e+ \geq$ 3jets data are consistent with results from the conventional analysis. A preliminary neural network analysis of $e+ \geq$ 3jets data yields $\sigma_{t\bar{t}}(m_{top} = 200 GeV/c^2) = 6.7 \pm 2.3$ (stat.) pb in agreement with our published (2) results($\sigma_{t\bar{t}}(m_{top} = 200 GeV/c^2) = 6.3 \pm 2.2\ pb$).

ACKNOWLEDGEMENTS

REFERENCES

1. F. Abe *et al.,* (CDF Collaboration) Phys. Rev. Lett. **74**, 2626 (1995).
2. S. Abachi *et al.,* (DØ Collaboration) Phys. Rev. Lett. **74**, 2632 (1995).
3. Pushpalatha C. Bhat (DØ Collaboration), FERMILAB-CONF-94-261-E. To be published in the proceedings of 1994 Meeting of the American Physical Society, Division of Particles and Fields (DPF 94), Albuquerque, NM, 2-6 Aug 1994.
4. H.E. Miettinen (DØ Collaboration), Proc. AIHENP (1995).
5. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
6. R. A. Fisher, Annals Eugenics 7(1936)179.
7. P.C. Mahalanobis, Proc. Nat. Inst. Sci. India, Part 2A, 49 (1936).
8. M. Kendall, A. Stuart and J.K. Ord, "The Advanced Theory of Statistics", Vol. 3, 4th ed., C. Griffin & Co. Ltd., London.
9. L. Holmstrom, S.R. Sain and H.E. Miettinen, submitted to Comput. Phys. Commun.
10. E.K. Blum and L.K.Li, Neural Networks 4, 511 (1991)
11. D.W. Ruck *et al.,* IEEE Trans. Neural Networks 1, 296 (1990)
12. J. Bantly, these Proceedings.
13. S. Abachi *et al.,* (DØ Collaboration), FEMILIAB-PUB-1995/020-E, submitted to Phys. Rev. D.
14. F.A. Berends, H. Kuijf, B. Tausk and W.T. Giele, Nucl. Phys. B357, 32 (1991).
15. S. Abachi *et al.,* (DØ Collaboration), Phys. Rev. Lett. 72, 2138 (1994).
16. JETNET 3.0 LUND Preprint, LU TP 93-29 (1993).