An Improved Search for Single Top Quarks using Bayesian Neural Networks

The DØ Collaboration
URL http://www-d0.fnal.gov
(Dated: June 1, 2007)

We report the results of an improved Bayesian neural network (BNN) analysis of 0.9 fb$^{-1}$ of DØ data. The observed excess of events over background has a $p$-value of 0.081%, assuming the background-only hypothesis, which corresponds to a significance of 3.1 standard deviations for a Gaussian density. The $p$-value computed using the Standard Model (SM) single top production cross section of 2.9 pb is 1.6%, corresponding to an expected significance of 2.2 standard deviations. Assuming the observed excess is due to single top production, we measure a single top quark production cross section of $\sigma(p\bar{p} \rightarrow tb + X, tqb + X) = 4.4^{+1.6}_{-1.4}$ pb.

Preliminary Results for Spring 2007 Conferences
I. INTRODUCTION

This note reports the results of an improved search for single top quark production in 0.9 fb\(^{-1}\) of DØ data [1]. We observe an excess of events over background which, if interpreted as due to single top production, yields a cross section of

\[ \sigma(p\bar{p} \to t\bar{b} + X, \; tq\bar{b} + X) = 4.4^{+1.6}_{-1.4} \text{ pb}. \]

In obtaining this result, we have assumed the Standard Model ratio between the \( s \) and \( t \) channel single top cross sections.

II. ANALYSIS DETAILS

In this analysis, the single top signal is enhanced relative to background using Bayesian neural networks (BNN) [2, 3]. A BNN is a function that approximates the discriminant

\[ D(x) = \frac{f(x|S)}{f(x|S) + f(x|B)}, \]

where \( f(x|S) \) and \( f(x|B) \) are the probability density functions for signal and background, respectively, and \( x \) denotes the variables that characterize an event. The signal, which comprises \( s \) and \( t \) channel single top, is modeled with \( s \) and \( t \) channel Monte Carlo events mixed in the SM ratio, while the background—comprising \( t\bar{t} \), \( W + \text{jets} \) and light-jet QCD—is modeled with MC events and real QCD events, also mixed in the predicted proportions. In order to maximize the sensitivity to a possible signal, the observed and MC samples were divided into 12 disjoint channels according to lepton flavor, number of \( b \)-tags and jet multiplicity: \( (\mu, e) \times (1\text{-}tag, 2\text{-}tag) \times (2, 3, 4) - \text{jet} \).

For each of the 12 channels, we selected a set of well-modeled variables using the Kolmogorov-Smirnov (K-S) test to assess the goodness of fit between the observed distribution of each variable and its expected distribution. The \( p \)-values for the tests were computed using the simulated distribution of the K-S statistic. This ensured that the \( p \)-values, as required, were distributed uniformly under the hypothesis that the expected and observed distributions agree. Figure 1 shows examples of the distributions of three variables: the first passed the criterion for inclusion in the analysis, the second failed, and the third was at the threshold for rejection, namely a K-S \( p \)-value of 0.1. Having thus arrived at 12 sets of well-measured variables, we ordered the variables of each set according to their discrimination importance (on a scale of 100) using an algorithm called RuleFit [4]. We kept variables with discrimination importance greater than 10, which resulted in the selection of between 18 to 25 variables depending on the analysis channel.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example_plot.png}
\caption{The plots show comparisons of the sum of expected background (pale shaded histogram) and expected signal (dark shaded histogram) to data (points with error bars). \textbf{First plot}: distributions and Kolmogorov-Smirnov (K-S) \( p \)-value for the best modeled variable, \( p_T^{(\text{jet3})} \), in the (electron, 1-tag, 4-jet) channel. \textbf{Second plot}: distributions and K-S \( p \)-value for the worst modeled variable, \( \cos^{(\text{jet1, alljets})}^{(\text{alljets})} \). \textbf{Third plot}: distributions and K-S \( p \)-value for the variable Missing \( E_T \), which has a K-S \( p \)-value near the rejection threshold.}
\end{figure}
We constructed 12 BNN functions, $n(x)$—one per analysis channel, each defined by the average

$$n(x) = \int n(x, w) p(w|T) dw,$$

$$\approx \frac{1}{K} \sum_{i=1}^{K} n(x, w_k), \quad (2)$$

over neural networks, $n(x, w)$, where $w$ denote the neural network parameters. Each network contained a single hidden layer with 20 nodes. As indicated in Eq. (2), the integrals (which have a dimensionality of the order of 400) were approximated by randomly sampling $K = 100$ points $w_k$ from the posterior probability density function $p(w|T) \sim p(T|w) \pi(w)$ defined over the space of network parameters. The function $p(T|w)$ is the likelihood of the training data, $T$, comprising 10,000 signal plus 10,000 background events, and $\pi(w)$ is the prior density, taken to be a product of zero-mean Gaussians, one for each network parameter.

In the previous BNN analysis [1], we allowed the width of each Gaussian in the prior to adapt to the level of noise in the training sample. However, we found that the network parameter values became too large, which yielded BNN functions that were not smooth. The background MC events we used are especially noisy because of the large dynamic range of their event weights. An obvious way to reduce the effect of events with abnormally large weights is simply to eliminate them from the training sample. However, we chose to keep all events, but to reduce the influence of those with very large weights by limiting the growth of the network parameter values. This was done by fixing the width of each Gaussian in the prior to one of a small set of values determined from a study of the distribution of parameter values in neural networks trained using back-propagation. The use of this non-adaptive, and therefore noise-insensitive, prior led to a dramatic improvement with respect to the previous analysis [1]. The current performance of the BNNs is illustrated in Figure 2 for one of the 12 channels.

![Figure 2: BNN outputs normalized to unity (green: background, blue: signal), S/\sqrt{B} and S/\sqrt{S+B}, and the signal efficiency (\epsilon_S) versus background efficiency (\epsilon_B) curves for the electron, 1-tag, 3-jet analysis channel.](image)

All statistical aspects of the analysis were validated in detail using ensembles of pseudo-datasets.

### A. BNN Output Distributions

For each analysis channel, we applied the associated BNN function to the observed events and to samples of signal and background MC and QCD events that were different from those used in the construction of the BNNs. The distributions of BNN outputs are shown in Figs. 3-4 for the 12 analysis channels. The model distributions have been normalized to the observed integrated luminosity of 0.9 fb$^{-1}$, with the single top cross section set to the measured value of 4.4 pb in these and subsequent plots.

In order to check our modeling of background, we applied the BNNs to two background-dominated samples defined by the following criteria: 2 jets, 1 b-tag and $H_T < 175$ GeV for the “$W$+jets” sample and 4 jets, 1 b-tag, $H_T > 300$ GeV for the “$t\bar{t}$” sample. The first sample is mostly $W$+jets and almost no $t\bar{t}$, while the second is mostly $t\bar{t}$ and almost no $W$+jets. Figures 5 and 6 show the Bayesian neural network output distributions for these cross-check samples. We see good agreement between the predicted background and the observed data in both the samples for each of the electron and muon channels.
B. Bias Study

We studied the bias in the analysis method by applying the entire analysis chain to different ensembles of pseudo-datasets, each with a different value for the single top production cross section. A straight-line fit of the ensemble average of the estimated cross sections versus the input cross section yielded a slope and intercept of $0.98 \pm 0.02$ and $0.13 \pm 0.14$, respectively. We therefore concluded that the bias is negligible.
FIG. 5: BNN outputs from $W$+jets cross-check samples for the electron (left) and muon (right) channels.

FIG. 6: BNN outputs from $t\bar{t}$ cross-check samples for the electron (left) and muon (right) channels.

III. RESULTS

Each distribution in Figs. 3-4 was used to compute a posterior density $p(\sigma|\text{Data})$ for the single top cross section [1] as well as the posterior density when the 12 channels were combined. In addition, we computed the expected cross section for each of the 12 channels. An expected result is one obtained by setting the “observed” distribution of counts equal to the expected distribution. The expected results are shown in Table I and Fig. 7 and the observed results in Table II and Fig. 8.

<table>
<thead>
<tr>
<th>All tag and jet multiplicities combined</th>
<th>Lepton channel</th>
<th>$\sigma$ $\pm \Delta\sigma$</th>
<th>$\sigma/\Delta\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electron ($e$)</td>
<td>2.7 $\pm$ 1.8</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Muon ($\mu$)</td>
<td>2.7 $\pm$ 2.2</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>$e+\mu$</td>
<td>2.7 $\pm$ 1.5</td>
<td>1.8</td>
<td></td>
</tr>
</tbody>
</table>

TABLE I: Expected results for the electron, muon and combined channels.

<table>
<thead>
<tr>
<th>All tag and jet multiplicities combined</th>
<th>Lepton channel</th>
<th>$\sigma$ $\pm \Delta\sigma$</th>
<th>$\sigma/\Delta\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electron ($e$)</td>
<td>4.6 $\pm$ 2.0</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Muon ($\mu$)</td>
<td>4.5 $\pm$ 2.3</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>$e+\mu$</td>
<td>4.4 $\pm$ 1.5</td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

TABLE II: Observed results for the electron, muon and combined channels.
A. Significance of Excess

The posterior densities indicate that we have an excess of events over background, the significance of which can be quantified with a \( p \)-value defined as the probability to measure a cross section equal to or higher than some reference value. Figure 9 shows the distribution of cross sections from a background-only ensemble in which all systematic effects have been included. Using the SM signal cross section of 2.9 pb, we obtain a \( p \)-value of 1.6\%, which corresponds to a 2.2 standard deviation expected significance. Using the measured value of 4.4 pb, we obtain a \( p \)-value of 0.08\%, which corresponds to a significance of 3.1 standard deviations.

It is instructive to compute the power curve of the BNN method using the two hypotheses \( H_1 = \text{signal+background} \) and \( H_0 = \text{background-only} \). A power curve is a plot of the \( p \)-value, \( 1 - \beta \), for the alternative hypothesis \( H_1 \) versus the \( p \)-value, \( \alpha \), for the null hypothesis \( H_0 \); that is, it is a plot of the probability to accept the alternative hypothesis, if true, versus the significance level \( \alpha \). The power curves in Fig. 10 show that the BNN method, as anticipated, achieves a performance comparable to those of the decision tree (DT) and matrix element methods [1].
IV. SUMMARY AND CONCLUSIONS

We used Bayesian neural networks to separate single top quark signals from background in a sample of lepton+jets events selected from nearly 1 fb⁻¹ of Run II data. The BNN output distributions across twelve independent channels were combined using a binned likelihood and the single top \((tb + tqb)\) cross section was measured to be

\[
\sigma (p\bar{p} \rightarrow tb + tqb + X) = 4.4^{+1.6}_{-1.4} \text{ pb},
\]

using a Bayesian method [1]. This result is associated with a \(p\)-value of 0.08\%, that is, a 3.1 standard deviation significance. The \(p\)-value using the SM value of 2.9 pb for the signal cross section is 1.6\%, corresponding to an expected significance of 2.2 standard deviations. Figure 11 shows the observed BNN distribution summed over all 12 channels, superimposed on the summed signal + background distributions. We note the excellent agreement between
FIG. 11: The observed BNN output distribution summed over all 12 channels superimposed on the summed signal + background model. The plot on the right is a zoom of the region near a BNN output of 1, and the y-axis shown on a log scale.

the Standard Model prediction, including the $s + t$ single top signal, and the data.