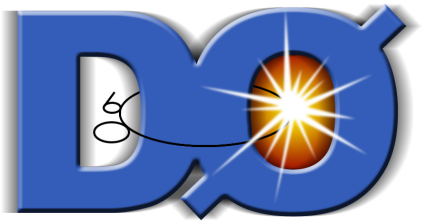
Optimization of the Sensitivity of the Standard Model Higgs Boson Analysis in the $WH \to l\nu bb$ Channel

Stephanie Hamilton
Michigan State University
SIST 2013, Fermi National Accelerator Laboratory
Supervisors: Dr. Michael Cooke and Dr. Ryuji Yamada

August 6, 2013



Abstract

During the summer of 2013, work continued on the analysis of the Standard Model Higgs boson in the $WH \to l\nu bb$ channel at the DØ detector at Fermilab. The focus of the summer effort was to achieve improved sensitivity by developing and utilizing new optimization tools that were aimed at improving multivariate analysis techniques. These tools, in combination with other strategies, resulted in an 9.05% improvement in the expected sensitivity to a Higgs boson with a mass of 125 GeV.

Contents

1 Introduction

The Higgs boson has been the subject of intensive searching by the high energy physics community at both Fermi National Accelerator Laboratory (Fermilab) and the European Organization for Nuclear Research (CERN) for nearly 50 years. It was finally discovered in 2012 by the ATLAS and CMS Collaborations at the Large Hadron Collider (LHC) at CERN, and with its discovery, all fundamental particles predicted by the Standard Model of Particle Physics have been observed. [1][2] The Standard Model describes all fundamental particles and their interactions with each other through the strong nuclear force, weak nuclear force, and electromagnetic force. It also describes the mediation of the aforementioned forces using gauge bosons. It is the most complete and accurate theory describing the fundamental particles in the universe that we have today.

The Higgs boson is not only predicted by the Standard Model, it is required. In a Higgs-less Standard Model, all fundamental particles are massless. Massless particles would not describe the world as we know it. The role of the Higgs boson within the Standard Model is to explain the electroweak symmetry breaking we observe happening early in the life of the universe and to provide an explanation for how massive particles acquire their mass [1].

The Fermilab Tevatron Collider accelerated protons and antiprotons to a center of mass collision energy of 1.96 TeV. Most collisions that occurred in the Tevatron were "unexciting" collisions in which one proton simply scattered off of an antiproton without creating a new particle. Occasionally, more direct collisions would have enough energy to produce a Higgs boson at a mass of 125 GeV.[2] These more exciting events were selected for permanent storage using a complex trigger system that was able to reduce the number of collisions recorded per second from 1.7 million per second to about one hundred.

[1]There has not been an observed occurrence of a right-handed neutrino. Since the gauge bosons of the weak force, the two W's and the Z boson, are left-handed, only left-handed neutrinos will be produced when they decay. In order for a Higgs boson to be the source of the mass of the neutrino, there must also be a right-handed neutrino. There are experiments that are actively looking for this phenomenon.

[2]While the units of momentum and mass are technically GeV/c and GeV/c^2 respectively, with c being the speed of light, it is convention in high energy physics to work in a unit system where c=1.

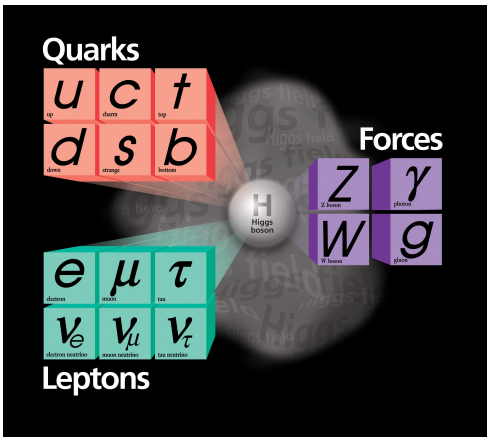


Figure 1.1: The Standard Model of Particle Physics

Higgs boson production at the Tevatron is dominated by two major production mechanisms. The first is gluon-gluon fusion, in which two gluons collide with enough energy to form a Higgs boson. The second and less likely production method is associated production, in which a Higgs boson is radiated by either a W or Z boson (See Figure 1.2).

The Higgs boson can then decay by several different methods, with the likelihood of each determined by the mass of the Higgs boson. While the Standard Model requires a Higgs boson, it does not specify the mass of the particle. Once the mass is determined, however, the Standard Model completely determines the particle's properties. Since the mass was not predicted, physicists were forced to split the search into different mass regions. A low mass Higgs boson (115 GeV to 135 GeV) would prefer to decay into a bottom-antibottom quark pair, $b\bar{b}$, while a high mass Higgs boson (135 GeV to 200 GeV) would prefer to decay into a pair of W bosons. The likelihood of each decay process based on the mass of the Higgs boson is called the branching ratio (See Figure 1.3). As reported by both ATLAS and CMS, the two general-purpose experiments at CERN, the excess of events has been seen to occur at a mass of approximately 125 GeV, placing the observed particle in the low mass regime. [1][2] The $WH \rightarrow l\nu bb$ channel remains the favored channel for studying a low mass Higgs boson at the Tevatron because requiring the W boson to decay into a lepton, specifically an electron or a muon, and a neutrino greatly reduces $b\bar{b}$ background. See Figure 1.4 for

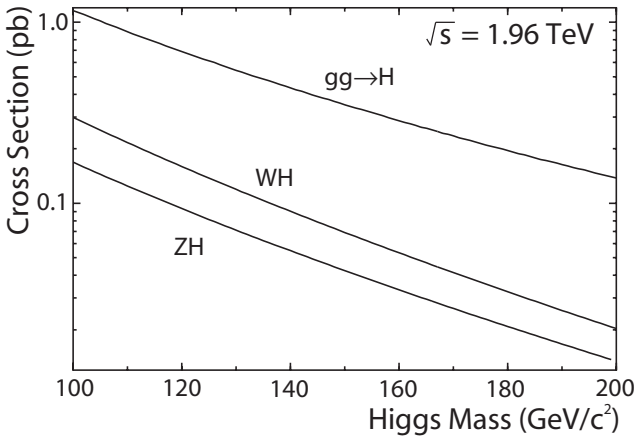


Figure 1.2: Production Cross Sections of the SM Higgs Boson [3]

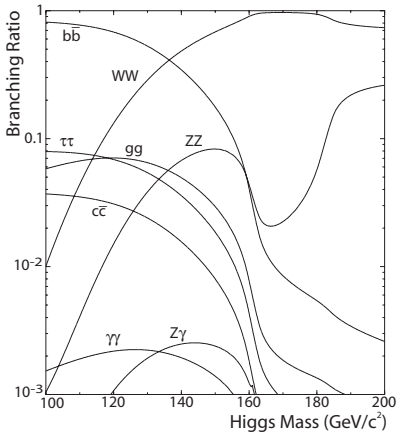


Figure 1.3: Decay Branching Ratios of the SM Higgs Boson [3]

a Feynman diagram describing the decay process of this channel. This is one of the most sensitive channels of the Tevatron.

The WH channel has a very low signal-to-background ratio and thus high integrated luminosities and sophisticated analysis techniques (detailed later in this paper) are required in order to separate signal events from background events in the most effective way possible. This paper details ongoing efforts to optimize and refine the analysis of the $WH \rightarrow l\nu bb$ decay channel.

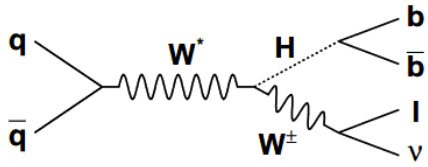


Figure 1.4: Feynman Diagram for $WH \rightarrow l\nu bb$ [3]

2 Materials and Methods of Research

2.1 The DØ Detector

The DØ detector consists of three major components specifically designed to accurately and precisely measure each aspect of a collsion inside the detector. The innermost component is the tracking system, comprised of the Silicon Microstrip Tracker (SMT) and the Central Fiber Tracker (CFT). The SMT consists of barrels and wedges of silicon sensors placed in both the central and forward regions of the detector in order to cover low and high pseudorapidity η[3]. The CFT is located outside the SMT and is composed of scintillating fibers that produce light as a charged particle passes through them and simultaneously shifts the light into the green spectrum, the range optimal for electronic readout. The next layer is the calorimeter, responsible for stopping everything except neutrinos and muons. The calorimeter consists of layers of uranium and liquid argon. The uranium stops particles by causing them to turn into a spray of lower-energy particles, while the liquid argon is ionized by charged particles, which allows us to detect the showers of particles and measure their energy. The final layer of the detector is the muon system, which is designed to tag muons that leave a track in the central tracker but leave little energy deposited in the calorimeter as they pass through it.

The electronic information encoded in the signals sent by each part of the detector is then utilized by systems of computers to reconstruct events. Particles can be identified according to the tracks and energy deposits they leave in the detector. An electron passing

[3]"Pseudorapdity" is defined as $\eta = -\ln\left[\tan\frac{\theta}{2}\right]$, where θ is the polar angle measured from the proton beam axis

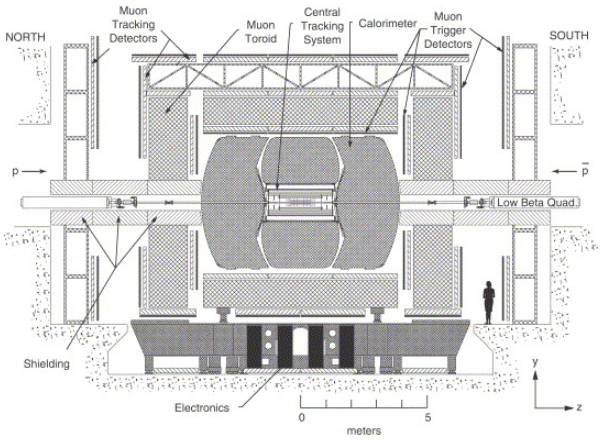


Figure 2.1: The DØ Detector [4]

through the detector will leave a track in the tracker as well as a concentrated energy deposit in the calorimeter. A photon looks very similar to an electron in the calorimeter, but it does not have a charge and so will not leave a track in the tracker. Quarks and gluons undergo a process known as hadronization almost immediately. The nature of the strong force is to become stronger as distance increases, which is opposite of what happens with the gravitational or electromagnetic forces. Therefore, when attempting to separate a pair of quarks, there is a point at which it takes less energy to simply create new particles than to continue to increase the distance between the quarks. This spray of new particles can contain charged particles, which will leave tracks in the tracker. The particles then go on to leave a wide energy distribution in the calorimeter. This type of signature is commonly referred to as a jet. Muons are not highly ionizing particles and tend to travel straight through the detector without depositing much energy anywhere and leave a long track through each part of the detector. Neutrinos are very weakly interacting particles and almost always travel through the detector without interacting with anything. Therefore, they are measured as missing energy in the event.

2.2 ROOT

ROOT is an object-oriented library and program based off of the programming language C++ that is maintained by CERN. This program is the primary way in which high energy physicists interact with and analyze data events and Monte Carlo (MC) events. Monte Carlo is the name given to the theoretical simulations of data events. ROOT is used to make and analyze histograms of data, along with making four-vector computations and utilizing statistical tools to analyze the data. ROOT uses the CINT interpreter, which allows the user to interactively input C++ commands. ROOT's main data container is called TTree, with subcontainers TBranch and TLeaf, and these containers act as a way for the user to see the raw data that are stored in files with a .root extension. These files contain information about the kinematic properties of all particles relevant to the event, missing transverse energy, b-tagging, and detailed particle or jet identification information.

2.3 b-tagging

b-tagging is the process by which jets originating from bottom quarks are identified. The signature of a jet of this type is a secondary vertex that is offset from the primary vertex of the event by a few millimeters, indicating a particle with a lifetime on the order of a couple picoseconds. This secondary vertex must also have an impact parameter that is slightly displaced from the primary vertex. The b-tagging efficiency is defined as the amount of tagged b jets over the amount of true b jets. This number is determined by the data and used for Monte Carlo simulations. The "fake rate" is defined as how often an up, down, charm, or strange quark is tagged as a bottom quark. Figure 2.2 shows the plot of b-jet Efficiency versus Fake Rate for the DØ detector.

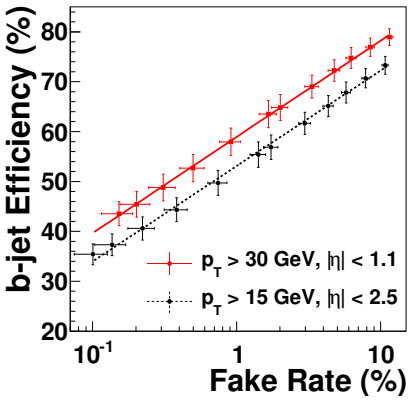


Figure 2.2: Efficiency of b-tagging vs. Fake Rate [5]

2.4 TMVA and Multivariate Techniques

Multivariate analysis (MVA) techniques are utilized at DØ to combine several moderately discriminating variables into one strongly discriminating variable. ROOT provides a library called TMVA that contains all of the tools needed for the application of MVA methods. TMVA provides a variety of classifiers, such as Neural Networks, Support Vector Machines, or Decision Trees that one can use.

There are several ways in which this type of analysis can be done, but the favored method at DØ is to use Boosted Decision Trees (BDTs). BDTs stem from the idea of Decision Trees (DTs), in which a cut is made on an input variable and the full sample is divided into subsamples that either pass or fail this cut. Subsequent cuts are made on different variables until the DT reaches a stop criterion. The final "leaves" of this tree then have specific signal-to-background ratios and are classified as either “signal” or “background” depending on the majority of events that end up in the leaf. In a BDT, this concept is extended from one to several trees that then form a forest. The trees use the specific signal-to-background ratios from previously trained trees as weights in order to correct for misclassified events. The trees are then finally combined into a single classifier that is the weighted average of the individual decision trees. [6]

TMVA also provides a number of different method options that can be varied and tuned in order to produce better classifiers. The tuning of these options is extremely important because only finite statistics are available. Varying the method options allows us to find the ideal combination that extracts the most information possible out of the limited statistics. The statistics in Higgs analyses are limited and the ultimate goal of any Higgs analysis is to optimize the multivariate analysis as best as possible with the statistics that are available. It is possible to vary the number of trees contained in a Random Forest, specify the number of nodes any tree may have, define how the trees will be boosted or pruned of insignificant nodes, or specify the maximum number of levels a tree can have. These are but a few of the numerous method options TMVA provides.

2.5 COLLIE

Part of our analysis process is using the output from TMVA as an input to a program called COLLIE, which, among performing other tasks, returns a upper exclusion limit on the cross section of Higgs production. Specifically, COLLIE produces the ratio of the maximum cross section consistent with the data to a 95% Confidence Level (C.L.) to the cross section that is predicted by the Standard Model. This limit depends on the mass that is being considered. Prior to 2012, when the mass of the Standard Model Higgs boson was unknown, a mass range could be considered excluded if the 95% C.L. limit for that range was below one times the Standard Model cross section. Since the Higgs boson observed in 2012 has a mass near 125 GeV, that is the mass point that was used throughout the course of our analysis.

3 Summer Work

3.1 My Work

My task for the first portion of the summer was to develop a tool that could be used to optimize an MVA training over several different values for various training method options.

I spent a significant amount of time working to find the method options that would be most useful to optimize over. In the end, there were four method options I decided to vary the values for.

1. `NTrees` - This method option varies the number of trees that TMVA creates to put in the forest. With a low number of events in a training sample, a lower number of trees can be helpful.
2. `Shrinkage` - This method option defines the learning rate of the boosting algorithm of the trees and ranges from 0.0-1.0. A small shrinkage would require more trees to be grown but can significantly improve the prediction of the training.
3. `NNodesMax` - This method option specifies the maximum number of nodes that any tree can have. Less nodes can be useful when there are a low number of events in the training sample.
4. `GradBaggingFraction` - This method option defines the fraction of events that will be used in each iteration of growing a tree, when one is using random subsamples of all events. [6]

When an MVA is trained, a ROOT file is produced that contains a multitude of information about the results of the training. Some of this information was used to gauge how effective an MVA might be as a discriminator. The first useful piece of information is the Signal Acceptance vs. Background Rejection curve that TMVA produces. This curve is also called the Receiver Operating Characteristic (ROC) curve. The ROC curve states how many of the background events would be rejected if a certain percentage of signal events was accepted. Essentially, a higher area under this curve indicates a higher amount of signal discrimination. The goal with optimization was to get the highest integral of this curve that was possible. See Figure 3.1 for an example curve.

The second useful piece of information is a signal sensitivity metric. In particular, we were interested in the ratio of the number of signal events to the square root of signal plus background events, or $S/\sqrt{S+B}$. This metric is indicative of the significance of a cross-section measurement of the signal for a given cut on the discriminant. In order to obtain this value, the signal and background efficiency curves produced by TMVA were used (See

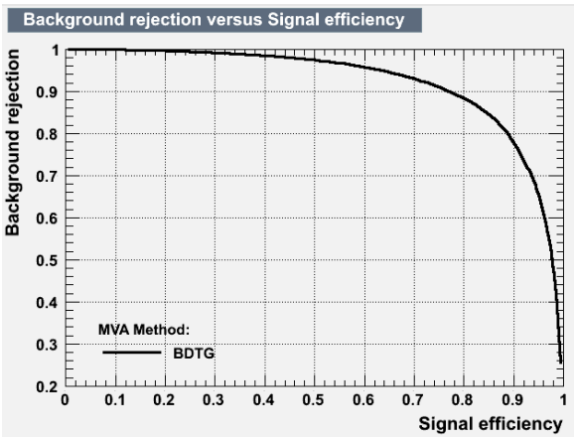
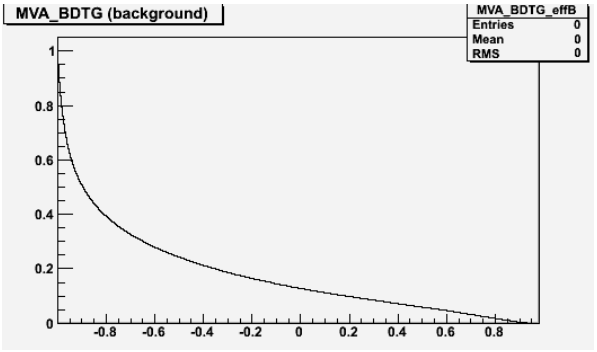
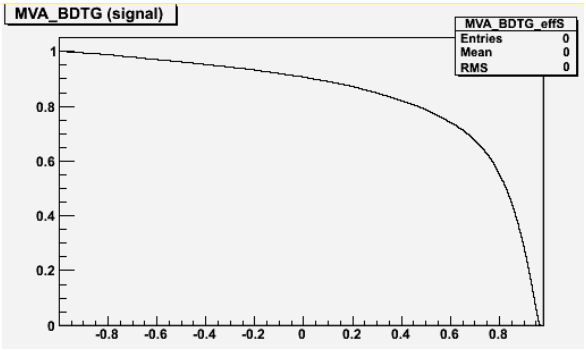


Figure 3.1: Example Signal Acceptance vs. Background Rejection Curve



(a) Example Background Efficiency Curve

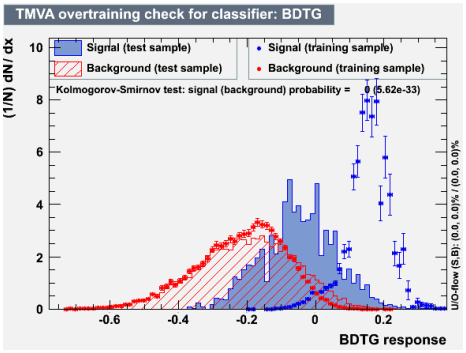


(b) Example Signal Efficiency Curve

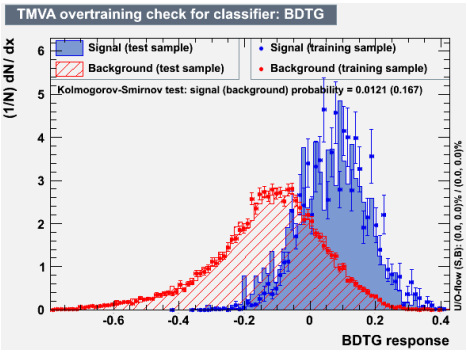
Figure 3.2: Background and Signal Efficiency Curves

Figure 3.2 for examples). These curves, together with appropriate normalization factors, can be used to determine the total number of signal or background events that would be kept after a cut on a final discriminant. This information can then, in turn, be used to determine where the best value of $S/\sqrt{S+B}$ occurred for a given cut.

The third useful piece of information produced by TMVA is the overtraining plot. When an MVA is trained, TMVA splits the sample into three statistically independent subsamples: one for parameter optimization, one for overtraining detection, and one for performance validation. The first subsample becomes the train sample, while the last two are merged to form the test sample. [6] Overtraining occurs when TMVA begins to make decisions based



(a) Severely Overtrained TMVA output



(b) Reasonable TMVA Output

Figure 3.3: Examples of Overtrained vs. Not Overtrained TMVA Outputs

on statistical fluctuations rather than on the physics properties of the training samples. A reasonable measure of the amount of overtraining is the result of a Kolmogorov-Smirnov (KS) test calculated for the signal and background distributions of the training and testing samples. The KS test essentially gives the probability of whether two histograms look like they originate from the same distribution or not. TMVA trains an MVA on the training sample and applies the MVA to the independent test sample, then compares the resulting distributions by way of a KS test. An MVA was considered overtrained if the KS test for either the signal or the background distributions was below 1%. (See Figure 3.3 for example plots)

When an MVA is trained using the optimization process that was developed, a ROOT file is again produced, but it now contains subdirectories labeled according to the specific combinations of the four options that were used for that round of training. Each of these subdirectories then contains the information stated above. Depending on the number of different values for the options one desired to optimize over, there could easily be on the order of two hundred different combinations, each with its own subdirectory in the ROOT file.

For this reason, I developed a macro that would use this ROOT file as input and print out all of the information stated above in an organized fashion. In addition to this, it also

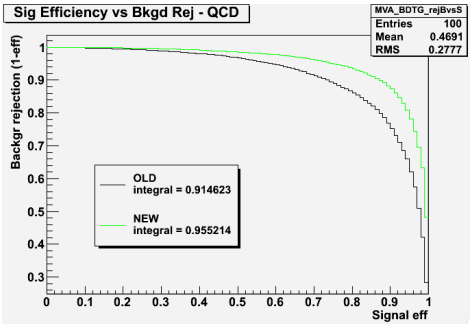
stated which combination of options yielded the best result for each piece of information, with the precondition that that particular combination had already passed the KS test for both signal and background. In order to pass the KS test, both the signal and background distributions had to have results above 1%. Using this macro, it became much easier to sort through the information contained in the ROOT file and thus determine the optimal settings to use to train the MVA.

Once the optimization process could easily be run and my macro was working, I began to utilize the tools to optimize important secondary MVAs that were aimed at separating the Higgs boson signal from a specific type of background. There were four secondary MVAs for each of the big backgrounds of the $WH \to l\nu bb$ channel: $t\bar{t}$ (a top-antitop quark pair), V+jets (a W or a Z accompanied by jets), VV (any combination of W's and Z's, also called diboson), and QCD (a jet that is misidentified as an electron, also called multijet). Each of these processes can have a signature that looks exactly like our signal, and therefore it was very important to develop strong discriminators that were able to separate these background processes from our signal. The optimization process allowed us to develop the strongest discriminators possible, maximizing the amount of signal we were able to keep for a certain amount of background that was rejected.

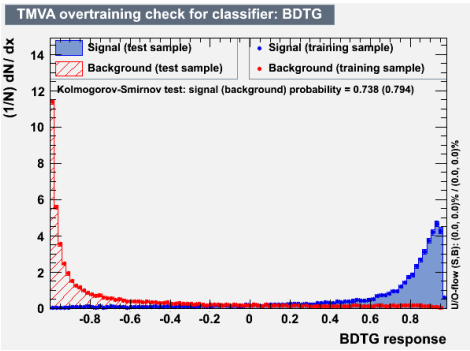
The combination of new variable selection techniques (detailed later) with the new MVA optimization process resulted in significant improvements in the discriminating power of each secondary MVA. These improvements can be seen in Figure 3.4.

3.2 Challenges

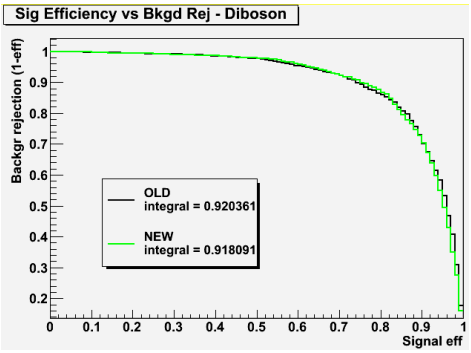
There were a few challenges, apart from the usual debugging of code, that I encountered during the summer. The first was expanding my knowledge of ROOT in order to develop the macro that was used to sort through the information that was returned during the optimization process. I needed to familiarize myself with the structure of a ROOT file, as well as learn which functions were allowed to be called with which types of ROOT objects.



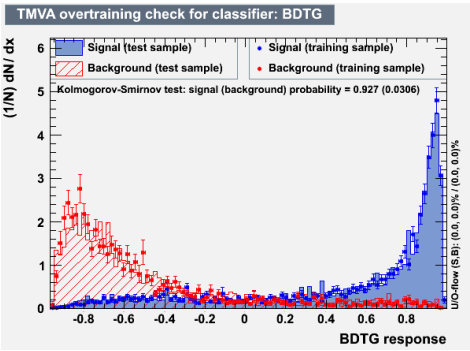
(a) New vs. Old Multijet MVA ROC Curves



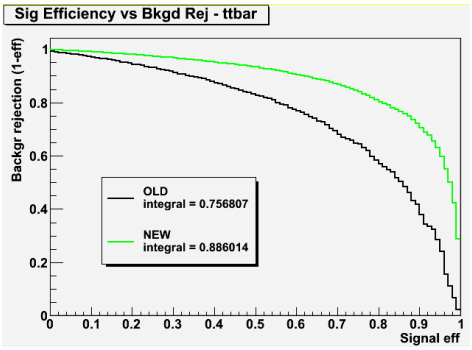
(b) Overtraining Plot, Multijet MVA



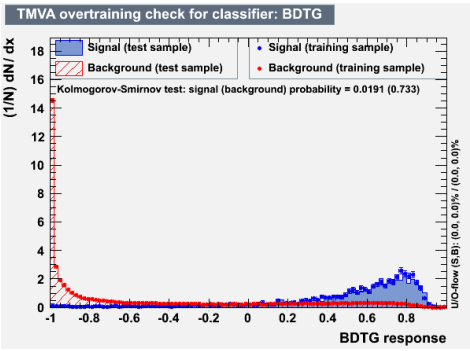
(c) New vs. Old Diboson MVA ROC Curves



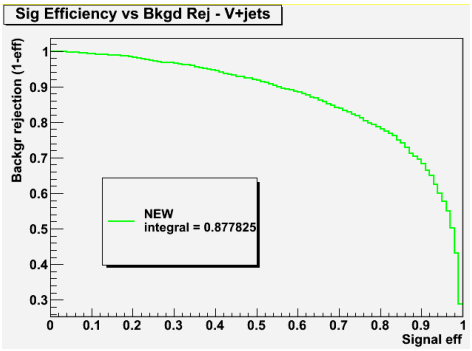
(d) Overtraining Plot, Diboson MVA



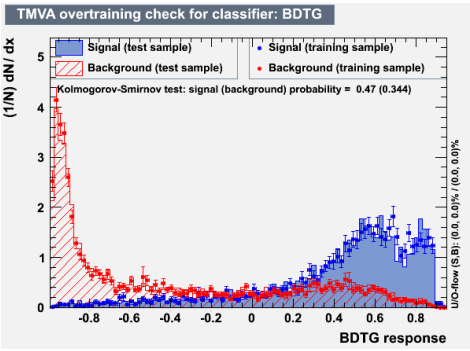
(e) New vs. Old $t\bar{t}$ MVA ROC Curves



(f) Overtraining Plot, $t\bar{t}$ MVA



(g) New V+jets MVA ROC Curve (Note: There was no old V+jets ROC curve to compare to)



(h) Overtraining Plot, V+jets MVA

Figure 3.4: Improvements in Secondary MVA Training

This took time, but I feel I am more well-versed with ROOT simply from developing this macro.

Another challenge I encountered was becoming familiar with each step of our analysis process. The framework that is used for the $WH \to l\nu bb$ analysis is a huge collection of files written in C++. In order to complete my project, I needed to become familiar with numerous files, as well as learn what each step of the process accomplished.

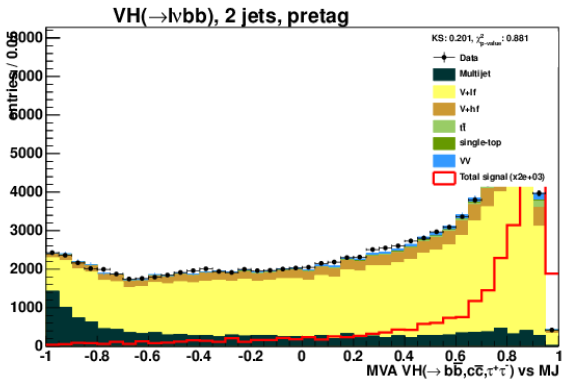
3.3 Other Work

A set of tools that was particularly useful in the optimization process was a class of our analysis framework developed by my colleagues, Will Johnson and Ben Rabe, that would ultimately provide a list of the best variables to use as inputs for a particular MVA. When run, their tools would calculate the values of a χ^2 and a KS test in order to determine whether the variables were well-modeled by the Monte Carlo. Using the results of these tests, the variables were separated into "GOOD" and "BAD" categories for each type of background, as well as for all backgrounds together. An integral test could then be used on the distribution of the variable for background versus its distribution for the signal. The integral test determined how much of these two histograms overlapped; therefore, a smaller integral test meant the variable had a higher separation power between background and signal and would be a potentially powerful discriminator in an MVA.

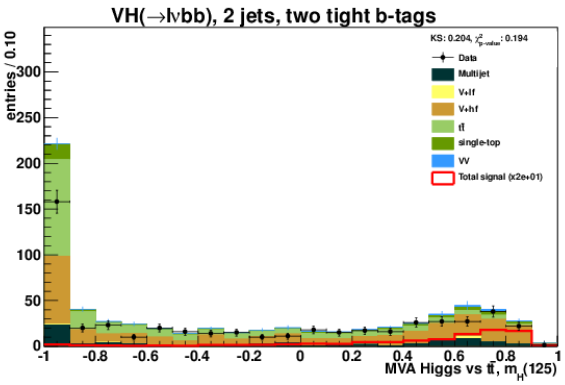
4 Results

All efforts during the course of the summer culminated in a final result that gave a measure of our sensitivity to a Higgs boson. The improvements that were made in the multivariate discriminators translated to increased discriminating power in the secondary MVAs as well as in the final MVA. Figure 4.1 contains plots of the improved MVAs.

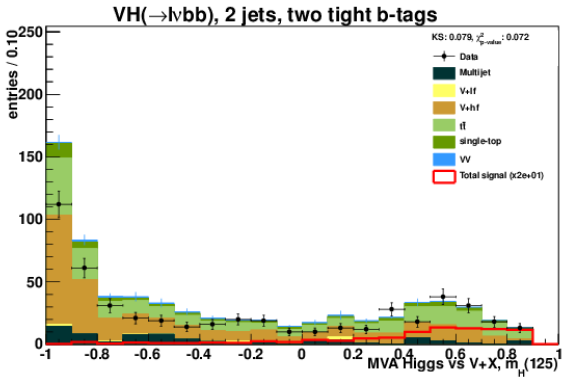
In addition to increased discriminating power, the improvements in the discriminators



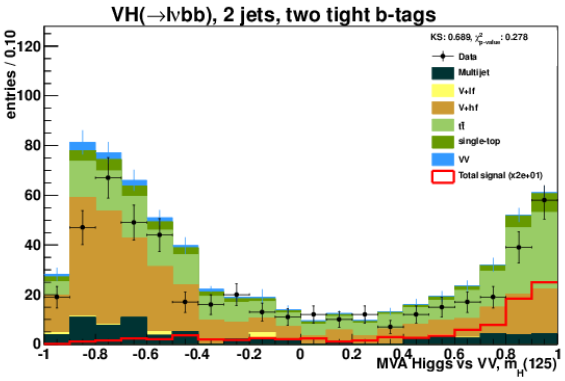
(a) Multijet MVA



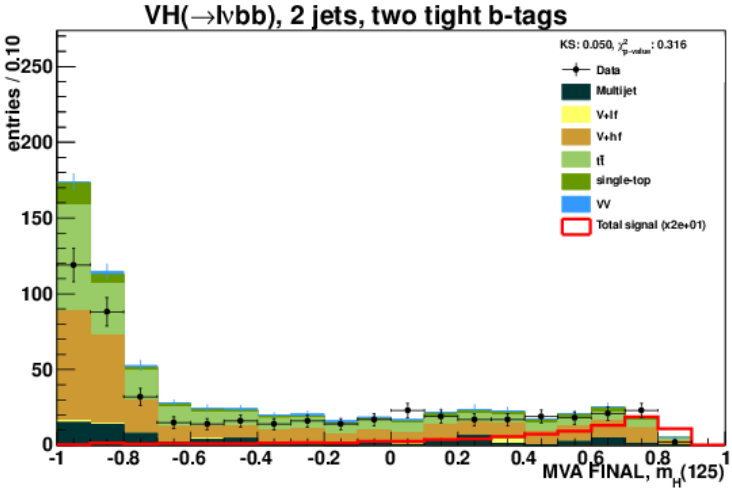
(b) $t\bar{t}$ MVA



(c) V+jets MVA



(d) Diboson MVA



(e) Final MVA

Figure 4.1: Output of the multivariate discriminators for the four major backgrounds after the optimization work of Summer 2013 (All plots are a work in progress)

Table 4.1: 95% C.L. limits on the SM Higgs boson production cross-section. The numbers in the second (old MVAs) and third (newly optimized MVAs) columns represent the ratio of the expected cross-section to the Standard Model cross-section. The fourth column states the percent difference between the second and third columns.

	Before Summer 2013	After Summer 2013	Percent Increase
MVA el	6.28	5.7	9.24%
MVA mu	6.52	5.88	9.51%
MVA el+mu	4.42	4.02	9.05%

also contributed to an increase in the expected sensitivity to the cross-section of a Higgs boson with a mass of 125 GeV in the $WH \rightarrow l\nu bb$ channel. The increase in expected sensitivity can be seen in Table 4.1.

5 Acknowledgements

I would like to thank my supervisors Dr. Mike Cooke and Dr. Ryuji Yamada for their teaching and guidance. I would also like to acknowledge my fellow summer students and coworkers, Alex Abbinante, Tony Podkowa, Ben Rabe, Will Johnson, and the rest of the WH Analyzers group for all of the hard work they put into this summer and into producing a meaningful result. I would like to thank Diane Engram and Linda Diepholz, as well as the entire SIST committee, for giving me this incredible opportunity. Finally, I would like to thank the DØ Collaboration and Fermi National Accelerator Laboratory for allowing me to come and work this summer.

6 Works Cited

[1] The ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC." [Online] Available: http://www.arxiv.org [Accessed July 24 2013].

[2] The CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC." [Online] Available: http://www.arxiv.org [Accessed July 24 2013].

[3] The CDF Collaboration, "Search for New Particles Decaying into $b\bar{b}$ and Produced in Association with W Boson." [Online] Available: http://www-cdf.fnal.gov [Accessed July 11 2013].

[4] The DØ Collaboration, "The Upgraded DØ Detector." [Online] Available: http://www.sciencedirect.com [Accessed July 17 2013].

[5] The DØ Collaboration, "b-jet Identification in the DØ Experiment." [Online] Available: http://www.arxiv.org [Accessed July 25 2013].

[6] TMVA Developers, "TMVA Users Guide," November 2009, [Online]. Available: http:tmva.sourceforge.net. pp. 27–28 and pp. 104–108. [Accessed July 17, 2013].

[7] The DØ and CDF Collaborations, "Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in Higgs boson searches at the Tevatron." [Online] Available: http://www.arxiv.org [Accessed July 16 2013].

[8] The DØ Collaboration, "The DØ Silicon Microstrip Tracker." [Online] Available: http://www.sciencedirect.com [Accessed July 17 2013].