Continuation of the Search for the Standard Model Higgs Boson in the $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ Channel

Anthony Podkowa

August 8, 2011

Abstract

During the summer of 2011, work began on developing intermediary stages of an analysis of the high-mass Standard Model Higgs boson in associated production with a W boson. This channel, $WH \rightarrow WWW \rightarrow l\nu.jj.jj$, is of particular interest since it has not yet been investigated by any other analysis group. Currently the wh_cafe analysis framework has been updated and utilized to produce plots comparing the data samples with Monte Carlo simulations of both the $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ signal and various background processes. Continued efforts are being made to modify the framework for this analysis.

Contents

1 Introduction

The Higgs boson is perhaps the most sought after particle in the field of High-Energy Physics (HEP) today. The object of Leon Lederman's book, *The God Particle*, the Higgs boson was postulated by Peter Higgs as the particle responsible for mass, and has been the object of many searches at Fermi National Accelerator Laboratory (Fermilab) and at the European Center for Nuclear Research (CERN). Its discovery is key to the verification of the Standard Model (SM), the overarching theory which describes all the known fundamental particles and their interactions. Though the Standard Model provides much information about the Higgs boson, it provides no information regarding its mass. The Higgs mass is significant since it effects the final decay products. The final decay products are what can actually be seen in particle detectors. Recent results from CMS and ATLAS[1] at the Large Hadron Collider in Geneva, Switzerland were able to exclude the Higgs mass between 155 to 206 GeV at a 95% confidence level[2][1].

Because of the way that the decay products can change based on the Higgs mass, Standard Model Higgs searches are divided into two main categories: low mass Higgs searches between 115 and 130 GeV, and high mass Higgs searches between 130 and 200 GeV. Depending on the mass of the Higgs, it can decay in a variety of ways. If the Higgs falls in the low mass range, it decays predominantly into bottom quark-antiquark pairs ($b\bar{b}$). Conversely, if the Higgs lies in the high mass range, the Higgs will decay primarily into a pair of W bosons. The likelihood of the each type of Higgs decay is called the branching ratio, and it is plotted in Figure 1.2.

The Standard Model Higgs can be produced via two methods: gluon fusion, where two gluons collide and produce a Higgs, or through associated production. In associated production, the Higgs is radiated by either a W or a Z boson. Our channel is concerned with the

[1]Results based on exclusions from the individual experiments, and have not officially combined as the Tevatron's Higgs exclusions.

[2]As per HEP convention, natural units are utilized throughout this paper. Though mass and momentum units are actually in GeV/c^2 and GeV/c respectively, the common way of representing these units is to set c=1 to simplify things, leaving the units in GeV.

Figure 1.1: Higgs Cross Section vs Mass [3]

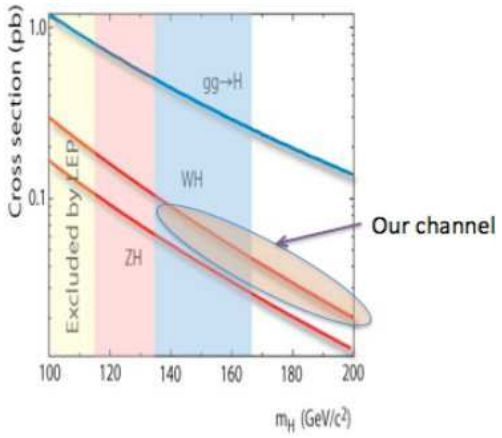
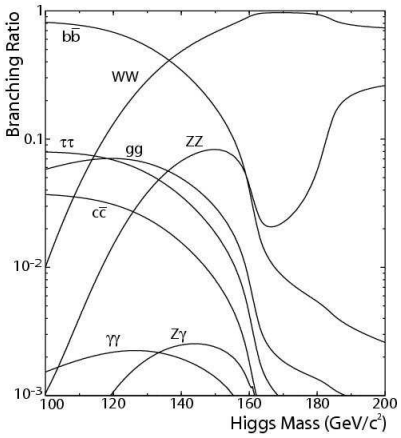


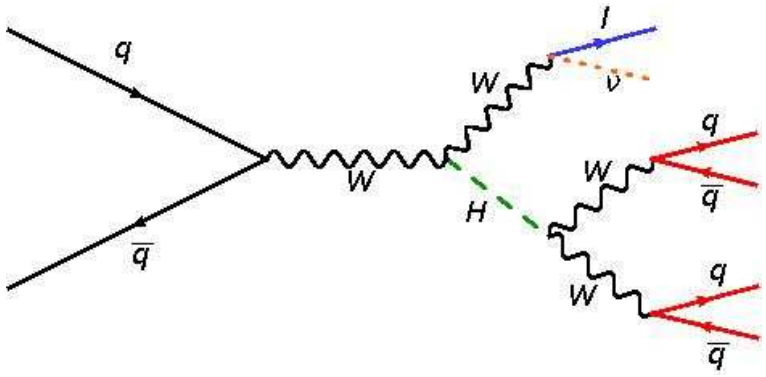
Figure 1.2: Higgs Decay Branching Ratio vs Mass[2]



production of the Higgs in association with a W boson, which has a cross-section around an order of magnitude lower than that of gluon fusion at the Tevatron (see Figure 1.1). This indicates that there should be ten times more Higgs events from channels involving gluon fusion than those involving associated production. However, gluon fusion channels also feature a higher cross-section for background processes as well, making each event harder to detect [3]. Thus searching for an associated production channel, such as ours, can still yield beneficial results. The Feynman diagram of the $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ channel is shown in Figure 1.3.

The purpose of this paper is to detail the continuing efforts at developing an analysis framework and processing the data and Monte Carlo simulations regarding the $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ channel. Efforts to update the framework to fit our analysis are discussed, as well as the development of software specific for this analysis.

Figure 1.3: Feynman Diagram for $WH \rightarrow WWW \rightarrow l\nu.jj.jj$



Generated using JaxoDraw (http://jaxodraw.sourceforge.net)

---

2 Materials and Methods

2.1 The DØ Detector:

The DØ detector is a highly sophisticated piece of machinery engineered for the purpose of identifying and tracking particles produced from the proton-antiproton collisions inside the Tevatron. It consists of a central tracking system, a calorimeter system and a muon detector system. The central tracking system is centralized within a 2 T solenoidal field, and consists of two separate tracking subsystems. Closest to the beam collisions is the Silicon Microstrip Tracker (SMT), which consists of four layered barrel silicon detectors and disks in the central region. In addition, there are large diameter disks positioned in the forward regions of the detector for tracking particles at high pseudorapidity (η)[3]. Outside of the SMT lies the Central Fiber Tracker (CFT), which is composed of scintillating fibers arranged within eight cylindrical supports wrapped concentrically around the beam pipe. Two doublets of scintillating fibers are overlaid on each cylinder, one parallel to the beam axis (z) and the other at an angle of $\pm 3^\circ$ with respect to z. The liquid argon and uranium calorimeter system consists of three regions: the Central Calorimeter (CC) and the two End Calorimeters (EC). The Central Calorimeter covers the region up to $|\eta| \approx 1$, while the End

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

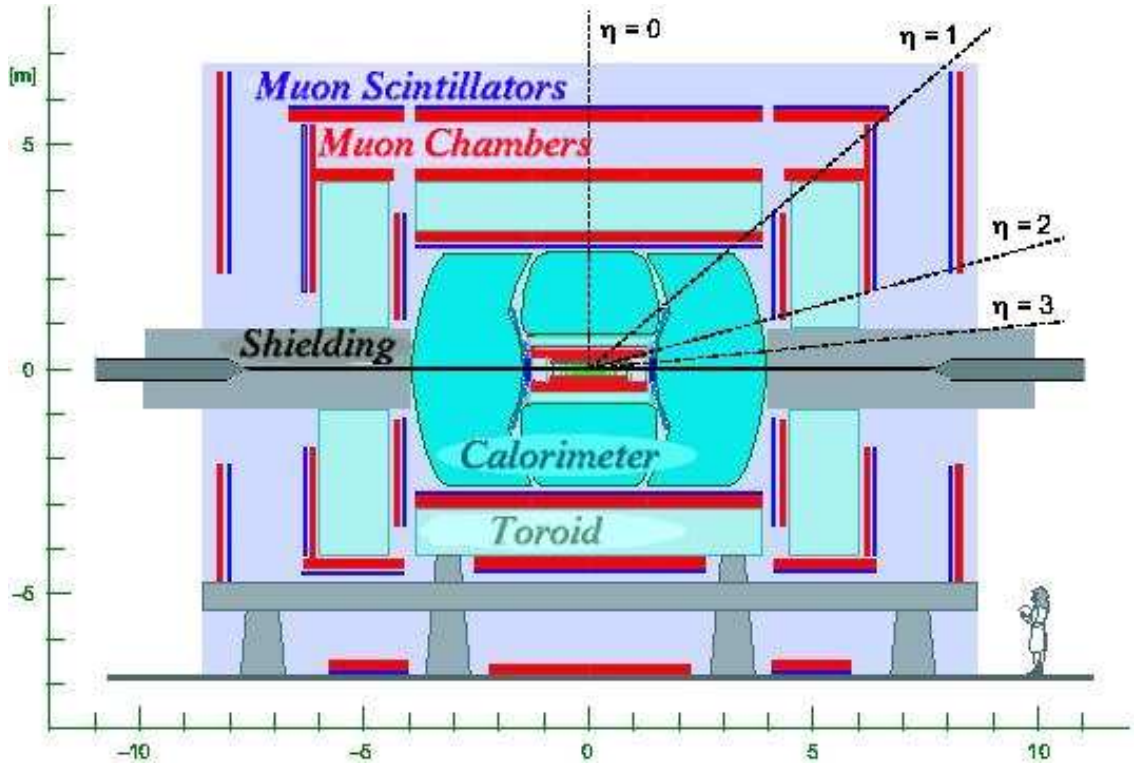


Figure 2.1: The DØ Detector

Calorimeters cover up to $|\eta| \approx 4$. The electromagnetic (EM) regions of the calorimeter are located closest to the collisions, followed by layers of fine and coarse hadronic layers outward.

2.2 C++

C++ is an object-oriented programming language derived from the C programming language. Originating in the late 1970s, the C++ language extended the functionality of C to include classes, which enabled a high level of abstraction whilst maintaining the high level of portability and low-level optimization. Many of these aspects continue to make C++ a popular language for software development. In the context of particle physics, C++ enables physicists to efficiently process data in the fractions of seconds after it is taken by the detector, while keeping the interface simple enough to write code to process the data.

2.3 ROOT

Developed and maintained by the European Center for Nuclear Research (CERN), ROOT is an object-oriented framework utilized by high-energy physicists worldwide. The core of the ROOT framework is based on the development of TTrees, classes that store histograms within binary ROOT files. TTrees and their DØ-specific derivatives, TMBTrees, provide the basic structure of our analysis output. Other secondary classes such as the TBranch, TClonesArray, and the TLeaf classes contribute to the organizational structure of ROOT files. TTrees for high-energy physicists usually contain information regarding the kinematic information of particles, such as the mass, missing transverse energy ($\not\!\!E_t$) and particle identification information.

2.4 MVA Processing

In order to effectively separate the signal MC from the various background processes it is often necessary to utilize Multivariate Analysis (MVA) techniques. Embedded in the ROOT framework is TMVA Library, which is designed for the efficient application of Multivariate methods in the context of HEP analyses. This library contains a multitude of MVA tools, which include a variety of classifiers, such as Artificial Neural Networks, Support Vector Machines, and Decision Trees. Within the context of the wh_cafe framework, $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ specific code was written to provide convenient access to the MVA methods.

2.4.1 Boosted Decision Trees

A Decision Tree (DT) is a binary tree structured classifier (see Fig 2.2). It operates by using a binary sort algorithm, making decisions on individual input variables one at a time until it reaches a stop criterion. This partitions the phase space into many different regions based off of the final leaves of the DT. Each partition of the phase space is then classified as either "signal" or "background" based on the majority of the training events that end up there. Boosting of DT's is an extension of this concept, which relies on training multiple DT's

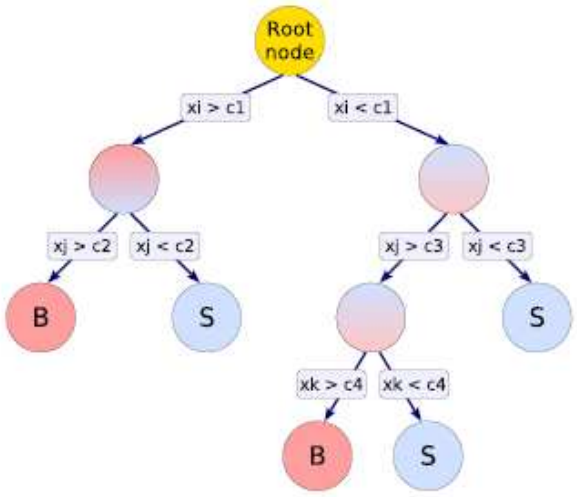


Figure 2.2: Example of a Decision Tree

---

which form a Random Forest (RF). Each tree in the RF reweights the misclassified events and retrains itself to obtain the optimal classification of the training sample. The resulting RF combines the DT's into a single classifier, which represents the weighted average of the individual DT's. Boosting is well known to stabilize the response of the DT's with respect to statistical fluctuations in the training sample and considerably enhance the performance vs a single tree.[5]

2.5 Shell Scripting

Remote utilization of the servers on the DØ Clued0 Linux cluster requires the usage of a shell, a simple interactive programming environment which processes text commands and returns human-readable output. The default shell for servers in the Clued0 Cluster is tcsh (TC SHell), a shell which attempts to mimic C style syntax. Though running the tcsh interactively can be extremely useful, it is often more productive to write scripts to automate repetitive tasks that require several shell commands. Shell scripts are often used in DØ to aid in the setup of of the various tools and run a variety of tasks. During my time here at Fermilab, I wrote several scripts in both tcsh and bash (Bourne Again SHell) to aid in

a variety of tasks, including a special run script for wh_cafe (See Section 2.6), variable retrieval from MVA weight files, and processing analysis output.

2.6 Analysis Framework

For our analysis, we chose to adapt an existing framework already in use at DØ: wh_cafe. The wh_cafe framework was developed by the WH Group for processing the $WH \to l\nu bb$ and $H \to WW \to l\nu jj$ channels. This presented itself with several advantages. First of all, we could utilize the WH Group's data and MC samples for our own purposes. Since the final states of both our channels are so similar ($l\nu jj$ vs. $l\nu jjjj$), we did not need to retrieve these samples ourselves. Last year, retrieving MC samples for our channel took a considerable amount of time and effort for just two samples. This year, we had all the samples provided for us, making it much easier for us to start working right away.

The second advantage of using the framework was that it was well understood. Since it had been in use for some time already by the WH Group, many of the errors and bugs that we ran into had already been encountered previously, and could be corrected rather quickly. Dr. Mike Cooke in particular was an invaluable resource for debugging many of the issues that we encountered along the way throughout the summer.

Finally, the framework already had code prepared for a complete analysis chain. Whenever we needed to write code that was specific to our own analysis, there was typically some example as to how we could approach the task already written for another channel. This particular aspect of the wh_cafe framework was extremely helpful for us while we were developing the code.

3 My Contribution

Unfortunately for our channel, for the majority of the summer the WH Group had a paper in review for publication. This meant that wh_cafe itself could not be directly altered by

our analysis group, since the framework itself had to be stable for last minute changes to the paper. This meant that in order to make persistent changes for our channel, we needed to maintain our own fork of the framework apart from official wh_cafe releases. For this reason, I volunteered to maintain and administer this fork of wh_cafe. My administrative responsibilities in this role required me to monitor changes that we added to the framework, ensure the quality of each release, and keep our fork in sync with the core wh_cafe release. In this role, I also aided by editing and integrating code developed by my colleagues. In particular, I integrated our W Reconstruction code written last summer into the framework and introduced several variables based on this code. These variables represent the first steps in MVA optimization for this channel.

In addition to this role, I was also responsible for key software development within the wh_cafe framework. Much of the functionality of the framework was restricted to the $WH \to l\nu bb$ and $H \to WW \to l\nu jj$ channels, and would not work for $WH \to WWW \to l\nu.jj.jj$. In particular, I was responsible for writing code to enable the processing of the Multijet and Final MVA's for this channel.

I also modified a large portion of the WH_LimitManager class in wh_cafe. This class provides the final output of the framework before viewing the final results. Though it was a relatively small class (under 200 lines of code) it was an essential part for setting the exclusion limits for the process. It was also the most ad hoc component of the wh_cafe framework. To adapt the class for our analysis, I had to make a significant amount of changes to get it to work.

3.1 Minimizing Multijet Background

Multijet background results from a group of jets that have been produced due to the strong interaction which mimics our signal. In order to match a final state similar to the $WH \to WWW \to l\nu.jj.jj$ signal, a group of five jets must be produced in the detector, with one of the jets getting misreconstructed as a lepton. Unfortunately, in our channel the final

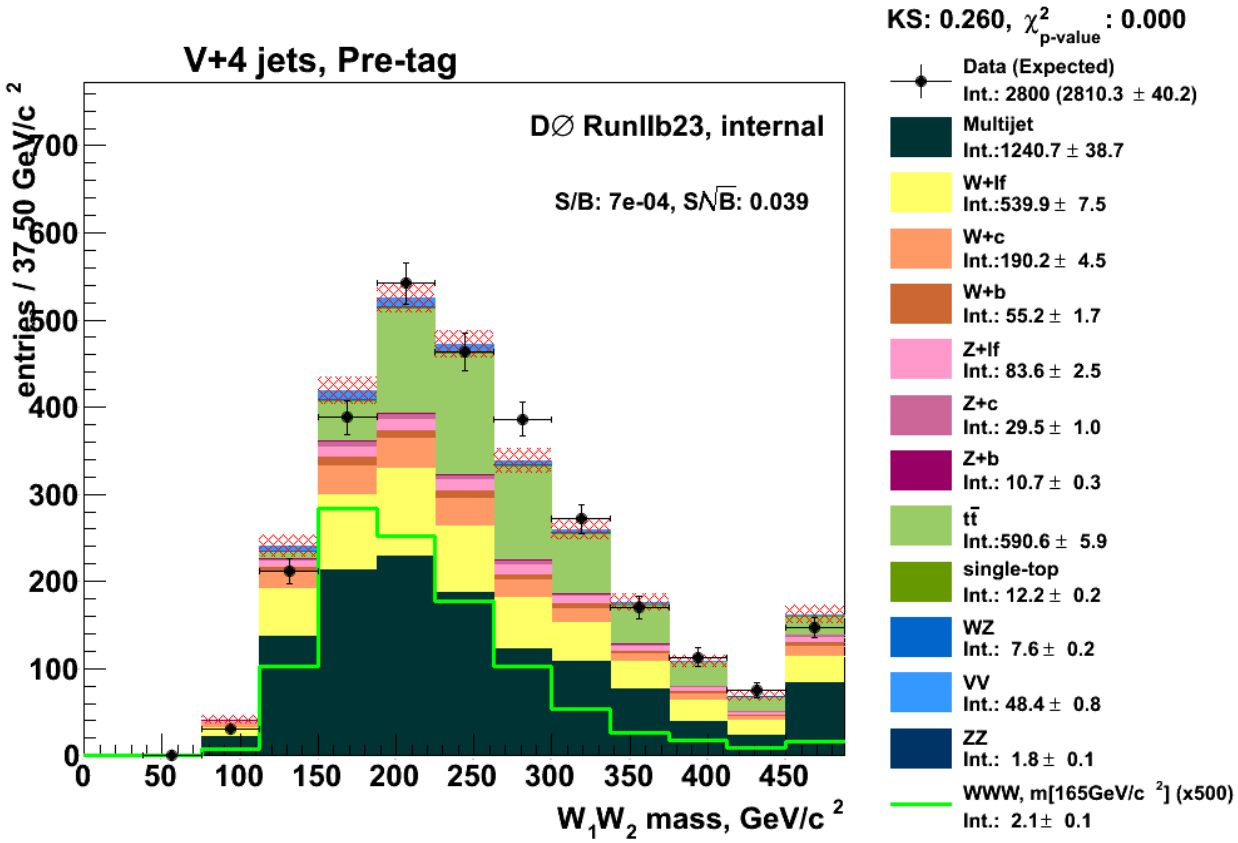
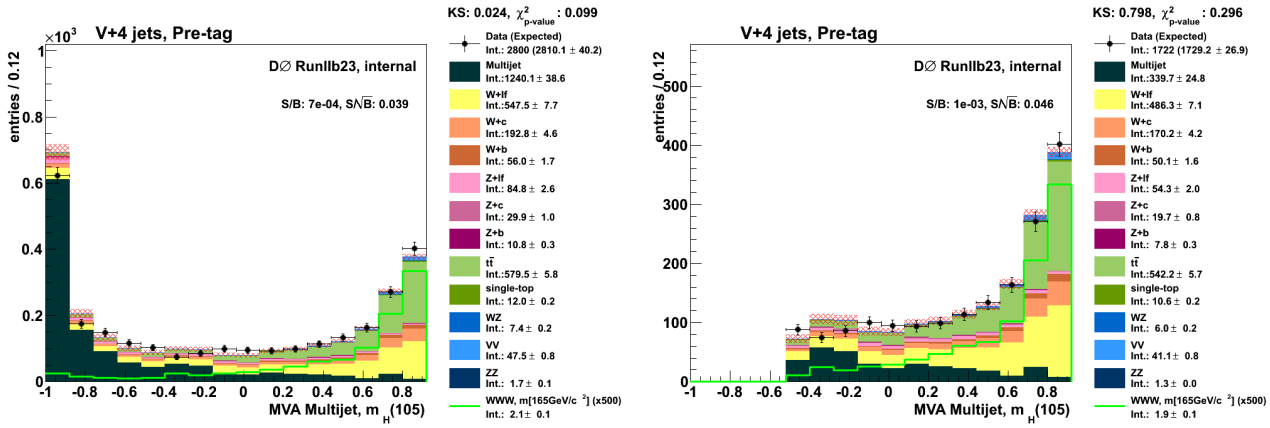


Figure 3.1: Mass of W_1 & W_2

Multijet background (dark green) is a dominant form of background for this channel. It corresponds to 1240 Events that we see in the data. Note that the $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ signal has been multiplied 500× to be visible on this plot. The signal itself represents only 2.1 events.



(a) Before MVA QCD Cut

(b) After MVA QCD Cut

Figure 3.2: Effect of MVA QCD Cut

Cutting on the Multijet MVA output resulted in a 66.3% reduction in Multijet Background with a 9.5% loss of signal, corresponding to a 52.2% improvement in the Signal to Background ratio. Note the change of scale from 1000 to 600 Events on the y-axis

---

decay state is mostly jets, which means that there are many opportunities for jets to "fake" a lepton.

The kinematics of Multijet phenomena are very complex, and can vary in different mass ranges. For this reason, Multijet background is one of the most difficult types of background to work with. To help discriminate Multijet background from signal, I made use of the newly created MVA code in wh_cafe to train and apply an MVA. The BDT method was particularly useful at developing a classifier to discriminate Multijet background from other sources. Using this MVA output, we would be able to generate a variable on which we could place a cut that would remove a significant amount of Multijet background while preserving the majority of our signal. For our analysis, we chose to make a cut on our 105 GeV Multijet MVA output (See Fig 3.1), since with the default variables the output of this classifier separated the largest amount of Multijet background from the signal. The overall outcome is shown in Figure 3.2.

3.2 Reconstructing W's from Jets

Since we can only detect the final state of the channel, we have to infer the intermediary steps of the process. As can be seen in Figure 1.3, a W boson can either decay into a lepton-neutrino pair, or a quark-antiquark pair. Thankfully, many of the variables we needed for the $W \to l\nu$ process were already provided in the framework by the WH Group. However, reconstructing the two W's from the $WW \to jjjj$ process required some thought. Since there were four jets in the signature, it was hard to tell what jets came from which W. There were three combinatoric possibilities: (1) $W \to j_1 j_2$ & $W \to j_3 j_4$, (2) $W \to j_1 j_3$ & $W \to j_2 j_4$, and (3) $W \to j_1 j_4$ & $W \to j_2 j_3$.

Eventually we decided that the best way to do this would be to sort the "W's" by their deviation from the average W mass of 80.399 GeV[4]. This was possible due to the ROOT TLorentzVector class. TLorentzVector is a special type of vector that is used to describe the particle tracks that are reconstructed from detector data. It has special methods that allows for easy access to particle kinematic properties, such as the mass, momentum, and angular properties. The wh_cafe framework provides us with TLorentzVectors associated with the lepton, neutrino, and each jet. Using these vectors, we could "add" the jet vectors together to produce W boson vectors. The algorithm is described below:

1. Generate each jet combination (12_34, 13_24, 14_23)
2. Calculate the mass of each jet pair.
3. Calculate Error in each W mass by using $\Delta m_{ij} = m_{ij} - m_W$, where $m_W = 80.399$ GeV[4]
4. Sum the errors together. $E[m_{ij\_kl}] = |\Delta m_{ij}| + |\Delta m_{kl}|$
5. Select the combination with the lowest summed error

When we have determined the selection of the W's, we label the least massive jet pair W_1 and the other W_2. We do this to help differentiate between the two W's. Depending on the mass of the W, there is a chance that the W is not "on-shell." This means that the W is not massive enough to live a perceptible amount of time—it decays the instant that it is produced. In this case, the physics governing the behavior of off-shell W bosons is

fundamentally different from that of on-shell W bosons. By requiring this criterion, W_1 is more likely to be off-shell and W_2 is more likely to be on-shell, giving greater significance to the variables that are derived from them. A complete list of W-based variables that I added to the framework is shown in Table 3.1.

3.3 Final MVA

Using these new variables in conjunction with several already defined in wh_cafe, we were able to train Final MVA's and analyze the output. Unlike the Multijet MVA, the Final MVA took into account all the different background processes for training. Unfortunately, many background processes, such as the $W + lf$ background, are very difficult to discriminate against. The $W + lf$ background consists of a W boson produced with light-flavor quarks, which appears very similar to our signal. Currently, there are no variables that have substantial discriminating power against this particular background. For this reason, the MVA performance does not appear to separate the signal and background as well as we did when trained only with signal and Multijet Background. An example of the Final MVA output is shown in Figure 3.3. Only after some substantial investigation into new variables will we produce better Final MVA output.

4 Results

Towards the end of the summer, the joint effort of the entire WWW team came together to produce preliminary results. In order to produce the sensitivity plots, we fed the Final MVA plots into the WH_LimitManager class. The LimitManager did the necessary computations to produce inputs to COLLIE, DØ's confidence limit analyzing software. After processing our data in COLLIE, we were able to generate the preliminary sensitivity plot as shown in Figure 4.1. The end result is sensitivity to the Higgs from 148–190 GeV to twenty times the Standard Model. Though this plot has not taken into account statistical and systematic

Table 3.1: List of Added W Variables

jp12_angle	jp1lnu_costheta	jp2_cf_sum_W2
jp12_costheta	jp1lnu_m	jp2lnu_angle
jp12_m	jp1lnu_mt	jp2lnu_costheta
jp12_mt	jp1lnu_sigma_12	jp2lnu_m
jp12_sigma_12	jp1_m	jp2lnu_mt
jp1_cf_12_W1	jp1_pt	jp2lnu_sigma_12
jp1_cf_21_W1	jp2_cf_12_W2	jp2_m
jp1_cf_sum_W1	jp2_cf_21_W2	jp2_pt
jp1lnu_angle		

In order to avoid conflict with names already in the wh_cafe framework, a new naming convention needed to be invented. The names needed to clearly state what physical quantities are being represented. In this convention 'jp' ("jet pair") represents a W reconstructed from jets and 'lnu' represents the W reconstructed from the lepton and neutrino. For example, 'jp12_angle' represents the angle between jet pairs 1 and 2, which we call W_1 and W_2. 'cf' stands for colorflow, which was studied in detail by Stephanie Hamilton. 'costheta' is the cos of the polar angle θ, which is used to derive η (See Figure 2.1). 'mt' and 'pt' refer to the transverse mass and momentum respectively. 'sigma_12' is a derived quantity based on the momentum and position of the particles with respect to each other.

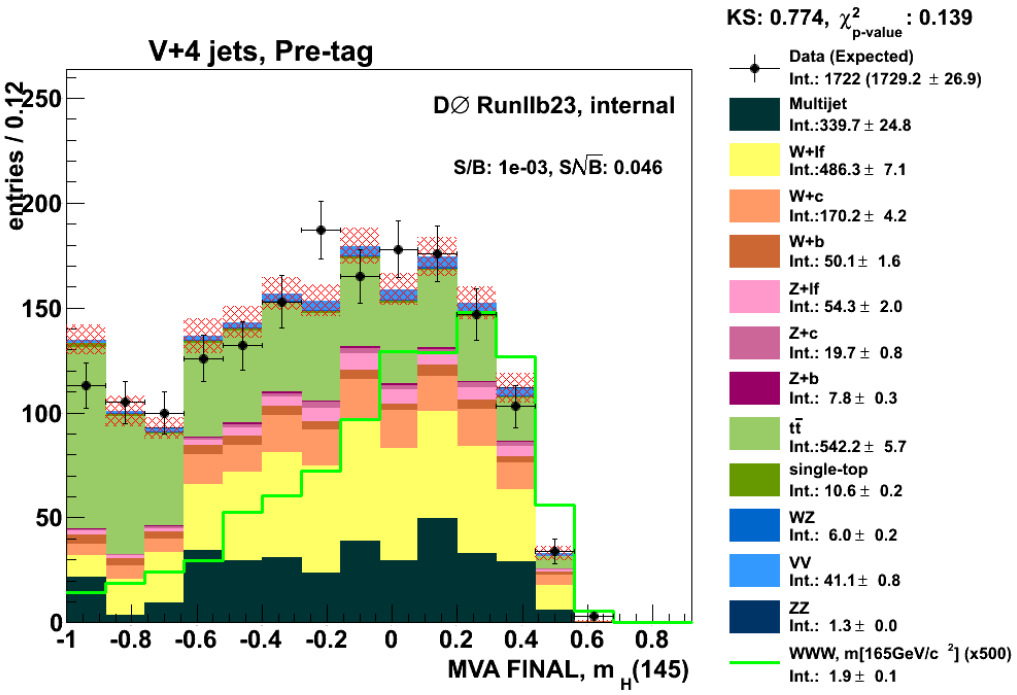


Figure 3.3: Example Final MVA Output

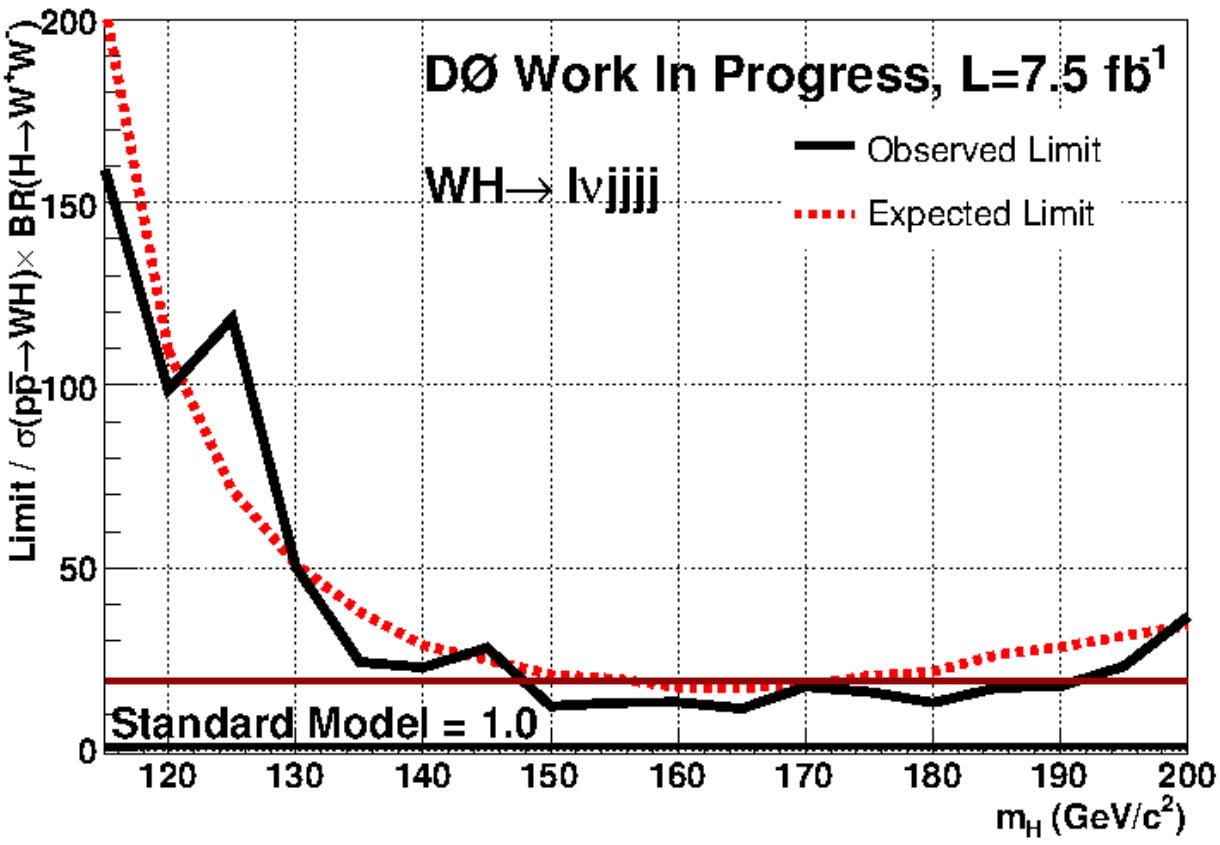


Figure 4.1: Preliminary Sensitivity Plot for $WH \rightarrow WWW \rightarrow l\nu.jj.jj$

The red line shown corresponds to the $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ process occurring at 20× the Standard Model.

---

uncertainties, it represents one of the final stages of the analysis.

5 Future Work

Now that we finally have a functioning framework for this channel, it could not be a more exciting time to be working on this analysis. The framework is stable enough that the analysis group can now begin to refine the analysis techniques. In particular, we need to start looking for new ways to discriminate the signal from the background. This begins by generating more variables within the framework. Our Final MVA's have a long way to go

before they can compare with the quality of the WH Group's Final MVA's, but they can only get better with time. During this coming semester, I plan on continuing to collaborate with the WWW team by remotely working on this analysis, hopefully optimizing it to the point where the sensitivity drops down to ten times the Standard Model or less.

6 Acknowledgments

During my time working for the Summer Internships in Science and Technology program, I have had the privilege of working with Dr. Ryuji Yamada, returning IMSA colleague Alex Abbinante, IPM intern Youssef Sarkis Mobarak, and fellow SIST intern Stephanie Hamilton. This analysis continues to receive much support from the DØ Collaboration. In particular, I would like to recognize Dr. Michael Cooke and the rest of the WH Group for their continued patience and guidance this summer.

7 Work Cited

References

[1] W. Murray, "Higgs Searches at the LHC" July 2011 [Presentation] Available: Europhysics Conference on High-Energy Physics, http://indico.in2p3.fr [Accessed July 28 2011].

[2] CDF, “Search for New Particles Decaying into bb and Produced in Association with W Boson.” [Online] Available: http://www-cdf.fnal.gov. [Accessed July 28 2011].

[3] Z. Hynes, “Search for the Standard Model, High-Mass Higgs Boson in the $WH \rightarrow WWW \rightarrow l\nu.jj.jj$ Channel,” August 2009. [Online]. Available: SIST Presentations 2009, http://sist.fnal.gov. [Accessed July 28 2011].

[4] Particle Data Group, “Particle Physics Booklet,” July 2008, [Booklet]. Available: http://pdg.lbl.gov/pdgmail. pp. 8.

[5] TMVA Developers, “TMVA Users Guide,” July 2008, [Online]. Available: http:tmva.sourceforge.net. pp. 104–118. [Accessed July 12, 2011].