2012

Validating Monte-Carlo Distributions in the Search for the Higgs Boson

Andrew Ackert

Abstract:
(Higgs, Monte-Carlo, kd-tree)
The goal of this research was to develop a method of validating multi-parameter Monte-Carlo (MC) simulations. These simulations mimic the proton-proton collisions at the Large Hadron Collider (LHC) that are being used to search for the Higgs boson. The basic idea is to construct partitions, count the number of data points that lie within each partition, and apply the partition and counting procedures to both collision data from the LHC and data created via Monte-Carlo simulations. In principle, the method developed can be used with any measure of dissimilarity between distributions with any number of parameters. In this study, we used Fisher's test (F-test) applied to 2 parameters of the data (btag and di-muon mass). Through the F-test it was concluded that, at least in these 2 parameters, the LHC data and the MC data were in agreement.

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS & SCIENCES

VALIDATING MONTE-CARLO DISTRIBUTIONS

IN THE SEARCH FOR THE HIGGS BOSON

By

ANDREW K ACKERT

A Thesis submitted to the
Department of Physics
in partial fulfillment of the requirements for graduation with
Honors in the Major

Degree Awarded:
Spring, 2012

The members of the Defense Committee approve the thesis of Andrew Ackert defended on Aprill 11, 2012

_____

Dr. Harrison Prosper
Thesis Director

_____

Dr. Mike Mesterton-Gibbons
Outside Committee Member

_____

Dr. Andrew Askew
Committee Member

_____

Dr. Todd Adams
Committee Member

I. INTRODUCTION

A. The Higgs Boson and the Origin of Particle Mass

The Standard Model of particle physics (SM) has been used with great success for the past several decades to describe the fundamental particles that make up the universe as we understand it today. One major breakthrough in creating the Standard Model was the formulation of the unified theory of the electromagnetic and weak forces by Glashow, Weinberg, and Salam (for which they won the 1979 Nobel Prize in Physics) [1]. The interactions of the electroweak field can be described by a mathematical quantity called a Lagrangian. The original electroweak Lagrangian of Glashow contains four massless electroweak gauge fields, associated with four bosons which are the field quanta. However, in nature there is only one massless electroweak boson: the photon [2]. (See the Appendix A for some details.)

The Higgs mechanism was incorporated by Weinberg and Salam into Glashow's theory in an attempt to explain the electroweak symmetry breaking that causes particles to become massive. The Higgs mechanism is based on ideas by Higgs, Brout, Guralnik, Englert, Hagan, and Kibble. The Higgs mechanism begins with particles that are inherently massless (as in the original electroweak theory of Glashow). According to special relativity, massless particles travel at the speed of light, that is,at luminal speed. However, when these particles interact with the Higgs field, they are slowed down to sub-luminal speed; that is, they acquire mass. The degree to which particles interact with the Higgs field determines the amount of mass particles attain. Particles that interact weakly with the Higgs field, such as neutrinos, attain very little mass and thus still move at near luminal speed. Photons, however, remain massless and therefore do not interact with the Higgs field. As with other fields, such as the strong, weak, and electromagnetic (EM) fields, the evidence for these fields lies with discovering the field quanta. For the Higgs field this is the Higgs boson.

According to quantum field theory, when a field has infinite range (such as the $1/r^2$ dependence of the EM field), the field's quanta (for example, photons) are massless. Consequently, since the Higgs field has an infinite range, its quanta should be similar to the EM quanta and be massless. But the self-interaction of the Higgs boson induces a mass onto itself, thereby limiting the field's range [3]. A similar effect is seen when the EM field comes upon a superconducting material. Normally, the EM field is infinite in range and can

penetrate through an object. However, due to the Meissner effect in superconductors (the exclusion of electromagnetic fields from the interior), the penetration depth for the EM field becomes finite [4]. Again, this can be interpreted as the field's photons attaining mass in the region of the superconductor's interior. By a similar process, when the W, Z, and Higgs bosons interacts with the Higgs field, their ranges become finite and therefore the bosons can be seen as acquiring mass [2].

B. Project Motivation

The search for the Higgs boson that is currently underway is designed to test whether this hypothesis about the origin of mass is correct. According to the Standard Model, the Higgs bosons can be created in high-energy proton-proton collisions, such as those being carried out presently at the LHC. At some stage, the search for the Higgs boson requires measuring its mass from the collision data. To determine which events (recorded data from particle collisions) are more likely to be those that contain the Higgs (the signal) and those that do not (the background, which can mimic the Higgs signal) a mathematical tool called a discriminant must be constructed. Using this tool, weights can be assigned such that the more signal-like events tend to have a weight near 1 and those that are background-like tend to have a weight near 0.

Traditionally, "data-driven" methods have been used to model the background events and discriminants are constructed using these events together with simulated signal events. In the data-driven methods, the background events are modeled using real collision data. However due to the elusive nature of the Higgs, there are two potential problems: 1) it is necessary to model the background events using events that are not the same as the ones which are used in the search for the signal and 2) the events used for modeling the background may contain signal events that have not been properly accounted for. In order to avoid these potential problems with the"data-driven" methods, the Florida State University High-Energy Physics (FSUHEP) group is exploring a different approach: to use simulated events for both the background and the signal. However, this method relies on a crucial assumption: that the simulations are an accurate representation of the real-world data. This provides the motivation for this project, which is to devise a means to validate the MC simulations.

The common tools that are utilized to compare data sets, such as the chi-squared test, fail or have great difficulty when it comes to dealing with data of higher dimensions. The usefulness of the method devised during this project is that it resolves the higher-dimensional data down to a set of 1 dimensional quantities, which is relatively easily handled by many tools of statistical analysis.

In order to reduce the dimensionality of the data, we used a "kd-tree." A kd-tree is a binary partitioning method that can partition a set of data with k data points in d-dimensions.[5]For the purposes of this research, the partitioning was restricted to 2 dimensions, but the process can be extended to any number of observables. The partitions are then used as "bins" whose boundaries are recorded. Knowing the bin boundaries, the bins can then be superimposed upon data sets, and the number of data points that fall within each bin can be counted. The counts for the respective data sets are then a simple list of numbers which can be compared via a statistical test such as the F-test, to see if the real and simulated data sets are inherently the same or not.

The F-test tests if sets of data obey the null hypothesis; that the difference between the counts are due only to random fluctuation, and not as a result of the two sets being intrinsically different. The test of this hypothesis is done by calculating an F-value (see Appendix B for the algorithm) for the sets of data, and comparing it to a critical F-value determined by the probability distribution function of F-values under the null hypothesis. If an F-value is above the critical F-value, then the null hypothesis that the sets are in agreement is rejected. In order to utilize the F-test however, we need the probability distribution function of the F-values.

C. Basic Method

This method can be broken down to several steps:

1. Construct a kd-tree for the LHC data.

2. Create bootstrap samples (see Section 3B) from the LHC data.

3. Using the samples, apply the kd-tree to count the data points that fall within the bins.

4. Calculate the F-values for pairs of samples in order to approximate the F-distribution.

5. Create a probability distribution of the F-values.

6. Use this distribution to compute critical F-values.

7. Apply the kd-tree to count and record the data points that fall within the partitions for the entire MC and LHC data.

8. Compare the F-value to the critical F-value to see if the null hypothesis is rejected.

II. THE DATA

The data from the LHC, collected with the CMS Experiment, and MC simulation were provided by Dr. Prosper. The CMS (Compact Muon Solenoid) experiment is one of two large multi-purpose particle detectors at the LHC (the other being the ATLAS experiment). The CMS consists of several layers of particle detecting structures wrapped concentrically around the particle beam axis. These structures are designed to detect different types of particles.

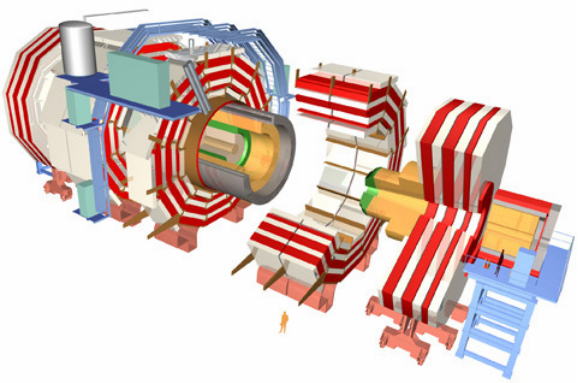

FIG. 1: Above is a 3D rendering of the CMS detector. [6]

Both real and simulated data sets consist of multiple observables per event, however, in this project we restricted attention to two observables: the di-muon mass $m_{\mu\mu}$ and the *btag* (btag explanation on page 6). The mass $m_{\mu\mu}$ is calculated from the sum of the 4-vectors of the two leptons produced in the reaction. In addition, the MC events have an event-by-event weight. These weights are numbers such that their sum over all of the events was equal to the number of events actually observed. The data provided were from proton-proton collisions resulting in the production of two muons (di-muons):

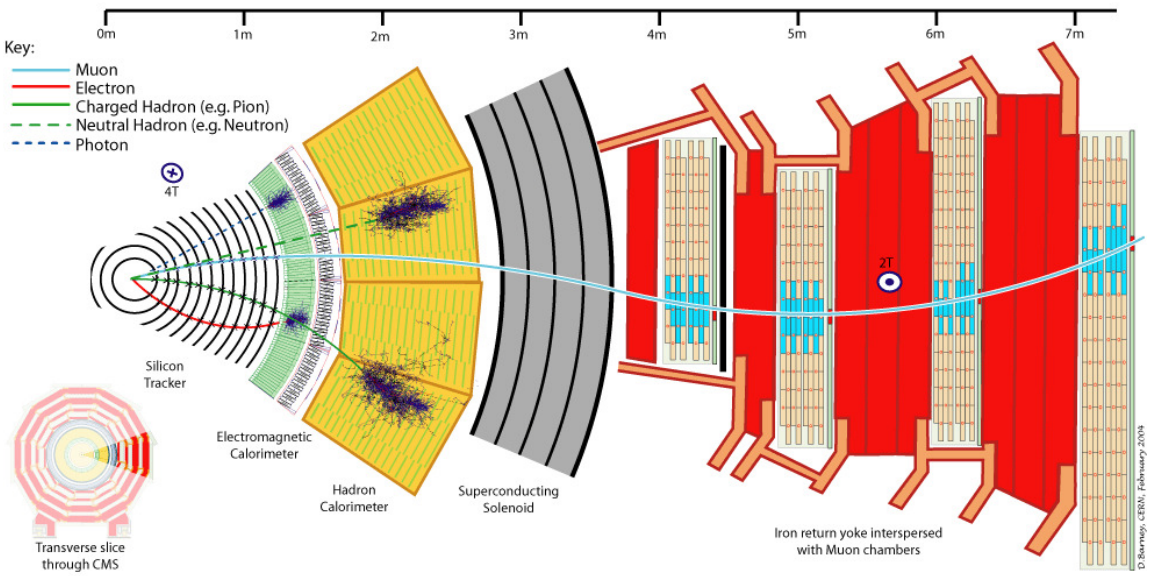

FIG. 2: Above is a slice of the detecting components of the CMS detector showing where different types of particles would be detected. [7]

$$p + p \rightarrow \mu^+ + \mu^- + X \tag{1}$$

The lifetime of the Higgs boson causes a difficulty. For a low-mass Higgs boson, the lifetime is predicted to be roughly 10^{-22} s. Therefore, the Higgs boson itself cannot be directly observed, only its decay products can. The most promising decay modes of primary research interest at the LHC are,

$$p + p \rightarrow H(\rightarrow \gamma\gamma) + X \tag{2}$$

$$p + p \rightarrow H(\rightarrow ZZ \rightarrow \ell^+\ell^- + Y) + X \tag{3}$$

$$p + p \rightarrow H(\rightarrow WW \rightarrow \ell^+\ell^- + Y) + X, \tag{4}$$

where X and Y denote all the other non-specified decay products, and $\ell^+\ell^-$ are two charged leptons. In the case, of ZZ, Y could be another lepton pair. In this project, we focused on the $ZZ \rightarrow \mu^+\mu^- + Y$ mode. However, what is much more statistically probable is that the intermediate step of the Higgs boson is skipped, and the proton collision results directly in ZZ being produced or, what is even more probable, $Z + X$ [8].

The *btag* observable is a measure of the b-jet content of an event. In practice, it is a measure of the degree to which the vertices of the jets are displaced from the collision point

of the proton beams. This quantity permits identification of processes that create b-quarks, such as the process:

$$\begin{aligned} &p + p \rightarrow t\bar{t} + X \\ &t \rightarrow b + W^+ \rightarrow \mu^- + \bar{\nu}_\mu + X \\ &\bar{t} \rightarrow \bar{b} + W^- \rightarrow \mu^+ + \nu_\mu + Y, \end{aligned} \tag{5}$$

Such an event would be recorded since it contains a μ^+ and μ^- along with the b-quarks. However, this would not constitute a signal since the target Higgs boson modes shown in equations (2) - (4) do not undergo such a process. The term "btag" comes from the fact that b-quarks will not show up directly in a detector due to their short lifetimes, but the quarks will however travel a certain distance from the collision point before decaying into a particle jet. Since b-quarks are contained in many reactions, identifying them can be very useful in reconstructing the collision reaction. [9] The arbitrary value of -1 is assigned to jets emanating from the interaction vertex (within error), with larger values of the *btag* observable indicating the jet vertex occurring further away from the collision point. These displaced jets are more likely to emanate from b-quarks and therefore less likely to be due to one of the Higgs boson signals listed in equations (2) - (4).

The difficulty in collecting possible signals of the Higgs boson is also increased due to the data recording rate at particle detectors. Particle beams collide millions of times each second, however the technology currently does not exist to store or analyze the massive amount of information at that rate. Instead, the detectors are programmed to record data only when certain conditions, "triggers," are met. Currently, the recording rate is approximately 500 events per second. Regardless of the technological limit to the recording rate, a staggering amount of data is collected at the LHC.

The number of events that occur, N, is given by the total cross-section, σ, multiplied by the integrated luminosity, $\mathcal{L}$. Here total cross-section is a measure of the probability of a particle collision yielding a particular kind of event, for example one containing a Higgs boson. Integrated luminosity is a measure of the number of particles that have crossed a unit area in a given amount of time. It is therefore the integral of the luminosity, which is a measure of the intensity of the beams. In particle physics, cross-section is usually given in barns ($1\text{b} = 10^{-24}$ cm^2).

Current results indicate that the SM Higgs mass must be somewhere between 110-140 GeV. As shown in blue in Fig. 1, the SM predicts a total cross section of $\sigma \approx 10$ pb for the

Higgs decay modes being studied, at center of mass energy = 7 TeV. This implies that in 2011, in which the CMS Collaboration collected $\mathcal{L} \approx 5200$ pb^{-1} of data [12], that roughly 52,000 Higgs boson events were created! Unfortunately, however, the number that can be identified is at least 10^3 times lower. Thus out of the millions of events captured by the LHC in 2011, only a very small fraction may be signals of the Higgs boson. Therefore, in the data-driven methods of estimating the background, it is fair to say that signal content is negligible. But, as we approach the Higgs discovery this will no longer be true.

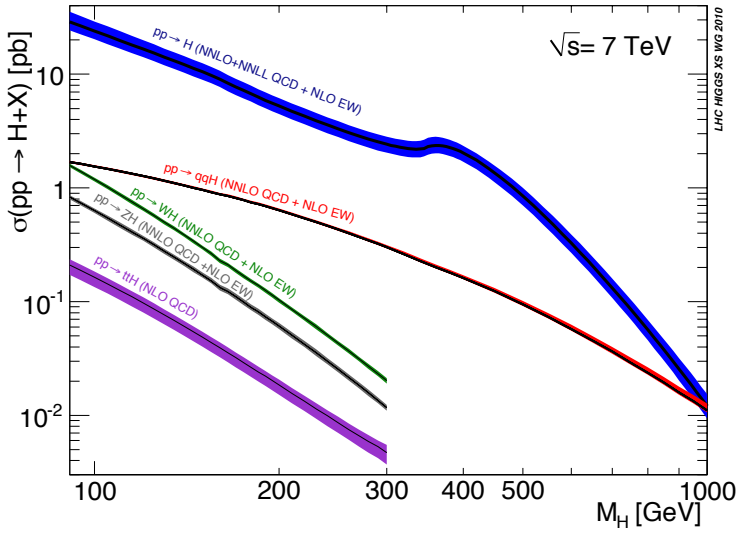

FIG. 3: Cross-section as a function of the SM Higgs boson mass at $\sqrt{s} = 7\ TeV$ [10]

Herein lies the benefit of utilizing MC simulations. With MC simulations, the exact number of Higgs signals can be controlled and manipulated, allowing the user to determine precisely what characteristics should be looked for to identify the Higgs signals, and how they differ from those of the background events. Utilizing this knowledge can give insight as to which real-world events are signals, and which are background. But, as noted above, it is vitally important that the MC simulations are trustworthy.

III. METHOD

A. Creating the kd-tree and counts

Kd-trees are a method of binary partitioning, in which each successive round of partitioning, new partitions are created from the previous round's partitions. For the first round,

1 partition separates the initial data set into two equal sub-sets. These two sub-sets, are then partitioned again such that after this second round of partitioning, there are 4 sub-sets. The kd-tree used in this project was restricted to partition data in 2 dimensions only along median points. The hope was that by using median partitions, the final subsets, or "bins" would contain approximately the same number of entries in each. Another design feature of this kd-tree is that after each round of partitioning in one direction, the next round of partitions would be perpendicular to the previous direction. An example of this is given in Fig. 2. Here the first round of partitioning occurred at x = 7 in the y-direction. This was because x = 7 was the median value. During the next round, the program focused on y-values in the two domains left and right of x = 7. During this second round, the left sub-set of data contained a median point at y = 4 and the right had a median of y = 6. These were the points along which partitions were created in the x-direction, denoted by the blue lines in Fig. 4. Had we chosen to stop the program after two rounds, this would have left 1 data point in 3 out of the 4 created sub-sets, with 0 data points inside the upper right sub-domain. In the example shown in Fig. 2, a third round of partitioning was carried out in order to show how the partitioning once again alternated to choosing x-values for the boundaries.

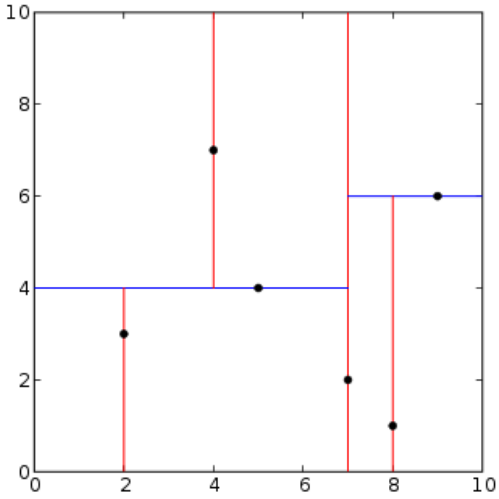

FIG. 4: Above is a simple example of a 2-dimensional Kd-tree.

Using the data from the LHC, the kd-tree program was applied to the first 10000 data entries using $btag$ vs $m_{\mu\mu}$. Staying in 2 observables aided in the ability to visualize how

the partitioning was being done and would require little augmentation to the programming in order to create a higher dimensional kd-tree. After 8 rounds of partitioning, 256 bins were created and the boundaries for these bins were then outputted to a data file to be applied for the remainder of the study. We choose 8 rounds of partitioning because this left us with a sufficient number of bins to apply to other data, while not having too many bins which would have resulted in lower counts per bin. The goal of the binning was to have as many bins as possible consistent with having a sufficient number of counts per bin such that fluctuations between the counts of different bins would be small compared to any deviation between the LHC and MC simulation data. We believe that the partitioning after 8 rounds was sufficient to meet this design goal.

Using ROOT's graphing macros, we plotted these first 10000 data points and superimposed them on the newly created bins. We expected that as the density of data points increased, so would the density of the bins, if our understanding of the kd-tree process was correct. This indeed turned out to be the case, as shown in Fig. 6.

Now that the bins have been established, the next step was to count how many data points lie within each bin. By design, the data points that lie along the boundary were not included in the count and we were left with a list of bins and their corresponding counts.

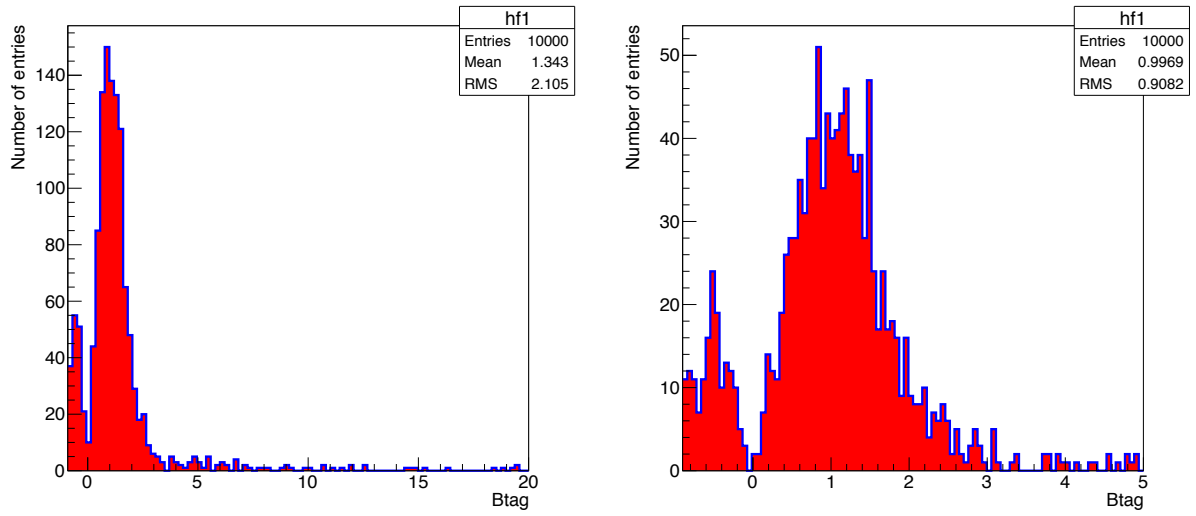

FIG. 5: Above are the btag distributions for the first 10000 entries.

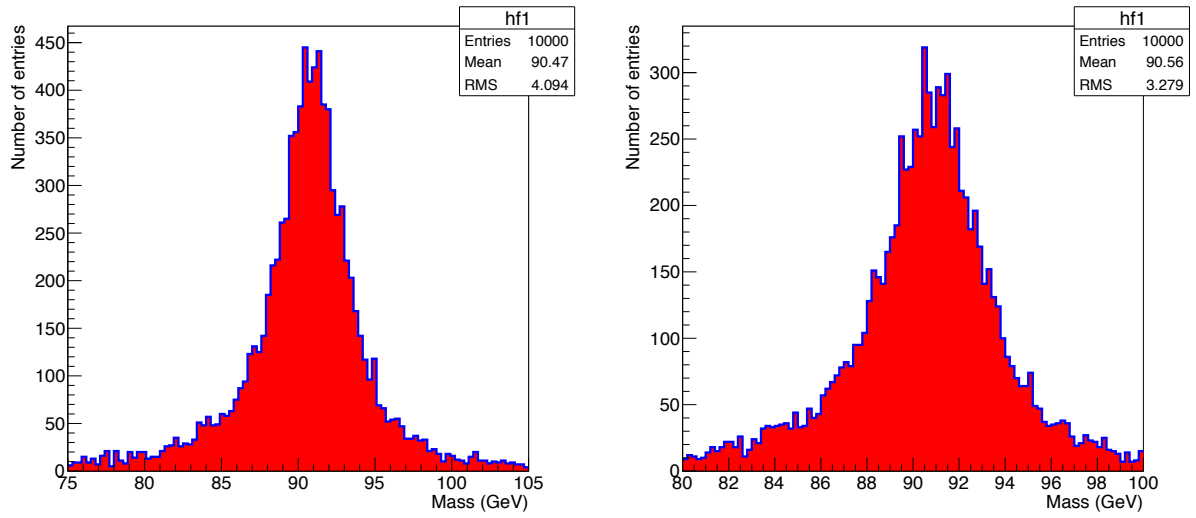

FIG. 6: Above are the di-muon mass distributions for the first 10000 entries.

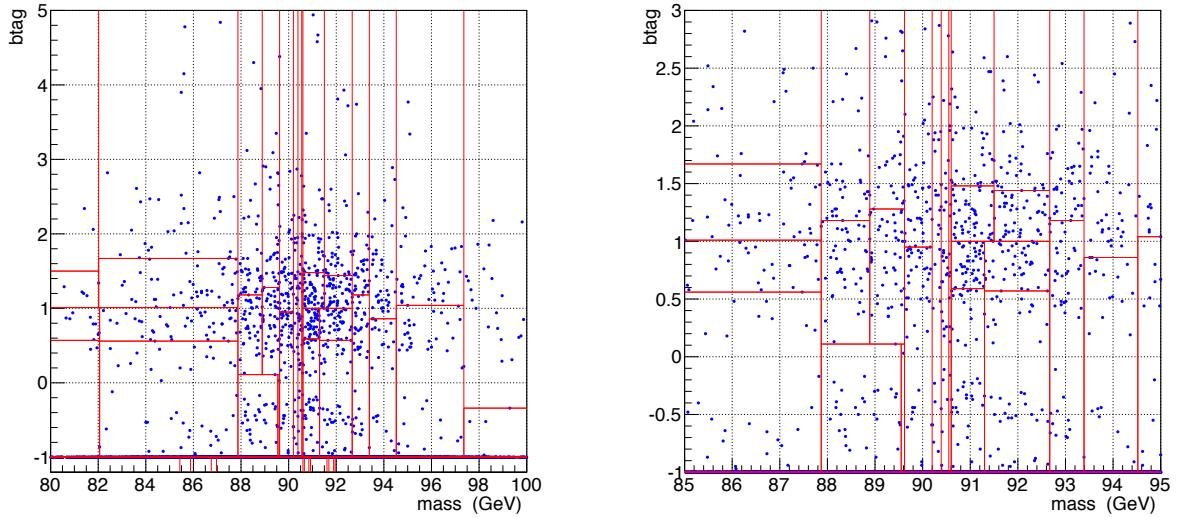

FIG. 7: Above is the kd-tree partitioning for the first 10000 entries of btag vs mass. As expected, since the data are clustered around 91GeV, so too do the bins.

B. Bootstrap sampling of LHC data

Random sampling with replacement, or "bootstrap sampling" as it is referred to, is the process of randomly selecting members of a set to create a new subset. The key is that when a member of the original set is selected, it is then able to be selected again. The bootstrap sampling allowed us to simulate the F-distribution empirically by applying the F-test to

many pairs of bootstrap samples. This allowed us to establish a distribution of F-values for sets that were inherently from the same larger set. Since they were from the same larger set, they satisfied the null hypothesis that any discrepancies between distributions were due to random fluctuations. This distribution was then used as the basis for computing the critical F-values to see if the MC data and LHC data satisfy the F-test.

To do this, the LHC data file was read in and the number of lines of data (N) was counted. A number between 1 and N was then assigned to each entry. Using a random number generator, an entry was selected and then the program stored that entry's information into a new data set. The program then randomly selected another number between 1 and N and repeated the process a specified number of times. It then outputted the newly created set to a new data file, or "bootstrap sample."

Using this sampling method, from the approximately 1.6×10^6 entries of the LHC data, 10^5 sets each containing 10000 data points were created. We then applied the original bin boundaries created from the first 10000 data points to each of the bootstrap samples. Using the counting program on each bootstrap sample, we attained 10^5 set of counts.

Upon examining different sets of the counts, it was discovered that some of the bins contained 0 count. This was largely due to data falling upon partition boundaries. For the bootstrap samples the vast majority of non-zero counts occurred in the same bins. This was acceptable since this maintained the design of the kd-tree, that the fluctuations between bin counts were negligible compared with potential discrepancies between sets.

These sets were used as the basis for creating the distribution of F-values, as the first 50000 entries were paired with the last 50000. The F-test was applied to each pair and the F-values recorded. Since the samples themselves were created randomly and not in any particular order, we limited any bias in the F-value distribution.

C. The F-test

Sir Ronald Fisher's test, colloquially known as the "F-test", was developed as a means to test whether the differences between the means among separate data sets was due to random fluctuations, or if the data sets differed to such an extent that they could be considered caused by different phenomena. The test was constructed such that when applied, a value was created for the sets of data in question and if this value fell below a certain critical value,

then the difference between the data sets was believed to be due to only random fluctuations (to within a certain degree of confidence that is determined by the critical value) [11]. The critical value is based upon the "alpha" value, where alpha is defined as the "tail" end of the area of a probability distribution function of the F-values:

$$\alpha = \int_{f_c}^{\infty} P(f)\, df \tag{6}$$

For the purposes of this project, we chose an alpha value of 0.05, such that any conclusions reached would be considered of 95 percent confidence level.

Various critical F-values (f_c in Eqn. (6)), are given in easily obtained charts, however these values are based on Sir Ronald's original research and only list critical F-values for selected degrees of freedom (see appendix B) and for Gaussian data. Therefore, we constructed our own F-value density function using the F-values from the bootstrap sample pairs. This density function was used to calculate our own critical F-values.

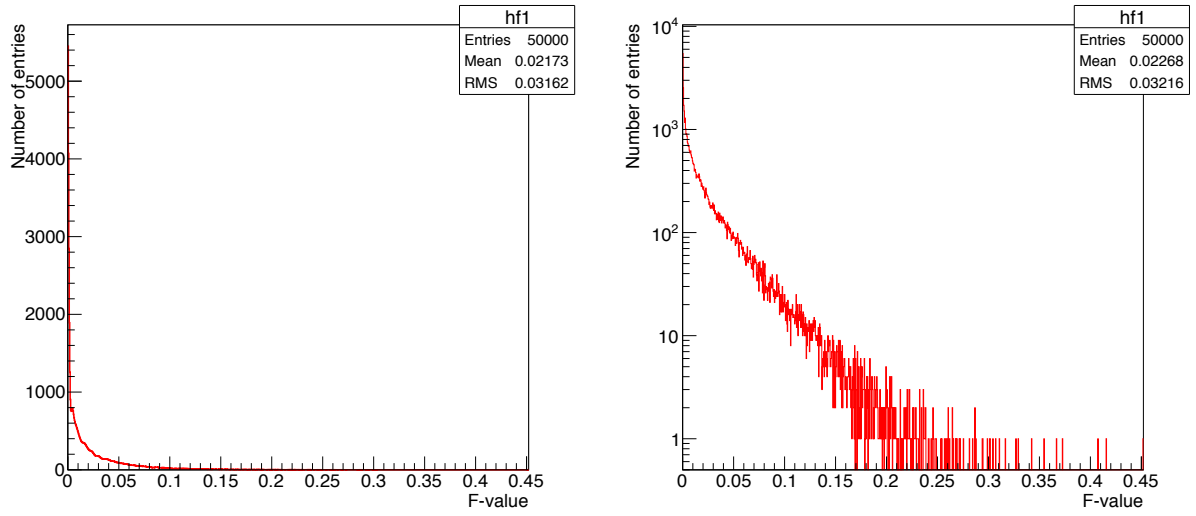

FIG. 8: Above is a histogram (left) of the F-values that resulted after the F-test was applied to the 10^5 bootstrap samples of the LHC data. The graph on the right is the left graph plotted on a log scale

Traditionally, the way to test the null hypothesis is to compare a F-value to a f-critical value. However, the way the ROOT histogram utility worked, it was simpler to feed into the computer a F-value and have it output the corresponding α value. Then, by seeing if the α value was larger than the desired α, it was equivalent to testing if the F-value was

below a desired critical F-value. The calculation of the α value was done by the following algorithm in Eqn. (7) applied to the values obtained from the histogram in Fig. 6:

$$\begin{aligned}
&\alpha = a_n + \frac{f_t - f_n}{w(a_{n+1} - a_n)} \\
&a_n = \text{the } n\text{th alpha value obtained from the histogram values} \\
&f_t = \text{the test F-value} \\
&f_n = \text{the nth F-value calculated from the histogram values} \\
&w = \text{the width of the interval between the F-values in the histogram} \\
&n = \frac{f_t}{w}
\end{aligned} \tag{7}$$

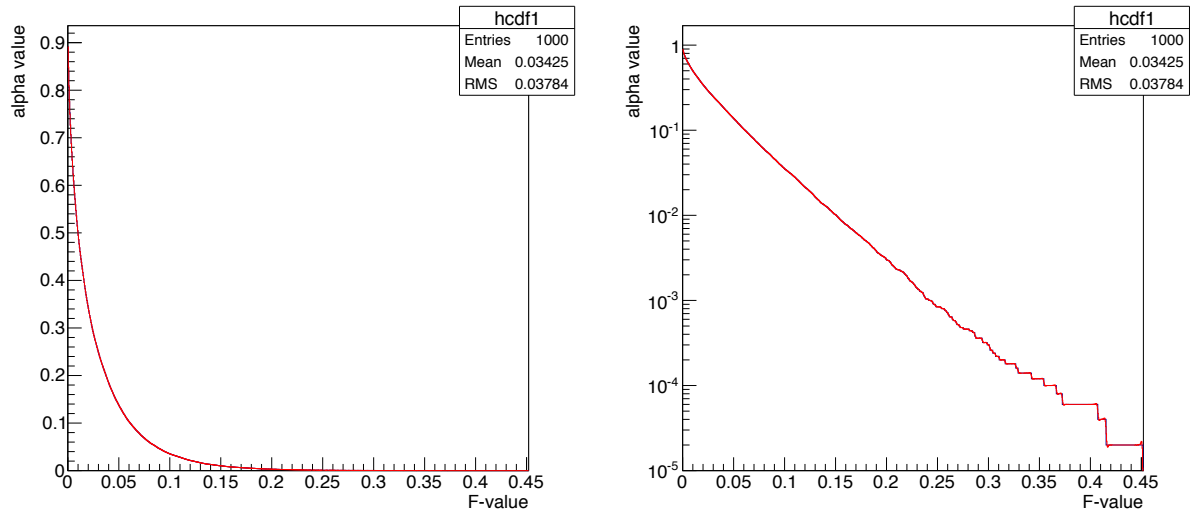

FIG. 9: The left graph shows α as a function of F-value. The right graph is the same data on a logarithmic scale. Alpha = 0.05 corresponds to an F-value of approximately 0.09.

As Fig. 9 shows, the lower the F-value, the greater the probability that the two data sets satisfy the null hypothesis of the F-test. Similar to the original F-distribution created by Sir Ronald, for two data sets that turn out to have zero difference between their standard deviations, the F-value was zero. In the case of the bootstrap sampling, this occurred roughly 400 out of 50000 times or roughly 0.8% of the time.

The last step for this project was to calculate the F-value between the entire LHC data set and the entire MC data set for the btag and $m_{\mu\mu}$ observables. We use the same partition that was used throughout the project, and created a list of counts for the two sets of data. For the MC data, the counts for each bin were the sum of the weights of the data points that

fell within the bin boundaries. Then the F-test was applied to these two final lists of counts. This produced an F-value of 0.0012. This F-value was then inserted into Eqn. (7) and the α value that resulted was approximately 0.78. (See Appendix B for F-value calculation).

IV. CONCLUSION/FUTURE

Since the alpha value obtained was 0.78, which was much higher than the alpha value of 0.05 that we were considering as the threshold for the null hypothesis, the F-test would suggest that to 95% confidence, the differences between the MC data and the LHC data are not caused by an intrinsic difference between MC data and LHC data. Hence, it would be reasonable to assume that the Monte-Carlo simulation in the two observables are a a close approximation to its real-world counterpart at the LHC.

The next step with this research would be to introduce a specific distortion to the MC data in order to see if the F-test is sensitive enough to notice and recognize that error. After this, the next step would be to expand the programs that were created for the various steps from 2 observables to more observables. This will be done by creating a modification to the kd-tree program in order to be able to create bins in higher dimension. Once that is completed, similar modifications to the binning and counting procedure will be added so that the program can apply higher order bins to the data in question, and be able to count which data points lie within the bins. Again, the usefulness of this process is that no matter what the dimensions of the data, the program will create appropriate bins and again undergo a counting procedure. The only issue is making sure there are enough events. The result will merely be a list of 1-parameter counts, regardless of the number of original observables. Then, we could again compare the entire MC data to the LHC data utilizing all the information available (in the form of higher dimensional analysis). Having done that, we could make a stronger claim as to whether the MC simulation is a suitable description of the LHC data.

From there, if we have established that the MC simulation is indeed accurate, a discriminant tool can be created such that given a function of signal counts $S(\tilde{x})$ and background counts $B(\tilde{x})$ (where $\tilde{x}$ is a multidimensional variable):

$$D(\tilde{x}) = \frac{S(\tilde{x})}{S(\tilde{x}) + B(\tilde{x})} = \begin{cases} \to 1 & \text{if more signal-like} \\ \to 0 & \text{if more background-like} \end{cases} \tag{8}$$

Once that is established, the discriminant ($D(\tilde{x})$) can aide in filtering which events are signals, and which events are backgrounds, and possibly aide in the measurement of the Higgs boson's mass.

Appendix A: Electroweak Symmetry Breaking

In the electroweak theory, fermions are specified as a left-handed weak isospin doublet:

$$\mathrm{L} = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \quad \text{(A1)}$$

with weak hypercharge $Y_L = -1$, and a right handed weak isospin singlet

$$R \equiv e_R \quad \text{(A2)}$$

weak hypercharge $Y_R = -2$

For the electroweak gauge group, $SU(2)_L \bigotimes U(1)_L$ (a gauge theory is a type of field theory in which Lagrangians are invariant under a continuous group of transformations), there are two sets of gauge fields: a weak isovector $\vec{b}_\mu$, spacetime-dependent, local, transform under a group of 2x2 unitary matrices with determinant 1 called $SU(2)$, with a coupling constant g, and a weak isoscalar A_μ that transform under U(1). Corresponding to these gauge fields are the field-strength tensors:

$$F^\ell_{\mu\nu} = \partial_\nu b^\ell_\mu - \partial_\mu b^\ell_\nu - g\varepsilon_{\jmath k\ell} b^\jmath_\mu b^k_\nu \quad \text{(A3)}$$

and

$$f_{\mu\nu} = \partial_\nu A_\mu - \partial_\mu A_\nu \quad \text{(A4)}$$

The overall Lagrangian can be written as the sum of the gauge Lagrangian and the Lagrangian for leptons:

$$\mathcal{L} = \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{leptons}}$$

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4} F^\ell_{\mu\nu} F^{\ell\mu\nu} - \frac{1}{4} f_{\mu\nu} f^{\mu\nu} \quad \text{(A6)}$$

$$\mathcal{L}_{leptons} = \bar{R} i\gamma^{\mu} \left(\partial_{\mu} + i\frac{g'}{2} A_{\mu} Y \right) R$$
$$+\bar{L} i^{\mu} \left(\partial_{\mu} + i\frac{g'}{2} A_{\mu} Y + i\frac{g}{2} \vec{r} \cdot \vec{b}_{\mu} \right) L \tag{A7}$$

The $SU(2)_L \bigotimes U(1)_Y$ contains four massless electroweak gauge bosons, $A_\mu, b^1_\mu, b^2_\mu, and, b^3_\mu$ in (A7), with nature having but one: the photon. The fact that these gauge bosons are observed to be massive means that the electroweak theory in this form is unrealistic. This was the motivation of the idea of "electroweak symmetry breaking" via the Higgs mechanism [2]

(Further discussion and greater formalism of the electroweak theory can be found in [2])

Appendix B: F-test Calculation

For two sets of data, with elements a_{1i} and a_{2i} the calculation of the F-value between these two sets is as follows:

Step 1. The mean of each set:

$$\bar{a_1} = \frac{1}{n_1} \sum a_{1i} \tag{B1}$$
$$\bar{a_2} = \frac{1}{n_2} \sum a_{2i} \tag{B2}$$
$$n_1 = \text{ number of entries in } a_1 \tag{B3}$$
$$n_2 = \text{ number of entries in } a_2 \tag{B4}$$

Step 2. The overall mean:

$$\bar{a} = \frac{1}{N} \sum \bar{a_i} \tag{B5}$$
$$N = \text{number of sets} \tag{B6}$$

Step 3. The "between-group" sum of squares:

$$S_B = \sum n_i (\bar{a_i} - \bar{a})^2 \tag{B7}$$
$$n_i = \text{ number of entries in set } a_i \tag{B8}$$

Step 4. The between-group degrees of freedom:

$$f_B = N - 1 \tag{B9}$$

Step 5. Between-group mean square value:

$$MS_B = S_B/f_B \tag{B10}$$

Step 6. "Within-group" sum of squares:

$$S_W = \sum_i \sum_j (a_{ij} - \bar{a_i})^2 \tag{B11}$$

Step 7. Within-group degree of freedom:

$$f_W = \left(\sum n_i\right) - N \tag{B12}$$

Step 8. Within group mean square value:

$$MS_W = \frac{S_W}{f_W} \tag{B13}$$

Step 9. F-value for the sets:

$$F = \frac{MS_B}{MS_W} \tag{B14}$$

---

[1] "The standard model." http://www-donut.fnal.gov/web_pages/standardmodelpg/TheStandardModel.html

[2] Quigg, Chris. "The electroweak theory." http://arxiv.org/abs/hep-ph/0204104

[3] Bednyakov, Giokaris. "On Higgs Mass Generation Mechanism in the Standard Model" arXiv:hep-ph/0703280v1

[4] "The Meissner effect." http://hyperphysics.phy-astr.gsu.edu/hbase/solids/meis.html

[5] "kd-tree" http://www.cs.sunysb.edu/~algorith/files/kd-trees.shtml

[6] "CMS experiment" http://public.web.cern.ch/public/en/lhc/CMS-en.html

[7] "CMS slice." http://www.bo.infn.it/~castro/research/CMS/CMS.html

[8] Schwarzchild, Bertram."The Large Hadron Collider yields tantalizing hints of the Higgs Boson"

Physics today, February 2012. pg16.

[9] Vanelderen, Lukas. "Challenging the Standard Model with the Compact Muon Solenoid in W/Z + jets studies and SUSY searches with b-jets."
pg 42.

[10] "LHC Higgs Cross Section Working Group." https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections

[11] "Sir Fisher's test." http://mathworld.wolfram.com/FishersExactTest.html

[12] "CMS luminosity." https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults

[13] "F-test by hand" http://www.stat.auckland.ac.nz/~wild/ChanceEnc/Ch10.wilcoxon.pdf