OBSERVATION OF THE HIGGS BOSON IN THE $H \to \gamma\gamma$ CHANNEL FROM DATA SIMULATED IN PYTHIA

JOE BENTLEY

School of Physics and Astronomy
Tutor: Kevin Ralley

ABSTRACT. In this paper the higgs data is analysed and we find the higgs with this statistical significance etc.

E-mail address: joebentley10@gmail.com.
Date: March 22, 2015.

1. Introduction

In this experiment particle collision data is analysed using multiple Python scripts to generate an invariant mass histogram and attempt to find the Higgs boson with some statistical significance. (WHAT STATISTICAL SIGNIFICANCE) The data is generated in PYTHIA as a flat text file, which is parsed by our parsing script to filter the events and calculate the invariant masses.

(TALK ABOUT INTEREST IN THIS EXPERIMENT:

- WHY IS IT HARD? - WHY WOULD YOU WANT TO DO IT? - BRIEF OUTLINE OF METHOD)

2. Theory

2.1. **Four Momentum.** The four momentum is similar to the classical three component momentum except generalised to four dimensional space-time, so it also has an energy component. [8] In the data supplied the four-momentum is given in the form (p_x, p_y, p_z, E), where p_z is aligned along the beam axis

The transverse momentum is the component of the momentum perpendicular to the beam axis. Since the z component of the four-momentum, p_z is aligned along the beam axis, the transverse momentum is given by,

$$(1) \qquad p_T = \sqrt{p_x{}^2 + p_y{}^2}.$$

The pseudorapidity is a measure of the angle between the particle and the beam axis, and is defined as,

$$(2) \quad \eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] = \frac{1}{2}\ln\left(\frac{|\mathbf{p}| + p_z}{|\mathbf{p}| - p_z}\right),$$

[8]

which can be calculated from the cartesian momentum components of the four momenta used in this experiment by,

$$(3) \qquad \eta = \sinh^{-1}\frac{p_z}{p_T}.$$

The rapidity is given by,

$$(4) \qquad y = \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right),$$

which in our case is equal to the pseudorapidity, as we are measuring photons which have zero rest mass, hence $E = |\mathbf{p}|$.

The azimuthal angle is the angle subtended by the x and y momentum components, so the angle between the momentum components that are perpendicular to the beam axis, in contrast to the pseudorapidity, which is between the particle trajectory and the beam axis, (MAKE SENSE?) (DIAGRAM?)

$$(5) \qquad \phi = \tan^{-1}\frac{p_y}{p_x}.$$

A lorentz invariant quantity measure of angular seperation which is a combination of pseudorapidity and azimuthal angle can also be defined,

$$(6) \qquad (\Delta R)^2 = (\Delta \eta)^2 + (\Delta \phi)^2.$$

The quantity ΔR is Lorentz invariant because both the difference in pseudorapidity $\Delta \eta$ and difference in azimuthal angle $\Delta \phi$ are Lorentz invariant. [6]

2.2. **Branching Fraction.** The branching fraction is the fraction of all particles which decay through a given decay mode, in this case the fraction of $H \to \gamma\gamma$ decays over all Higgs decay modes. For example, if the branching fraction for a given decay mode is 0.5, then on average half of the particles will decay through that mode.

2.3. **Higgs Decay Channels.** Since the Higgs is a very massive particle, it can decay through a large number of different decay modes: (GET SOURCE)

- Fermion-antifermion pair
- Pair of massive gauge bosons
- Pair of massless gauge bosons (through heavy virtual quark or massive gauge boson loop)

[3] [5]
(MAYBE A BIT MORE EXPLANATION ABOUT THESE DECAY MODES?! (GET ANOTHER SOURCE?!)

The decay channel investigated in this experiment is the last one, the decay of the Higgs into

two photons by a heavy virtual quark (top) or massive gauge boson (W boson) loop. This is significantly rarer than the other decay modes as seen in the small branching fraction of 0.0023 (so it occurs twice in every thousand decays) [2], but is used because the momenta of the photons can be measured much more accurately than other more massive particles, since the momenta of the photon is directly proportional to its energy, as the photon has no rest mass. (GET SOURCE) (HOW TO MEASURE PARTICLE MOMENTUM? LOOK AT ATLAS EXPERIMENT)

2.4. **Invariant Mass.** To find the mass of the particle(s) before a collision event, the invariant mass is used, which is the mass which remains unchanged through a Lorentz transform (GET SOURCE). The sum of the invariant masses of the products of an interaction is the same as the sum of the invariant masses of the original particles, so by calculating the invariant masses of the resultant particles of the collision events, we have calculated a series of invariant mass values corresponding to the inputs of the collision event. Since the Higgs invariant mass is known (ROUGHLY?!?), the frequency of values in that window should be higher.

The invariant mass is calculated using the relativistic equation for energy, in natural units this is given by,

$$(7) \qquad\qquad m_0^2 = E^2 - |\mathbf{p}|^2$$

This is calculated in the parsing program by summing the four momenta from the event and then taking the scalar product of the resulting vector with itself. Minkowski space-time has signature $(+, -, -, -)$ so the scalar product has the effect of the above equation. We then take the square root of the scalar product to find the invariant mass. [8]

The mass window is taken to be $115\,\mathrm{GeV}$ to $135\,\mathrm{GeV}$, which is roughly $\pm 10\,\mathrm{GeV}$ the expected Higgs mass, as measured. [1]

3. PYTHIA

The GPL-licensed PYTHIA was used to generate the collision data used in this experiment. [7] The event data is given in a flat text file format. The file consists of a series of events, which start with a header "Event n", the subsequent lines are the four momenta of the products of the collision event, each line in the format "p_x p_y p_z E", where p_z is aligned along the beam axis. Then there is another line "Event $n + 1$" with subsequent lines being four momenta events and so on.

4. Programming

The scripts used in this experiment were written in Python 3, and used the numpy (numeric Python) module for generating the histogram, and PyPlot (from matplotlib) for plotting the histogram and significance plots. Three different scripts are used in this experiment,

4.1. **parse.py.** The script parse.py parses the event data in the format described in section 3 into individual event objects which each contain a list of four momenta objects. In these event classes we have methods which do useful operations like calculating the invariant mass of the event using the four momenta. The four momenta class has methods that allow us to easily calculate the transverse momentum, azimuthal angle, and pseudorapidity, as well as allowing us to add and take the dot product of four momenta. The parsing scripts also contains the filters which are used to filter out Higgs events from background events, and are discussed in more depth in section 6. The invariant masses are calculated for each event in the event data file, and then the resulting list of invariant masses is saved using python's built-in pickle module, which saves python objects in a binary representation, saving space and improving read and write times of the invariant masses.

4.2. **significance.py.** This scripts optimises the filters used on the Higgs and background data by using a series of parameters with the filters and testing them on the entire data set; the significance is taken for the parameters and this is used to find the best filter parameters, as discussed in section 7. This script parses and filters the data using the parsing module.

4.3. **plot.py.** The plotting script generates the histogram of both the higgs and background event data together, called the combined mass histogram. The significance in the mass range

(WHAT MASS RANGE) is calculated by counting the number of Higgs events in the mass range and the number of background events in the mass range and then using eq. 9 to calculate the statistical significance. The script can also generate an invariant mass plot from the combined mass histogram.

4.4. **Testing.** Python's unit testing framework, unittest, was used to run testing scripts on the filters to ensure that they were working as intended. The tests took a pre-generated dataset and then ran this against each filter, checking the result to see if it is what was expected. This allowed us to debug a major issue in the second transverse momenta filter, which was generating large numbers of duplicate invariant masses, causing repeated values in the invariant mass histogram.

5. WEIGHTING

The data given included $10,000$ signal events and $1,000,000$ background events. This is a much smaller ratio of signal to background events than would actually be observed in a real collider experiment, due to the low branching fraction of the $H \rightarrow \gamma\gamma$. The theoretical ratio of the observed Higgs events to observed background events can be calculated by taking the ratio of the Higgs production cross section to the background cross section and multiplying it by the branching fraction of $H \rightarrow \gamma\gamma$,

$$(8) \quad \frac{\text{num. H}}{\text{num. back}} = \frac{\sigma(\text{H prod}) \cdot B_f \left(H \rightarrow \gamma\gamma\right)}{\sigma(\text{back})}$$

where num. H is the number of signal events, num. back is the number of background events, $\sigma(\text{H prod})$ is the Higgs production cross section, $\sigma(\text{back})$ is the background cross section, and $B_f \left(H \rightarrow \gamma\gamma\right)$ is the branching fraction of Higgs to diphoton.

The branching fraction is given by $B_f \left(H \rightarrow \gamma\gamma\right) = 0.0023$, and the cross sections are given by $\sigma(\text{H prod}) = 17.12 \, \text{pb}$, and $\sigma(\text{back}) = 140 \, \text{pb}$. The Higgs cross section and branching fraction are taken at roughly the expected Higgs mass $125 \, \text{GeV}$, [1] where the cross section is taken from the ABPS calculation for $\sqrt{s} = 14 \, \text{TeV}$ where s is the centre-of-momentum energy (the energy in the frame where the total

momenta is zero). [2] [4] (MORE JUSTIFICATION NEEDED). The background cross section is taken from the PYTHIA collision event generator.

6. FILTERS

There are many more background events than there are from the signal (Higgs) events in our data. After the weighting, there are roughly 100 Higgs events to $100,000,000$ background events, so in the combined invariant mass histogram the Higgs is practically invisible. Thus filtering needs to be applied for there to be any chance to be able to observe the Higgs in the invariant mass histogram. This is implemented in the form of event selection in which we attempt to remove the background data while retaining the Higgs data. A few types of filtering are used: number threshold, transverse momenta, energy, angular (azimuthal and pseudorapidity).

6.1. **Number.** The first filter applied a threshold to the number of four momenta in the collision event. It is required that the event must have at least two four momenta, as the $H \rightarrow \gamma\gamma$ will produce at least two photons, not just one. This filter should have a good effect to clean up a large amount of data which is definitely background, as we know that all events with one four momenta must be background.

6.2. **Transverse Momenta.** The transverse momenta is the component of the momentum perpendicular to the beam axis as described in section 2.1. The filter is applied in two passes over both the signal and background events. First the given event is discarded *unless* it has at least one four momentum with a transverse momentum above a given threshold, denoted p_1. Then all four momenta in the event with transverse momenta below a given threshold, denoted p_2, are removed from the event.
(JUSTIFICATION)

6.3. **Energy.** The energy of the four momenta is filtered similarly to how the transverse momenta is filtered. First the event is discarded *unless* it has at least one four momenta with an energy greater than a given threshold, denoted E_1. Then all four momenta with an energy less

than a given threshold, denoted E_2, are removed from the event.

6.4. **Angular.** The square of the difference in azimuthal angle and square of the difference in pseudorapidity are also filtered by finding the maximum difference squared in both the azimuthal and pseudorapidity for all the four momenta in an event. If the maximum difference squared of a given event is below a given threshold, for azimuthal denoted $(\Delta\phi)^2$ and for pseudorapidity denoted $(\Delta\eta)^2$, then that event is discarded.

6.5. **Final Selection.** There are two methods that were proposed to choose the final two transverse momenta to use to calculate the invariant mass. First it was decided to try an azimuthal angle based filter, where the most back-to-back photons (the photons with an azimuthal angle with the greatest difference) would be chosen to calculate the invariant mass. (WHY DIDN'T YOU DO THIS?!?) Instead all but the two highest transverse momenta photons are discarded, and the two remaining are used to calculated the invariant mass.

7. Optimisation of Filters

The filters available can be used to attempt to isolate the Higgs signal from the background signal to increase the Higgs statistical significance in the mass window. Without optimising the filter parameters the best parameters to choose are unknown, and so it is just guessing to choose any. By optimising the filters the best possible values are attained. The filter parameters "goodness" were measured using the statistical significance equation,

$$(9) \qquad \Sigma = \frac{\text{no. signal}}{\sqrt{\text{no. signal} + \text{no. background}}}$$

Each filter is applied to the signal (Higgs) and background events and then the above equation is used to find the statistical significance of the Higgs in those events by counting the number of signal and background events left after filtering. The filter parameters giving a higher statistical significance correspond to having a higher ratio of signal events to background events, so

this gives a better significance of the Higgs in the invariant mass histogram.

To find the best parameters to use for the filters a range of different parameters are tested for each filter. For example for the transverse momenta filter, the threshold is originally applied from $0\,\text{GeV}$ to $700\,\text{GeV}$ in steps of $100\,\text{GeV}$. The value that gives the best statistical significance is then chosen as the current candidate for the best optimised value. Then the value is further optimised by taking the currently best optimised value and then calculating the statistical significance for the values around it, for example checking the statistical significance at $\pm 10\,\text{GeV}$ the current value for the above example. After repeating this process a few times (HOW MANY TIMES?) the most optimum filter values to see the most Higgs event in the invariant mass histogram are found.

INCLUDE SIGNIFICANCE PLOT!!!

After running repeated optimisations over the mass range $0\,\text{GeV}$ to $700\,\text{GeV}$, the optimum cut values were found to be $p_1 = 70\,\text{GeV}$ and $p_2 = 90\,\text{GeV}$.

8. Invariant Mass Histogram

The numpy library is used to generate an combined (signal and background) invariant mass histogram from the invariant masses calculated in the parsing script. The statistical significance of the signal (Higgs) is calculated using eq. 9 over the mass window discussed in section 2.4, and gives a measure of how sure we can be that it is the Higgs in this mass range.
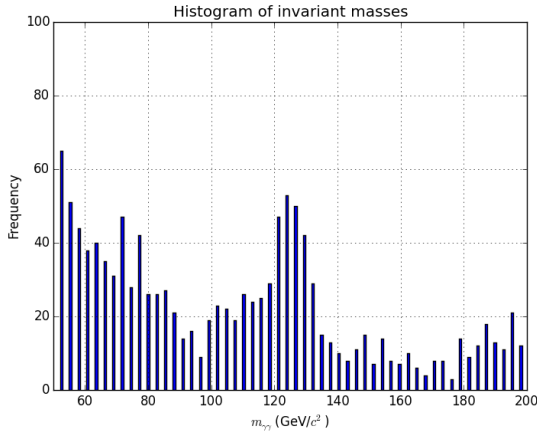
Figure 1. Combined invariant mass plot in the mass range $50\,\mathrm{GeV}$ to $200\,\mathrm{GeV}$ with optimised filter parameters applied. In the mass range the Higgs had a significance of $\Sigma = 0.33295$

(DISCUSS RESULTS!!!)
(TALK ABOUT HISTOGRAM RESOLUTION)

9. Results

10. Issues

There were a few issues encountered upon the way. Some were expected but a couple were not, and were harder to fix.

The clearest and most pressing issue was having to filter, since the Higgs was practically invisible in the unfiltered invariant mass plot (having a statistical significance of only $\Sigma = 0.01264$). Filters were written to attempt to remove the non-Higgs events from the dataset as discussed in section 6. Then the most optimum parameters had to be found that would give us the best ratio of signal to background events, as discussed in section 7.

When the invariant mass histogram was generated after finding some preliminary optimised parameters, invariant masses were being repeated. This was revealed by testing as discussed in section 4.4, and was due to a missing break statement. These events were being added multiple times in the second transverse momenta and second energy filters (section 6). The filter would iterate over each four momenta in an event, and if the transverse momentum/energy of the four momentum was greater than the threshold, that event would be added to the list of events to return from the filter. The missing break statement meant that even if a transverse momentum/energy above the threshold was found, the filter would continue to iterate through the event, repeatedly adding the same event to the output list. This was noticed when looking at the invariant mass histogram, there would be some bins that were much taller than adjacent ones, instead of a roughly homogenous distribution. This was due to those bins containing events with *many* four momenta over the transverse momenta/energy threshold, so those events would be added to the invariant mass histogram many more times. This was all fixed by adding a single break statement after a four momentum above the threshold was found, preventing the repeated adding.

Another issue, while not as urgent, was that the parsing script took a very long time to parse all the data. This was due to the entire remainder of the data file being passed to the parsing function every time an event was found, as the file would parse each line as a four momentum until it found another "Event n" heading. To fix this a file was created alongside the event data files containing the number of four momenta in each event. This meant that only the number of lines needed were passed to the event parser, instead of all the lines being passed, so instead of having to pass almost a million lines, only around ten lines would need to be passed for each event, which was significantly faster. Without this the parsing script would take many hours to run. Parallel processing was also used for a small performance gain. (EXPLAIN?)

11. Acknowledgments

12. Conclusion

References

[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. *arXiv*, 2012.

[2] S. Dittmaier et al. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. *arXiv*, 2011.

[3] Thomas R. Junk. Review of physics results from the tevatron: Higgs boson physics. *arXiv*, 2015.

[4] Luc Blanchet, Bala R. Iyer. Third post-Newtonian dynamics of compact binaries: equations of motion in the centre-of-mass frame. *IOP*, 2002.

[5] K. Nielsen. Higgs boson decay into two photons in an electromagnetic background field. *Phys. Rev. D*, 2014.

[6] Ross Diener, Stephen Godfrey, Travis A.W. Martin. Using final state pseudorapidities to improve s-channel resonance observables at the LHC. *Phys. Rev. D*, 2009.

[7] Torbjörn Sjöstrand; Stefan Ask; Jesper R. Christiansen; Richard Corke; Nishita Desai; Philip Ilten; Stephen Mrenna; Stefan Prestel; Christine O. Rasmussen; Peter Z. Skands. An introduction to pythia 8.2. `http://home.thep.lu.se/~torbjorn/Pythia.html`.

[8] Cheuk-Yin Wong. *Introduction to High-Energy Heavy-Ion Collisions*. World Scientific, 1994.