संकुल नवप्रवर्तन केंद्र, दिल्ली विश्वविद्यालय

CLUSTER INNOVATION CENTRE, UNIVERSITY OF DELHI

STATISTICAL METHODS TO DETECT HIGGS FROM PARTICLE ACCELERATOR DATA USING DIGITAL EVENT PROCESSING

Project Report

V.2 Digital Event Processing

Mentor

Dr. Kirti Ranjan

Centre for Detector and Related Software Technology

(In collaboration with CMS, CERN)

Department of Physics & Astrophysics

University of Delhi

Submitted By

Sahil Nakul Mathur

Rahul Yadav

Vasundhara

ACKNOWLEDGEMENT

We would like to extend our heartfelt gratitude to our mentor Dr. Kirti Ranjan. Without his unconditional guidance and support this project wouldn't have been possible. We thank him to spare his valuable time to attend to our petty doubts. His notes and documentation enhanced our clarity over the subject. We also acknowledge the help of our course teachers Mr. Abhijeet Parmar, Ms. Apurva Agarwal and Mr. Amit Gautam. They have been an incredible support and guiding force throughout the execution of the project. We appreciate the tremendous help given by our program co-ordinator Dr. Bibhu Biswal and Academic Convenor Dr. Shobha Bagai and Dr. Pankaj Tyagi. Last but not the least we would also like to thank our friends and family for their unconditional support and love.

-Authors-

CONTENTS

INTRODUCTION

Particle physics is the branch of physics which studies the fundamental or elementary particles and the interactions which hold them together. In the last century, the interplay of the theory and collider based experiments have already provided a great insight into the fundamental structure of matter. The Standard Model of particle physics evolved as a very successful theory for describing the behaviour of smallest particles and their interactions. However, there are plenty of open questions in the Standard Model and hence information from new collider experiments, like the Large Hadron Collider (LHC) at European Organization for Nuclear Research (CERN), Switzerland may provide key information related with the search for new particles and theories beyond the Standard Model. LHC has a general purpose detector called the CMS experiment which looks for these new particles.

Last year LHC came up with a discovery of the Higgs like boson, the so called God's particle, which may solve the mystery of mass generation at the smallest scales. Since the signature of new particles can also be mimicked by the other known particles so it is important to separate the Event coming from the new particle from the Background or Background created by other particles. Accordingly, the aim of the present project is to understand and implement statistical methods, which can help separating Event from Background in the particle physics experiment.

ELEMENTARY PARTICLES

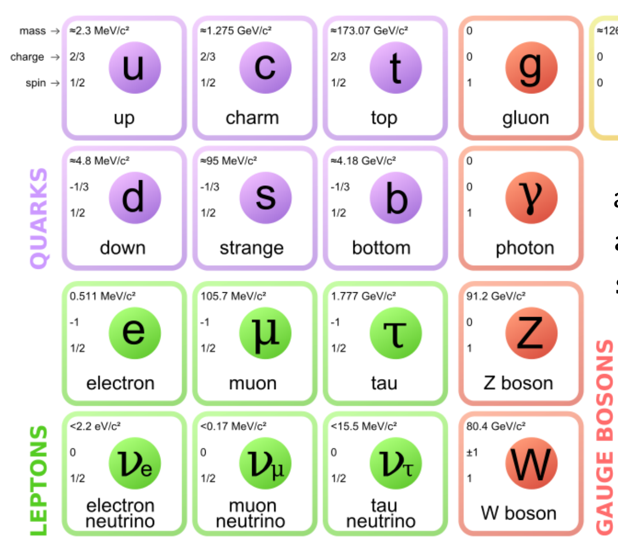

Figure 1: An Introduction to Elementary Particles

Elementary-particle physics deals with the fundamental constituents of matter and their interactions. In the past several decades an enormous amount of experimental information has been accumulated, and many patterns and systematic features have been observed. Highly successful mathematical theories of the electromagnetic, weak, and strong interactions have been devised and tested. These theories, which are collectively known as the standard model, are almost certainly the correct description of Nature, to first approximation, down to a distance scale 1/1000th the size of the atomic nucleus. The Standard Model of particle physics contains 12 flavors of elementary fermions, plus their corresponding antiparticles, as well as elementary bosons that mediate the forces and the Higgs boson, which was detected by the two main experiments at the LHC (ATLAS and CMS). However, the Standard Model is widely considered to be a provisional theory rather than a truly fundamental one, since it is not known if it is compatible with Einstein's general relativity.

HIGGS PARTICLE

The Higgs boson or Higgs particle is an elementary particle initially theorised in 1964, and tentatively confirmed to exist on 14 March 2013. It appears to confirm the existence of the Higgs field which is pivotal to the Standard Model and other theories within particle physics. It would explain why some fundamental particles have mass when the symmetries controlling their interactions should require them to be massless, and—linked to this—why the weak force has a much shorter range than the electromagnetic force. The discovery of a Higgs boson should allow physicists to finally validate the last untested area of the Standard Model's approach to fundamental particles and forces, guide other theories and discoveries in particle physics, and potentially lead to developments in "new" physics.

The existence of the Higgs field – the crucial question could be proven by searching for a matching particle associated with it, which would also have to exist—the "Higgs boson". Detecting Higgs bosons would automatically prove that the Higgs field exists, which show the Standard Model is essentially correct. There no way to discover whether Higgs bosons actually existed in nature either, because they would be very difficult to produce, and would break apart in about a ten-sextillionth (10−22) of a second. Although the theory gave "remarkably" correct answers, particle colliders, detectors, and computers capable of looking for Higgs bosons took over 30 years (c. 1980 – 2011) to develop.

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

At CERN, the European Organization for Nuclear Research, physicists and engineers are probing the fundamental structure of the universe. They use the world's largest and most complex scientific instruments to study the basic constituents of matter - the fundamental particles. The particles are made to collide together at close to the speed of light. The process gives the physicists clues about how the particles interact, and provides insights into the fundamental laws of nature.

The instruments used at CERN are purpose-built particle accelerators and detectors. Accelerators boost beams of particles to high energies before the beams are made to collide with each other or with stationary targets. Detectors observe and record the results of these collisions.

Figure 2: CERN laboratory

Founded in 1954, the CERN laboratory sits astride the Franco-Swiss border near Geneva. It was one of Europe's first joint ventures and now has 20 member states. It comprises of a 26.7 km tunnel having more than 9300 Superconducting Magnets, 1232 Dipoles (15m), 448 Main Quads, 6618 Correctors. Its Operating temperature is about 1.90 K. Scientists at CERN are conducting experiments to answer some pertinent question related to Matter and Universe.

Some of them includes:

- What is the origin of mass
- What could be the dark matter that keeps together the clusters of galaxies
- Why the main interactions are so different in strength
- Why gravity is not included so far in our picture
- How many are really the dimensions of our world

The answer to some of these questions is probably hidden in the so far unexplored TeV region. The bare Standard Model could be consistent with massless particles but matter particles range from almost 0 to about 170 GeV while force particles range from 0 to about 90 GeV.

HIGGS PARTICLE

The Higgs boson or Higgs particle is an elementary particle initially theorised in 1964, and tentatively confirmed to exist on 14 March 2013. It appears to confirm the existence of the Higgs field which is pivotal to the Standard Model and other theories within particle physics. It would explain why some fundamental particles have mass when the symmetries controlling their interactions should require them to be massless, and linked to this why the weak force has a much shorter range than the electromagnetic force. The discovery of a Higgs boson should allow physicists to finally validate the last untested area of the

Standard Model's approach to fundamental particles and forces, guide other theories and discoveries in particle physics, and potentially lead to developments in "new" physics.

The existence of the Higgs field – the crucial question could be proven by searching for a matching particle associated with it, which would also have to exist—the "Higgs boson". Detecting Higgs bosons would automatically prove that the Higgs field exists, which show the Standard Model is essentially correct. There no way to discover whether Higgs bosons actually existed in nature either, because they would be very difficult to produce, and would break apart in about a ten-sextillionth (10−22) of a second. Although the theory gave "remarkably" correct answers, particle colliders, detectors, and computers capable of looking for Higgs bosons took over 30 years (c. 1980 – 2011) to develop.

EXPERIMENT

On October 8, 2013 the Nobel Prize in Physics was awarded jointly to François Englert and Peter Higgs "for the theoretical **discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles**, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider". The Higgs boson is believed to give other particle their mass. The astonishing world of particle physics and application of Digital Signal Processing in it inspired us to work on this project. The Pseudo-LHC data is acted upon by suitable filters which efficiently segregate background and data. The entire project was inter disciplinary and is a beautiful amalgamation of Statistics, DSP and Particle Physics.

The LHC accelerates the protons to energies as high as 3.5 TeV. When 2 such protons beams travelling in opposite directions collides, a huge number of elementary particles are observed. The Higgs has a mass 126 times to that of a proton. The proton is actually made up of two up and one down quark so the entire mass of proton is subdivided into 3 parts and hence the probability of generating a particle of higher mass is very low. It has been observed that one Higgs particle is found in one of the trillion proton- proton collisions. The Higgs has very short half-life ($\sim 10^{-23}$seconds) so the sensors of course cannot detect its presence directly. The Higgs decays into two photons of equal energy. Therefore, from the data generated in LHC, the researchers try to trace back the energy of the photons and see to what TeV region does it belong to.

SIMULATING EVENTS

Simulating various possibilities of events that occur at the collider requires a pseudo event or date. Pseudo data must also obey laws and constraints that physics must impose to generate a legitimate simulation of events. We know that the LHC's data was a superposition of the cumulative distribution curves of exponential and Gaussian type. Some events coming from collision of protons produce a Gaussian probability distribution which is an indicator of the Higgs while other events which we call background are coming from other interactions. These events produce an exponential decay probability distribution. Hence to detect the presence of Higgs, the project considers the exponential curve represents the background and the Gaussian curve represents data. The number of data points in the background curve is much more than the numbers of points in the event curve. This is because the probability of occurrence of the Higgs is far less than probability of occurrence of other events. The technique to develop such pseudo events and creating a virtual collider scenario, the use of Monte Carlo as a tool is very important.

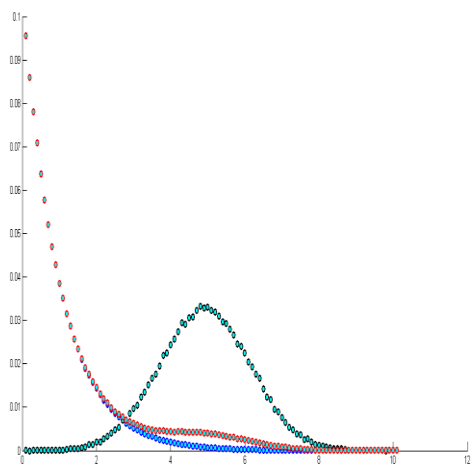

Figure 3: Event, background and pseudo data.

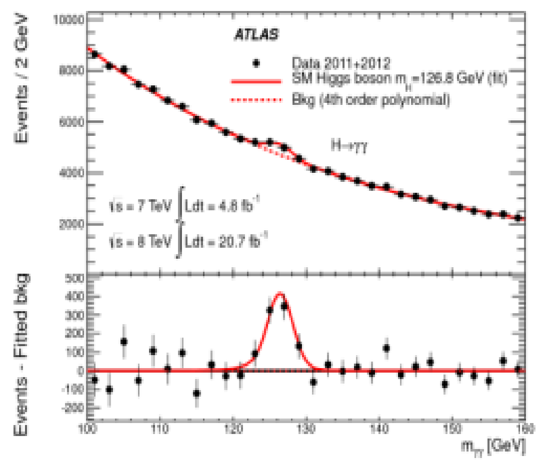

Figure 4: The event and the LHC data
Source: Scientific background on the Nobel Prize in Physics 2013

MONTE CARLO

Monte Carlo methods are used as a fundamental tool in computational statistics. At the heart of the Monte Carlo simulations is random number generation. The Monte Carlo method generates suitable random numbers and observes that fraction of the numbers obeying some property or properties. The name Monte Carlo comes from the city on the Mediterranean with its famous casino and a Monte Carlo calculation implies a statistical method of studying problems based on the use on random numbers, similar to those generated in the casino games of chance.

Our first introduction to a Monte Carlo technique was using the Ising Model. It is a mathematical model of ferromagnetism. This defines discrete atoms in a lattice and the magnetic dipole moments of atomic spins which are taken to be either 1 or -1. The interaction of the neighbouring atoms is considered on a particular atom after the spin is randomly changed.

Monte Carlo methods require a reliable set of random numbers but truly random numbers are hard to obtain. Most Monte Carlo techniques use pseudo random numbers. Our entire effort to generate pseudo data was based on generating random numbers whose probability distribution functions are known. The techniques were inverse transform method and Box- Muller transformation for exponential and normal curves respectively. These algorithms of generating such curves are a very elementary step towards Monte Carlo generation. This step just teaches us the way of generation and use of random numbers and probability theory which are so important in Monte Carlo techniques.

INVERSE TRANSFORM METHOD

Sampling of random numbers from distributions is done by applying a transformation to non – uniform variates. It is a basic method for pseudo-random number sampling, i.e. for generating sample numbers at random from any probability distribution given its cumulative distribution function (cdf). This is basically mapping a number in a particular domain to a probability between 0 and 1 and then inverting that function. Therefore the problem is that there is a random variable X whose distribution can be described by the cumulative distribution function P and we want to generate values of X which are generated according to this distribution.

If X is a scalar random variable with a continuous cumulative distribution function P_X , then the random variable

$$U = P_X(X) \tag{1}$$

has a U (0,1) distribution.

Using this basic definition the inverse can also be stated. We can derive a simple relationship between uniform random variable U and a random variable X with distribution function P

$$X = P_X^{-1}(U). \tag{2}$$

To implement the inverse transform technique, a random number U is generated with a uniform distribution in the interval [0,1], compute the value of X such that P(X) = U and then take X to be the random number drawn from the distribution described by X.

In the same way the distribution can be applied to discrete variables.

X is the discrete random variables with probabilities

$$P(X = x_i) = p_{i,} \quad i = 0,1 \ldots \quad \sum_{i=1}^{\infty} p_i = 1. \tag{3}$$

We have to generate a realization X and hence first U is generated from U(0,1) and then set X = x_i if

$$\sum_{j=0}^{i-1} p_j \leq U < \sum_{j=0}^{i} p_j \quad (4)$$

The implementation of the same in the code is as follows.

$$R(i) = -\log(1 - e^{-rand(1)}) \quad (5)$$

BOX – MULLER TRANSFORMATION

A transformation which transforms from a 2 dimensional continuous uniform distribution to a 2 dimensional bivariate normal distribution (or complex – normal distribution). If x_1 and x_2 are uniformly and independently distributed between 0 and 1, then z_1 and z_2 as defined below have a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

$$z_1 = \sqrt{-2 \ln x_1} \cos 2\pi x_2 \quad (6)$$

$$z_2 = \sqrt{-2 \ln x_1} \sin 2\pi x_2 \quad (7)$$

IMPLEMENTATION

The application to the above two methods are implemented on the algorithmic front to obtain probability distribution using non uniform random number generation.

GENERATING DISCRETE RANDOM NUMBERS

The discrete random numbers were generated by using Inverse Transformation for the exponential numbers and the Box – Muller method for the exponential numbers. The numbers are generated in such a way that the frequency distribution of the numbers is exponential and Gaussian respectively.

The inverse transform method gives a value to the random variable by using random variable. In our case we have made use of the MATLAB function `rand()` to generate a random number between 0 and 1. The Box – Muller method makes use of the same function and gives us a pair of numbers. In both the cases the function is acted upon by a particular set of rules and gives us the result of the corresponding transformation.

Physically, the exponential curve which is the frequency distribution of the background represents the output from the background processes. The normal curve is the frequency distribution of the event and represents the 2 output photons of which are left after the unstable Higgs breaks down into these. The X – axis of the graphs represents the probability and the Y – axis represents the energy.

EXPONENTIAL PROBABILITY DISTRIBUTION

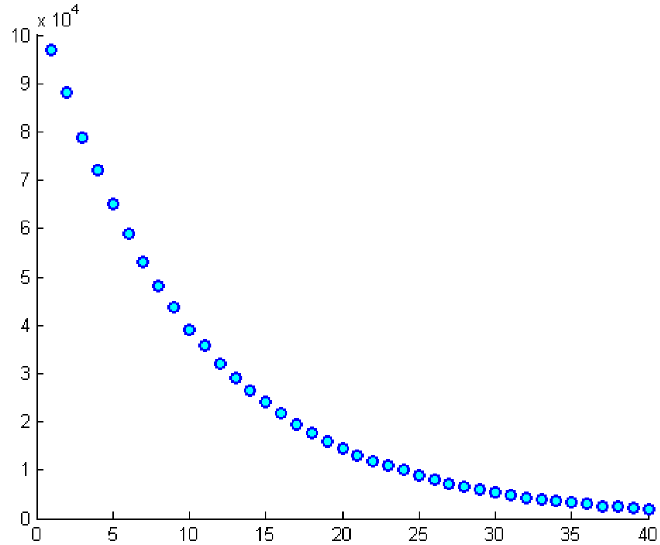

Figure 4: Exponential Probability Distribution

```
% To generate exponential random
numbers
n= 10^6;
R=zeros(n,1); % Matrix for Background
(Exponential Distribution)
for i=1:n
    R(i)= -log(1-exp(-rand(1)));
%Inverse Transform Method
end
```

GAUSSIAN PROBABILITY DISTRIBUTION

```
% To generate normal random
numbers
R2=zeros(n,1); % Matrix for Event
(Normal Distribution)
for i=1: n/2
    flag=0;
    while flag==0
        u1=2*rand(1)-1;
        u2=2*rand(1)-1;
        rsq= u1^2 +u2^2;
        if rsq<1
            flag=1;
            fac=sqrt(-
2*log(rsq)/rsq);
            R2(i)= (u1*fac) ;
%Box Muller Transformation (First
Random Number)
            R2(n2-i)= (u2*fac) ;
%Second Random Number
        end
    end
end
```

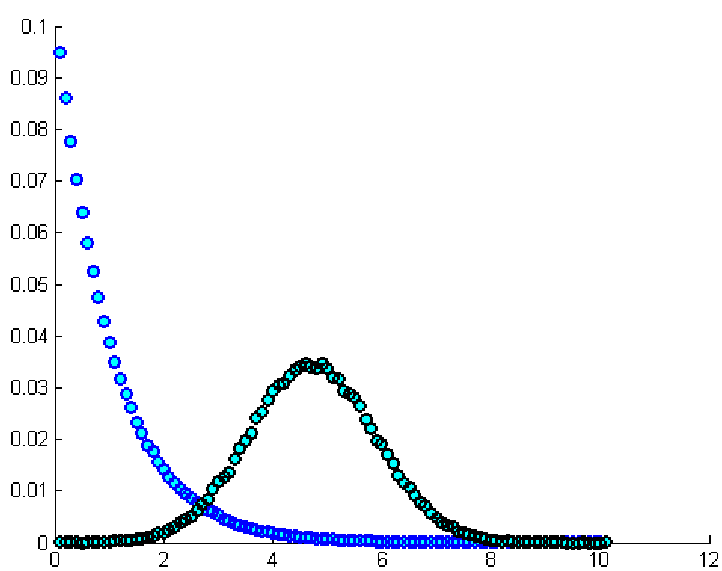

Figure 5: Gaussian and Exponential probability Distribution

MIXING OF EVENT AND BACKGROUND

Mixing of Events to obtain real like data must be done in the same ratio as the probability of the occurrences of Higgs and the background. Hence the following plot is obtained. The pseudo data is obtained by superposition of the 2 graphs. As already explained the exponential graph and Gaussian distribution give us the background and the event respectively. The 2 graphs when combined give us the pseudo data. The pseudo data is the data generated by the random numbers similar to LHC data and is used for computation. To look closely at the nature of the graph of the pseudo data the parameters (λ, μ, σ, k) of the exponential and the Gaussian distributions are varied and the nature of the final curve is observed for different parameters.

```
%% Mixing the random numbers :: Generating Pseudo Data.
R4= [R;R2]; % Since the numbers are Pseudo Random, just adding the matrices
will create Pseudo Data.
```

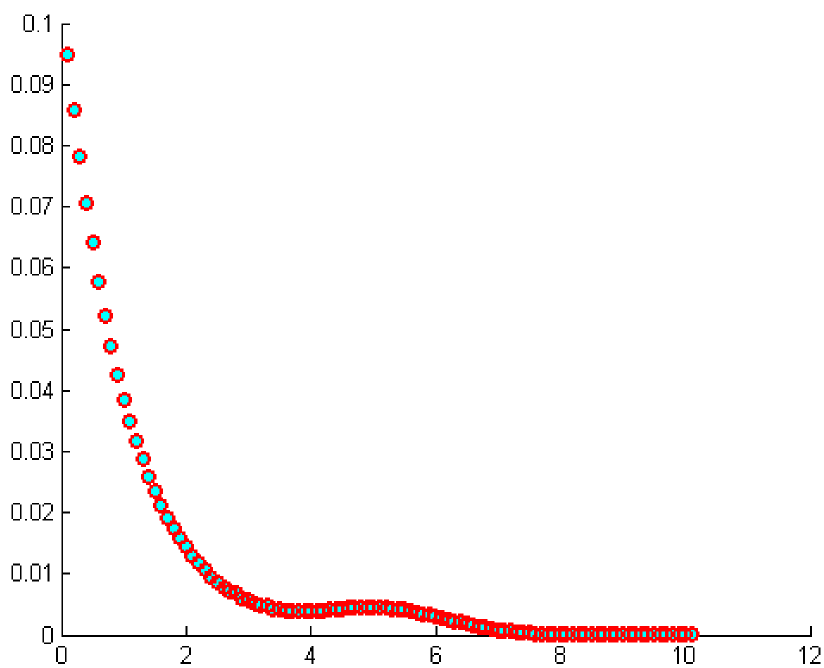

Figure 6: Pseudo Data

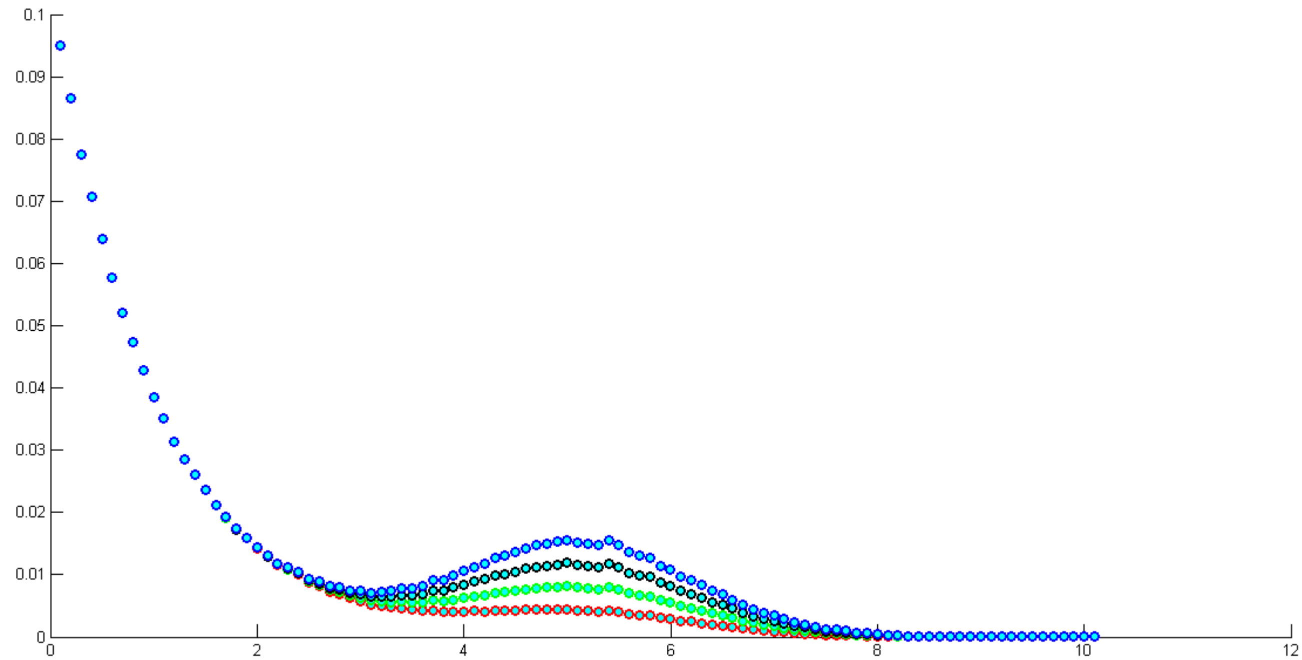

Figure 7: Pseudo Data with varying λ

SEPARATING DATA

The data obtained should now be looked at independently without any knowledge of what the background and event were. The type of the distribution is known but the parameters are unknown. Knowing nothing at all about how the pseudo data was obtained the task is to separate out the event and the background. In very vague terms it can be commented that the exponential curve can be fitted using least squared fit and for the Gaussian distribution curve fitting is to be used. The Bayesian and the frequentist approach can be used to detect the Gaussian curve from the pseudo data.

A very preliminary approach was implemented to separate the background and event. The event i.e. the Gaussian curve was assumed to have $\mu = 0$ and $\sigma = 1$. The Gaussian curve was hence symmetric about the X – axis and hence the part of the event where the values of X are negative is as it is used to represent the actual event. For obtaining the right hand side of the Gaussian curve the ratio of probability of event to background (λ) is used and the points corresponding to Gaussian curve are obtained from pseudo data. The obtained curve is fitted using `polyfit()` function of MTALAB. This is a very rustic approach to do things and there are a lot of demerits to this approach. However, this is a very simple way to explain things.

RESULT AND ANALYSIS

The project had a theoretical basis and the result was based on observation and prior knowledge. Initially the data was generated which was a replica of the data obtained as the result of experiments at LHC. The data was a result of the superposition of the two events which were mixed in the same ratio as the probability of their occurrence. The other important result is extracting two events from the pseudo data and relating to the physical implications by detecting the Higgs which shows peculiar probability distribution as compared to other background. The Gaussian curve of the data obtained from the LHC data as obtained in figure 4 is somewhat similar to the curve obtained from the `polyfit()` function in MATLAB the extracted curve is a very simple technique applied to roughly understand how to go about things. The actual implementation is yet to be done.

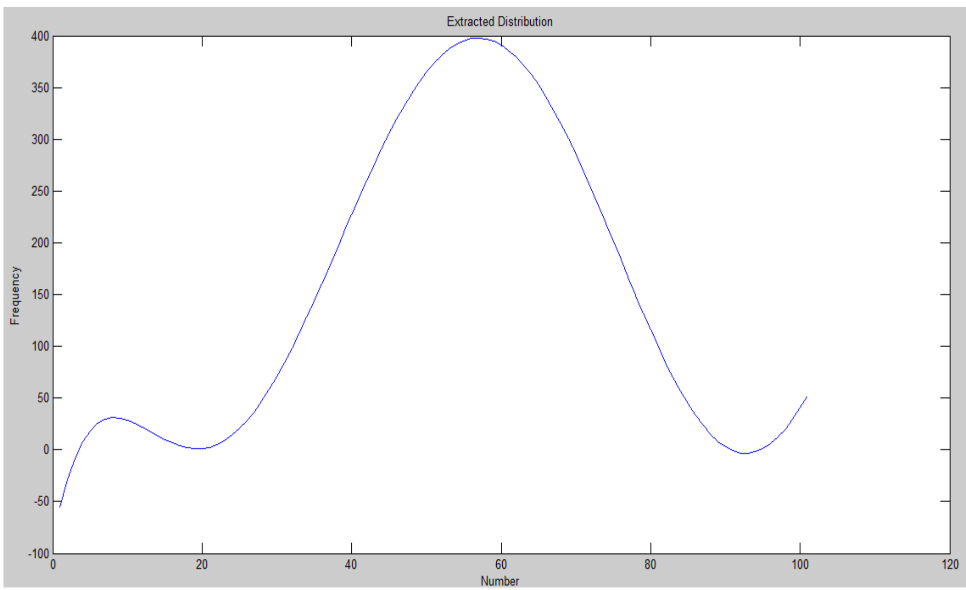

Figure 8: Extracted Gaussian distribution

FURTHER PROPOSALS

The implementation of extracting the event from the data is still not achieved to a confident level. The implementation has still not taken a concrete shape. The curve fitting and least squared fit techniques have to be studied and the implementation has to be carried out. Also the deeper insights of particle physics have to be understood so as to be in a position to understand the physical implications what the project work carried out. The beauty of the experiments can only be understood after going deep into the related particle physics.

APPENDIX

```
%% HOW THEY FOUND HIGGS?
% Project Mentor - Dr. Kirti Ranjan, University of Delhi
% Sahil Nakul Mathur, Vasundhara Mehta, Rahul Yadav

clear
clc
n =10^6; %number of random numbers of background or noise to be generated
n2=10^4; %number of random numbers of event or signal to be generated
nbin= 100; %number of bins

% To generate exponential random numbers
R=zeros(n,1); % Matrix for Background (Exponential Distribution)
for i=1:n
    R(i)= -log(1-exp(-rand(1)));      %Inverse Transform Method
end
% To generate normal random numbers
R2=zeros(n2,1); % Matrix for Event (Normal Distribution)
for i=1: n2/2
    flag=0;
    while flag==0
        u1=2*rand(1)-1;
        u2=2*rand(1)-1;
        rsq= u1^2 +u2^2;
        if rsq<1
            flag=1;
            fac=sqrt(-2*log(rsq)/rsq);
            R2(i)= (u1*fac) ;    %Box Muller Transformation (First Random
Number)
            R2(n2-i)= (u2*fac) ;  %Second Random Number
        end
    end
end

%% Mixing the random numbers :: Generating Pseudo Data.
R4= [R;R2]; % Since the numbers are Pseudo Random, just adding the matrices
will create Pseudo Data.
[l,m]= size(R4);


%% To see the pattern made by the distributions
R3=zeros(n,1);
for i=1:n-1
    R3(i)=R(i+1);
end
R3(n)=R(1);
R5=zeros(n2,1);
for i=1:n2-1
    R5(i)=R2(i+1);
end
R5(n2)=R2(1);
figure,subplot(2,1,1),plot(R3,R,'+');
       subplot(2,1,2),plot(R5,R2,'*');
```

```
%%  Plots of the generated distributions
% Exponential Distribution
% Finding the maximum and minimum random number.
ymin2= 10^7;  %to calculate minimum
ymax2= -10^8; %to calculate maximum
for i= 1:n
    a= R(i);
    if a<ymin2
        ymin2=a; %minimum random number
    end
    if a>ymax2
        ymax2=a; %maximum random number
    end
end

binsize2 = (ymax2-ymin2)/nbin;
pn2= zeros(nbin+1,1); %matrix that will store the frequency of random number

for i=1:n
    t2=((R4(i)-ymin2)/binsize2)-mod(((R(i)-ymin2)/binsize2),1); %counting the
frequency of random numbers
    pn2(t2+1)=pn2(t2+1)+1;
end

xaxis2=zeros(nbin+1,1);
for i=1:nbin+1
    xaxis2(i)=ymin2+(i*binsize2);
end
% Normal Distribution
% Finding the maximum and minimum random number.
ymin3= 10^7;  %to calculate minimum
ymax3= -10^8; %to calculate maximum
for i= 1:n2
    a= R2(i);
    if a<ymin3
        ymin3=a; %minimum random number
    end
    if a>ymax3
        ymax3=a; %maximum random number
    end
end

binsize3 = (ymax3-ymin3)/nbin;
pn3= zeros(nbin+1,1); %matrix that will store the frequency of random number

for i=1:n2
    t2=((R2(i)-ymin3)/binsize3)-mod(((R2(i)-ymin3)/binsize3),1); %counting
the frequency of random numbers
    pn3(t2+1)=pn3(t2+1)+1;
end

xaxis3=zeros(nbin+1,1);
for i=1:nbin+1
    xaxis3(i)=ymin3+(i*binsize3);
end
```

```
figure,subplot(2,1,2), plot(xaxis3,pn3);
title('Events(Guassian Distribution)');
xlabel('Number');
ylabel('Frequency');
subplot(2,1,1), plot(xaxis2,pn2);
title('Background(Exponential Distribution)');
xlabel('Number');
ylabel('Frequency');

%% Extracting the signal from noise (Normal Distribution from Exponential)
% Step 1: Since exponential function is not giving any negative values
implies  those values are of event.
% Step 2: From the remaining positive matrix, ignore values corresponding to
R4[i]== - log(1-(exp(-rand(1))).
% Step 3: Use curve fitting techniques on values obtained from Step 1 and
Step 2.
% Mixed Data
% Finding the maximum and minimum random number.
ymin1= 10^7;  %to calculate minimum
ymax1= -10^8; %to calculate maximum
for i= 1:l
    a= R4(i);
    if a<ymin1
        ymin1=a; %minimum random number
    end
    if a>ymax1
        ymax1=a; %maximum random number
    end
end

binsize1 = (ymax1-ymin1)/nbin;
pn1= zeros(nbin+1,1); %matrix that will store the frequency of random number

for i=1:l
    t2=((R4(i)-ymin1)/binsize1)-mod(((R4(i)-ymin1)/binsize1),1); %counting
the frequency of random numbers
    pn1(t2+1)=pn1(t2+1)+1;
end

xaxis1=zeros(nbin+1,1);
for i=1:nbin+1
    xaxis1(i)=ymin1+(i*binsize1);
end

figure,plot(xaxis1,pn1);
title('Mixed Distribution');
xlabel('Number');
ylabel('Frequency');

%% Extracted Data
Final_Matrix=[];
for i=1:l
   if R4(i)<=0
       Final_Matrix =[Final_Matrix; R4(i)];
```

```
    elseif R4(i)>0 && R4(i)<4
        prob=rand(1);
        if prob<0.01
        Final_Matrix =[Final_Matrix; R4(i)];
        end
    end
end
[len_final,klm]=size(Final_Matrix); %length of final_matrix

%% Finding the maximum and minimum random number.
ymin= 10^7;  %to calculate minimum
ymax= -10^8; %to calculate maximum
for i= 1:len_final
    a= Final_Matrix(i);
    if a<ymin
        ymin=a; %minimum random number
    end
    if a>ymax
        ymax=a; %maximum random number
    end
end
%% Plots
binsize = (ymax-ymin)/nbin;
pn= zeros(nbin+1,1); %matrix that will store the frequency of random number

for i=1:len_final
    t2=((Final_Matrix(i)-ymin)/binsize)-mod(((Final_Matrix(i)-
ymin)/binsize),1); %counting the frequency of random numbers
    pn(t2+1)=pn(t2+1)+1;
end

xaxis=zeros(nbin+1,1);
for i=1:nbin+1
    xaxis(i)=ymin+(i*binsize);
end

%figure,plot(xaxis,pn,'*')
p_fit = polyfit(xaxis,pn,6);
figure,plot(p_fit(1).*(xaxis.^6) + p_fit(2).*(xaxis.^5)+
p_fit(3).*(xaxis.^4)+ p_fit(4).*(xaxis.^3)+  p_fit(5).*(xaxis.^2)+
p_fit(6).*xaxis + p_fit(7))
title('Extracted Distribution');
xlabel('Number');
ylabel('Frequency');
```