EP219: Data Analysis and Interpretation

Report: Assignment 4
Team Darth Analysis

OCtober 29 2018 to November 4, 2018

Contents

1 Problem Statement

Consider a dark matter direct detection exper- iment that is designed to measure the recoil energy of nuclei being scattered by dark matter particles. The measured recoil energies (ER) range from 0 40 KeV and the total number of events are reported in 1 KeV bins. The published data is attached in the text file "recoilenergydata EP219.csv" showing the number of events as a function of recoil energy (For the first bin 0.5 KeV is the central value of the bin, so the first bin corresponds to recoil energies between 0-1 KeV). We want to analyze this data to look for a dark matter signal! Unfortunately, there are a large number of background processes that could also contribute to dark matter scattering.

The dark matter signal spectrum has the following triangular form as a function of the recoil energy of the nucleus (ER).

$$\frac{dN}{dEr} = \sigma * 20 * (E_R - 5KeV) \; for \; 5KeV < E_R \; < \; 15KeV$$

$$= \sigma * 20 * (-E_R + 25KeV) \; for \; 15KeV < E_R \; < \; 25KeV$$

$$= 0 \; otherwise$$

Here the signal strength depends on a single parameter σ which is the dark matter- nucleus scattering cross-section measured in femto-barns (fb) (1 fb = 10^{-39} cm^2).

The background rate is exponentially falling with energy and has the form,

$$\frac{dN}{dEr} = 1000 * exp(-\frac{E_R}{10KeV})$$

1. Make a clearly labelled histogram of the data.

2. Assuming background only processes, calculate the mean number of events that you would expect to see in each bin. Make a histogram of this expected background.

3. Assuming cross-sections of 0.01 fb, 0.1 fb, 1 fb, 10 fb, 100 fb, calculate the mean number of events that you would expect to see in each bin assuming background and signal. Make histograms for each of these cases. In which cases do you expect to tell by eye whether or not you have a dark matter signal?

4. Find the log likelihood function of the cross-section log L() and plot it. De- scribe in detail the process used to arrive at this log likelihood function.

5. Use this log likelihood function to find the maximum likelihood estimate (MLE) of the cross-section. Also report a 1-σ interval of cross-sections that are consistent with the data.

2 Python Code

Here's our python code for to extract data into an numpy array and then changing the values of the columns of the array to plot the required histograms.

```
# -*- coding: utf-8 -*-
"""
Created on Fri Nov  2 18:37:48 2018

@author: Raunak
"""
import numpy as np
import pandas as pd
from scipy.integrate import quad
import math
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.interpolate import spline

def darkmatterfunction(x):
    if  x<15 or x>25:
        return 0
    else:
        return sigma*20*(25-x)

def backgroundrate(x):
    return 1000*(math.exp(-x/10))

data=1
```

```python
backgroundsignalcalc=[0]*40
darkmattersignalcalc=[[0 for j in range(40)] for i in range(5)]



data = np.genfromtxt("recoilenergydata_EP219.csv", delimiter=',', skip_header = 1, usecols = (0,1))
dt=np.transpose(data)


err=0
for i in range(40):
    backgroundsignalcalc[i],err=quad(backgroundrate, data[i][0]-.5,data[i][0]+.5)

    for j in range(5):
        sigma=10**(j-2)
        darkmattersignalcalc[j][i],err=quad(darkmatterfunction,data[i][0]-.5,data[i][0]+.5)
        darkmattersignalcalc[j][i]+=backgroundsignalcalc[i]
        j+=1

plt.bar(dt[0],dt[1])
plt.title('Measured signal')
plt.xlabel('Energy in kEV')
plt.ylabel('Number of events')
plt.show()


plt.bar(dt[0],backgroundsignalcalc)
plt.title('Calculated Background signal')
plt.xlabel('Recoil energy in kEV')
plt.ylabel('Number of events')
plt.savefig('BackgroundSignal.png')
plt.show()

```

```

for j in range(5):
    sigma=10**(j-2)
    print("Total signal for sigma=", sigma)
    plt.bar(dt[0],darkmattersignalcalc[j])
    plt.title('Calculated signal')
    plt.xlabel('Energy in kEV')
    plt.ylabel('Number of events')
    plt.savefig('Signal for Sigma.png')
    plt.show()







#
#We call the array containing predicted values of N for sigma=0.01 as N_1 (which has 40 elements) and so on
#N is the array obtained from csv file for no. of events (Not really)

j=1
N=np.transpose(darkmattersignalcalc)
L=[0]*5
m=M=[[0.0]*5]*40
L[j] = M[0][j]

for j in range(0,5):
    for i in range(0,40):
        m[i][j]=(N[i][j] + abs(N[i][j]- data[i][1]))
        M[i][j]=data[i][1]*math.log(m[i][j])-math.log(data[i][1])-m[i][j]    #Calculated in log itself since product is too big
        L[j] = L[j]+M[i][j]        #LOG Likelihood

```

```python
s=[-2,-1,0,1,2]
plt.plot(s,L)
plt.title('Maximum likelihood Estimate')
plt.xlabel('Log Sigma')
plt.ylabel('Log likelihood')
plt.savefig('LogLikelihood.png')
plt.show()

s1=np.array([-2,-1,0,1,2])
xnew=np.linspace(s1.min(),s1.max(),300) #interpolating the given data into a smooth curve to find the 1-sigma interval
L_new=spline(s,L,xnew)
q=0
j=0
for i in range(len(L_new)):
    j=abs(L_new[i]-(L[0]/math.sqrt(math.e)))  #since even after interpolation, the values of the likelihood don't exactly drop by sqrt(e), we found the only value present in the interpolated set which has an error of 54 points of L_max/sqrt(e).
    if j<54.75:
        q=xnew[i]
        print(q)#the value of the order of sigma(parameter) for the 1-sigma interval
        print("The Sigma interval Value")
    else:
        continue

plt.plot(xnew,L_new)
z=plt.axvline(q)
g=plt.axvline(-2)
plt.xlabel("Log sigma")
plt.ylabel("Value of Log Likelihood")
plt.show()
```

Finally we used the interpolate function from Scipy and interpolated the data points to fit a smooth curve through them for the sigma interval.

3 Histograms

1. **Question a**:- Histogram Plot for the measured data

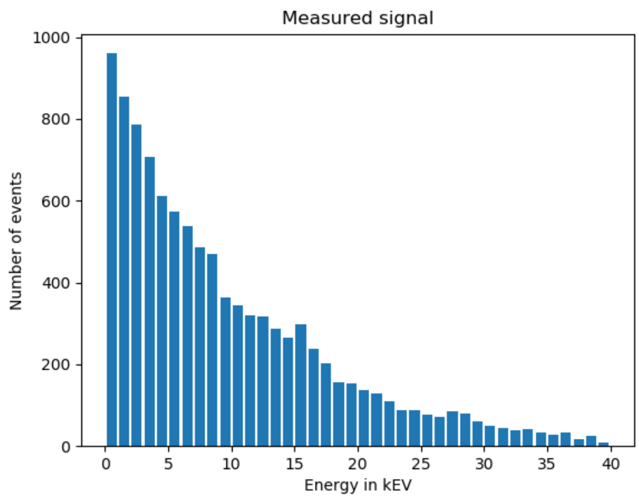

Figure 1: Measured Signal

2. **Question b**:- Histogram Plot for the background data

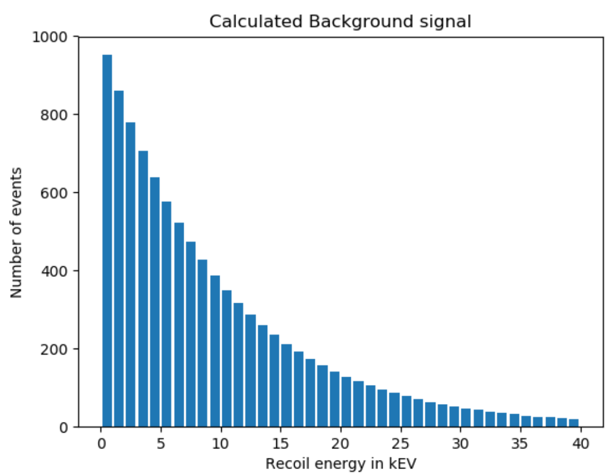

Figure 2: Background Signal

3. **Question C.1**:-Histogram Plot for $\sigma = .01$

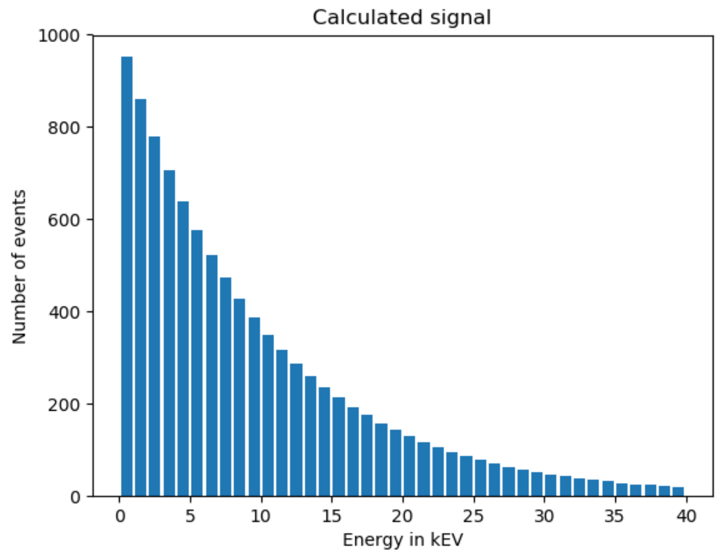

Figure 3: Calculated Signal for σ=0.01

4. **Question C.2**:-Histogram Plot for $\sigma = .1$

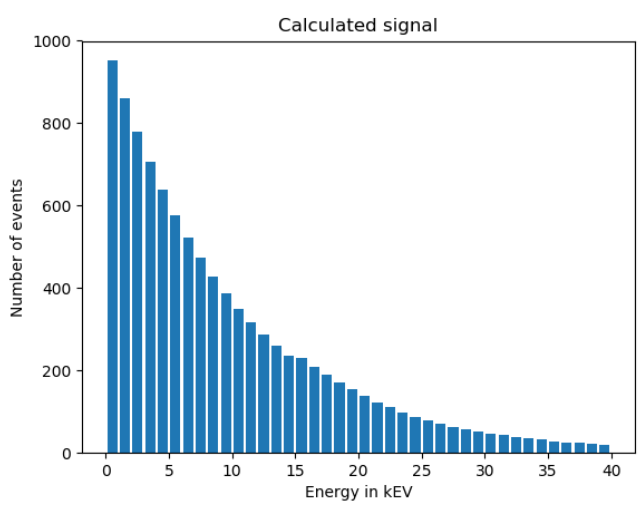

Figure 4: Calculated Signal for σ=0.1

5. **Question C.3**:-Histogram Plot for $\sigma = 1$

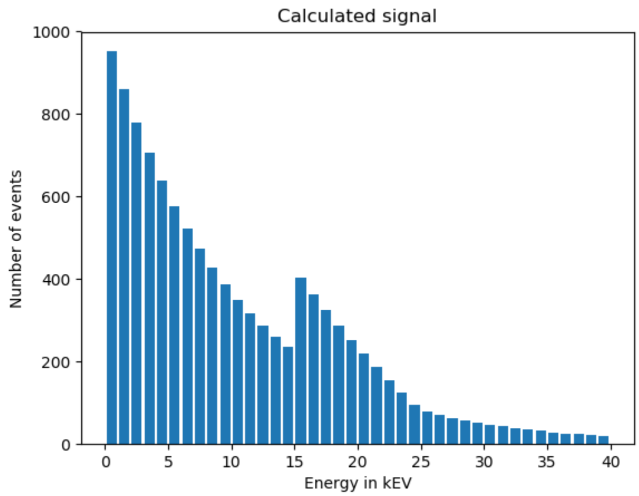

Figure 5: Calculated Signal for σ=1

6. **Question C.4**:-Histogram Plot for $\sigma = 10$

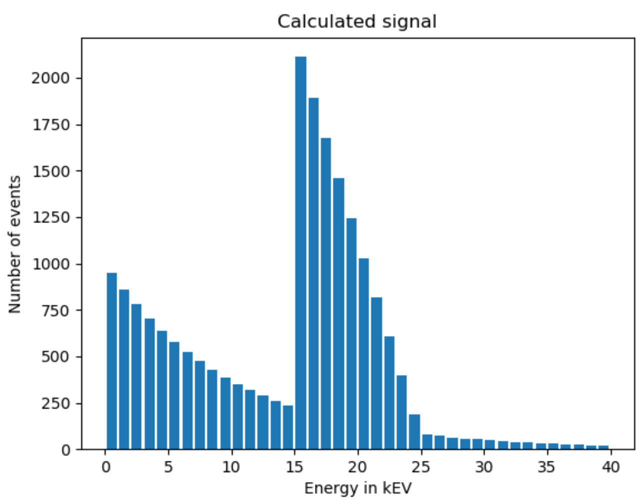

Figure 6: Calculated Signal for σ=10

7. **Question C.5**:-Histogram Plot for $\sigma \ = \ 100$

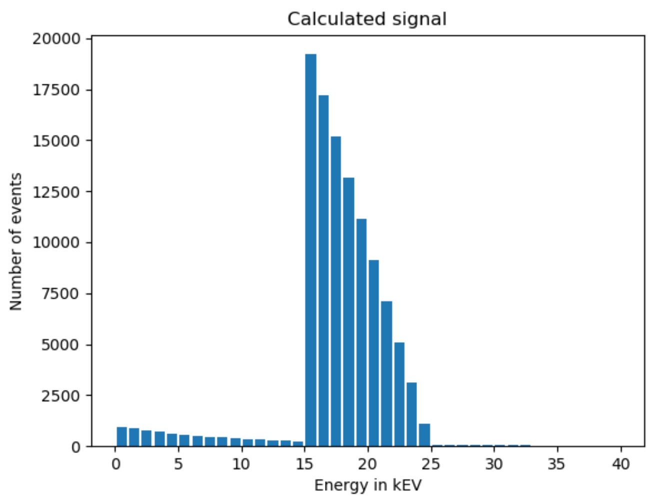

Figure 7: Calculated Signal for σ=100

We can clearly see the dark matter signal for $\sigma \ = \ 1, \ 10, \ 100.$

4 Log Likelihood

4.1 Log Likelihood Function

We claim that the Log Likelihood is of the following form:-

$$log\ \mathcal{L}\ =\ \sum_{n=1}^{40} m_i(log(B_i\ +\ dm_i))\ - B_i\ - dm_i\ - log(m_i!)$$

where

1. m_i is the value of the measured number of events at the ith bin.
2. B_i is the value of the number of events due to background distribution(calculated from the given distribution of Background Signal)
3. dm_i is the value of the number of events due to dark matter signal. It depends on the parameter σ.

This corresponds to the $\mathcal{L}$ being a product of possion distribution probabilities with a mean at each bin being $B_i\ +\ dm_i$ and m_i being the observed the measured number of final signal at each bin.

Reason for assuming a possionian functional form:-
Here the number of bins if fairly large(nearly 40), so for each individual event, the probability of falling into a particular bin,p_i, is very small. This correspsonds to saying that for each nuclei, the probability of having an energy between E_R and $E_R + 1$ (both in KeV) is very low. Also, the total number of particles is very large. Therefore, under small Bernoulli probability of the particle falling in a particular bin and large number of particles, we can approximate the distribution of particles in ith bin as Poissionian distribution around a mean of $B_i\ +\ dm_i(\sigma)$.
Intrinsically we have used the fact that a binomial distribution B_p^N under large N and small p, with Np = constant, tends to a Poissionian distribtuion.

5 Maximum Likelihood Estimation

From the histogram plots, we see that $\sigma\ = 1,\ 10,\ 100$ serve as a very poor estimation for the parameter. Also we see that $\sigma\ =\ 0,01\ and\ 0.1$ are nearly

identical in terms of approximating the measured signal. However, they fail to account for the abrupt changes in the measured signal completely and more or less resemble the background signal. Hence, among the given orders of parameters,we infer that σ is definitely not of the order of 0, 1 or 2.
Report of the findings from the data:-

1. MLE:-
   The value of σ was **0.01.**

2. The 1-σ interval was $[0.01,10^{1.06}]$. The interval extends on the RHS of the MLE.

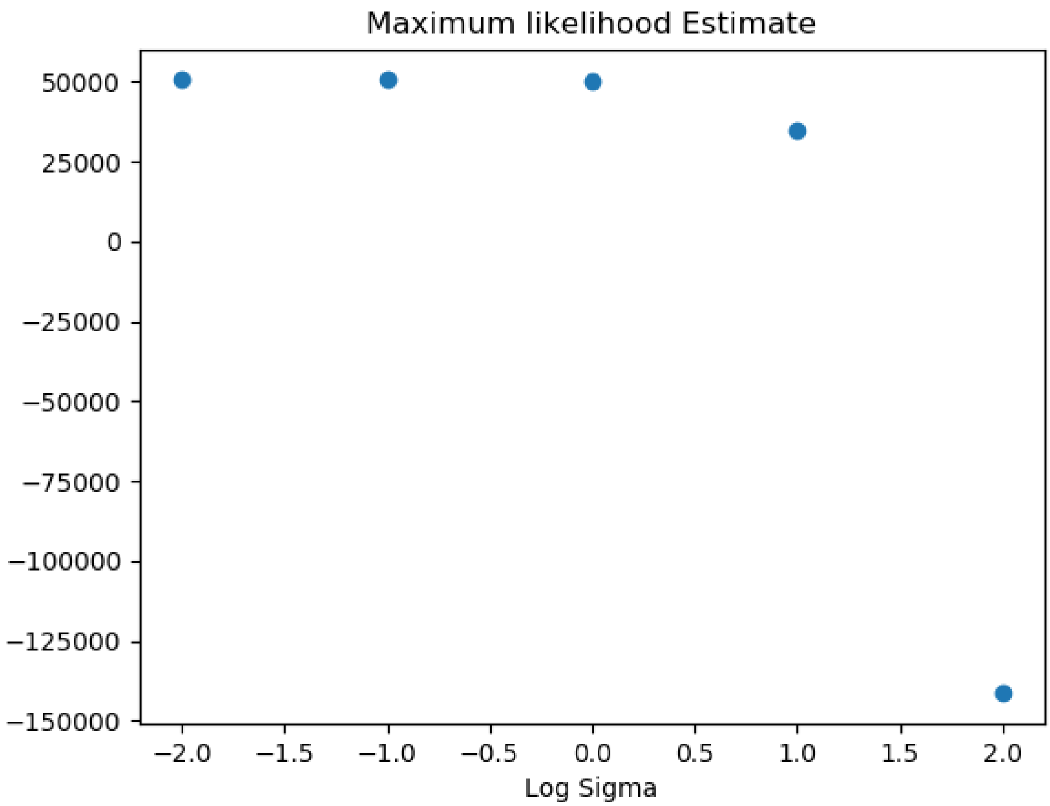

Figure 8: LogLIkelihood

After interpolation using a smooth curve, the graph looked like this:-

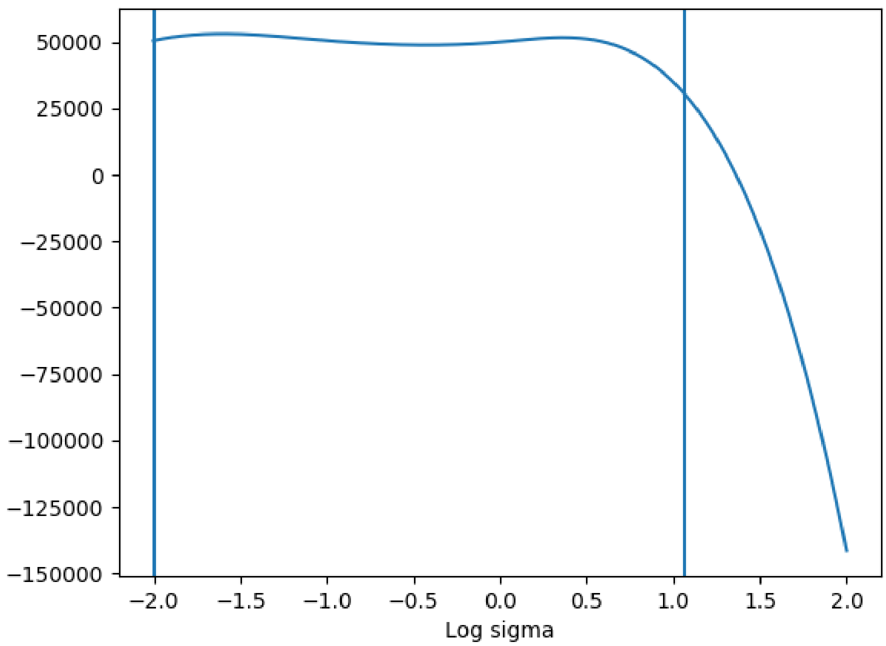

Figure 9: Interpolated Graph of Likelihood

The vertical line represents the 1-sigma interval from -2.0 to 1.065.

6 Team Responsibilties

- Project Leader - Aakash Marthandan
- Programmer - Raunak Dutta
- Web Manager - Ayush Bhardwaj
- Report Writer - Guru Kalyan Jayasingh

7 Website and Resources

This the link to our website.
Also, we have uploaded our code to Github repository. The link for the same can be found here.