



McGill

PROJECT 3 - GAMMA-RAY ASTRONOMY

Basic Analysis of Gamma-ray Observations

Name: Michael Jafs

Student Number: 261084068

Contents

Contents	2
1 Data Exploration	3
1.1 Loading and Visualizing the Gamma-ray Data Set	3
1.2 Source Identification and Event Distribution	4
1.3 Crab Pulsar Offset	5
2 Statistical Detection	5
2.1 Part 1: "ON" Events	5
2.2 Part 2: SkyCoord Separation Justification	6
2.3 Part 3: Determination of "OFF" Event Regions	6
2.4 Part 4: "OFF" Region Position Justification	7
2.5 Part 5: Significance Calculation	7
2.6 Part 6: Significance Interpretation	7
3 Cut Optimization	8
3.1 Part 1: Mean Scaled Width and Length Distributions	8
3.2 Part 2: Cut Optimized Significance	8
3.3 Part 3: Cut Significance Heatmap	9
3.4 Part 4: Cut Improvements	10
4 References	11

1 Data Exploration

1.1 Loading and Visualizing the Gamma-ray Data Set

After loading the Gamma-ray data set into python, the sky positions in right ascension (RA) and declination (DEC) of all contributing event detections can be seen in Figure 1. This plot serves as a nice indicator of event positions and makes identifying the center of the cluster of events straightforward, however, since many of the events are overlapping, the overall distribution is challenging to make out. In Figure 2, a 2D histogram shows again the RA and DEC of the collection of events, but alleviates confusion as to the nature of the distribution in the center of the cluster. From the 3D surface plot of the 2D histogram (right plot in Figure 2), we can even better identify the way in which the distribution appears to be peaked around roughly 40 counts with a single brighter collection of events (consisting of ~ 60 counts) at a particular RA and DEC combination. Since this brighter hotspot appears around the Crab pulsar coordinates, this is likely our gamma-ray signal of interest.

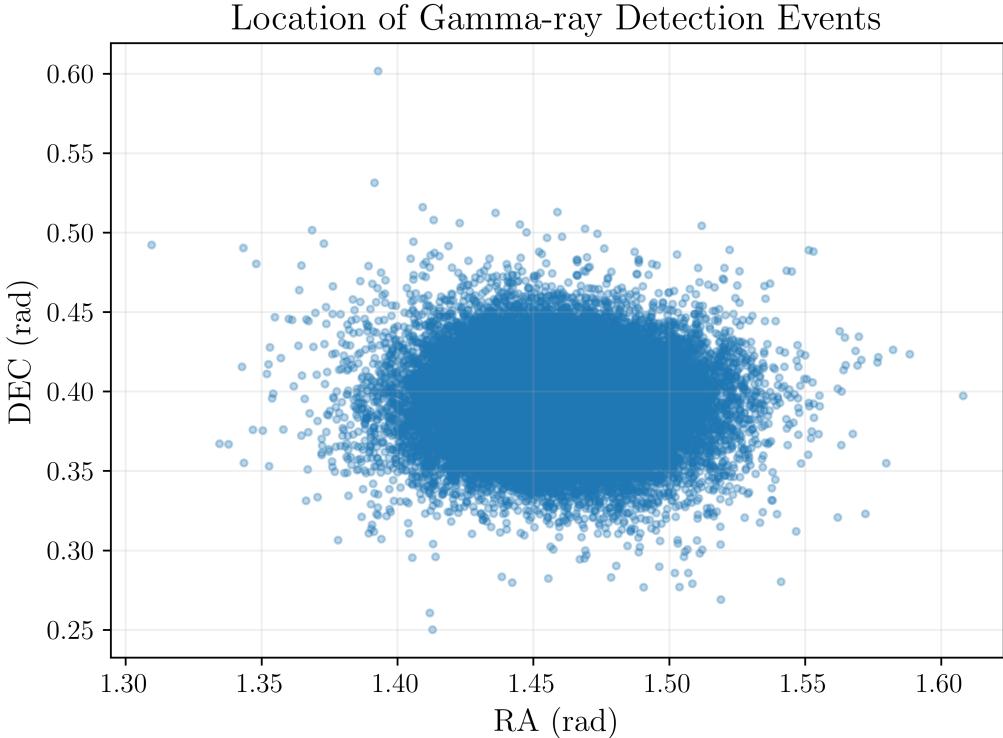


Figure 1: Sky positions in right ascension (RA) and declination (DEC) of Gamma-ray events from the Crab pulsar collected using VERITAS.

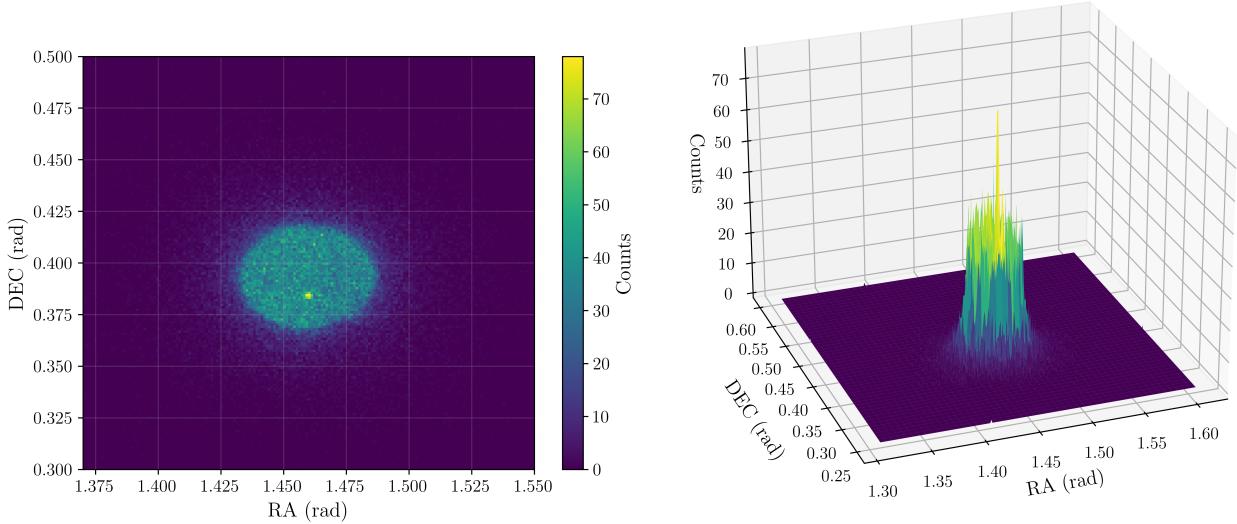


Figure 2: (Left) Histogram of Gamma-ray events from the Crab pulsar first shown in Figure 1. (Right) 3D surface plot realization of the identical plot shown on the left. From this perspective, the statistical distribution is easier to identify. The one bright peak of events at the Crab pulsar RA and DEC coordinates is also highly evident from looking at the 3D surface plot.

1.2 Source Identification and Event Distribution

Using the *gamma-sky* map available at (<http://gamma-sky.net/>), the source was identified to be the well known Crab pulsar. This was determined by locating the approximate RA and DEC at the center of the cluster in Figure 1, converting the RA and DEC to galactic coordinates and finding the source that best matched the corresponding coordinates. The output from *gamma-sky* (including the Crab pulsar coordinates) was:

```

Basic Info
Common name: Crab pulsar
Gamma names:
Fermi names:
Other names: PSR 0531+21, V* CM Tau
Location: gal
Class: psr
TeVCat name: TeV J0534+220p (TeVCat ID: 129, TeVCat2 ID: 06r8mb)
TGeVCat name: (ID: -9223372036854776000)
Discoverer: magic
Discovery date: 2008-11
Seen by: magic, veritas
Reference:

Position Info
SIMBAD

RA: 83.633 deg
DEC: 22.014 deg

```

GLON: 184.557 deg
GLAT: -5.784 deg

As for the distribution of events, this is best shown in Figure 2. From both the left and right plots, we can see the distribution looks almost uniform in number of events out to a particular RA and DEC. In other words, in the hotspot region, the number of events seems to be mostly similar aside from what looks to be random fluctuations until a certain radius where the number of events falls off sharply. That said, examining the 3D plot on the right side of Figure 2, shows there to be a slight trend in more counts as we approach the center of the collection of events, leading to something that looks like what would have been a Gaussian distribution with the detector being saturated after a certain point. Though the distribution may be mostly circular, it appears to be slightly elongated along the declination axis. Finally, there is also a strong pulse of events toward the center in RA and center-bottom in DEC of the hotspot, this is likely gamma-ray photons from the Crab pulsar as the RA and DEC coordinates closely align with those shown above.

1.3 Crab Pulsar Offset

The pointing of the observation is:

- RA = 83.6329 deg
- DEC = 22.5258 deg.

The offset of the Crab pulsar from the pointing observation is then simply:

- $\text{RA}_{\text{offset}} = 83.633 - 83.6329 = 0.0001 \text{ deg}$
- $\text{DEC}_{\text{offset}} = 22.5258 - 22.014 = 0.5118 \text{ deg.}$

Since the offset in the RA dimension is closest to zero, this is the coordinate offset we take to be zero for the duration of the project.

2 Statistical Detection

In this section, we wanted to estimate the statistical significance of the measured excess in events. This was done by estimating the number of events in the region assumed to contain the source and comparing to the number of events in a collection of “background” regions.

2.1 Part 1: ”ON” Events

To determine the number of events corresponding to the excess, a function was written to sum the total number of events within a 0.1 degree radius from the Crab pulsar RA and DEC position. This number is known as the number of “ON” events or (N_{on}) and was found to be 566. The algorithm worked by looping through each event sky position and for each (RA, DEC) pair, calculating the angular separation using astropy’s SkyCoord “separation” method. Given the length of the events data table ($\sim 90,000$) entries, a parallel implementation using numpy’s ProcessPoolExecutor was written to accelerate the total compute time by a maximum factor of ~ 8 depending on the range of cpu cores used (a maximum of 22 were available).

2.2 Part 2: SkyCoord Separation Justification

Instead of using the euclidean distance formula:

$$dist = \sqrt{\Delta_{RA}^2 + \Delta_{DEC}^2} \quad \text{with} \quad \Delta RA = RA_1 - RA_2, \quad (1)$$

as mentioned in the previous subsection, the built in SkyCoord separation method was used to calculate angular distances. The reason is that although the distance formula above is correct for flat space, it breaks down over large enough distances when working on the surface of a sphere. Since RA and DEC are angular coordinates, we cannot simply add their squares to compute their distances. Instead, we require the more precise great-circle distance which calculates the distance between two point on a sphere. It is this latter calculation that astropy uses when determining the angular distance of two sky positions. If the two sky positions are far enough apart, distances using the great-circle method versus using Eq. 1 will begin to diverge by a noticeable amount. Therefore, any distance that is sufficiently large on the celestial sphere would imply that using Eq. 1 is inappropriate. This is crucial for the calculation of events in a particular region, since using the approximate (euclidean distance) formula could result in an over-counting of events.

2.3 Part 3: Determination of “OFF” Event Regions

Just as we determined the number of “ON” events by summing the number of events within a 0.1 degree radius of the Crab pulsar pointing, we determined the number of “OFF” events by counting the number of events with 0.1 radii of pointings in the 3 other cardinal directions from the telescope pointing. To be more concrete, as described in the section on the offset calculation, the Crab pulsar was offset from the telescope pointing by an amount RA=0 degrees and DEC=0.5118 degrees. This means that we should select our background regions to be an equal separation from the telescope pointing and in the 3 other cardinal directions as shown in Figure 3.

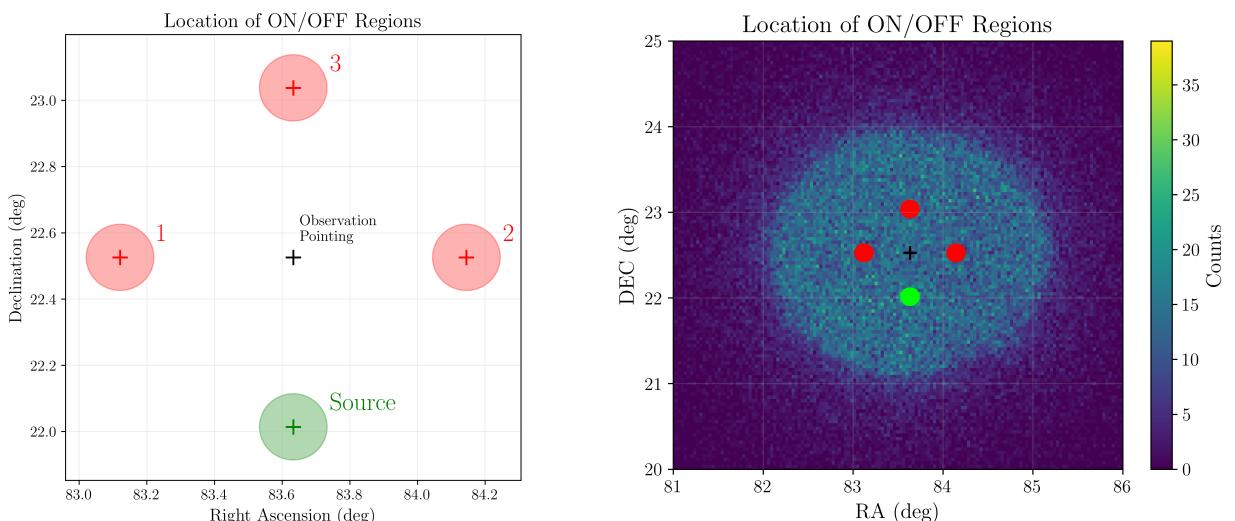


Figure 3: (Left) Locations of both ON and OFF regions with the size of each 0.1 degree region drawn on for scale. (Right) Each region shown on top of the events histogram from Figure 2 clearly outlining the position of each region relative to the collection of background events. In both plots, the observation pointing is denoted with a black cross at the center of the image.

The left plot in this figure clearly denotes the location of each “OFF” region and also highlights the size of the 0.1 degree regions relative to their locations. On the right, the “ON” and “OFF” regions are painted on the events histogram from Figure 2 to show their location relative to the spread of background events. Running this part of the computation, revealed the number of “OFF” events to be 321, 309, and 348 for regions 1, 2, and 3 respectively.

2.4 Part 4: “OFF” Region Position Justification

Instead of choosing “OFF” regions at random provided they contain some number of background events, they were chosen such that their centers were the same angular distance from the telescope pointing as the source. In order to claim a result significance, we need to be able to establish a null hypothesis that, for example, the number of “ON” events are from background events and not source event counts. This means that each region chosen should not allow geometrical inconsistencies that allow for the potential of different answers depending on where the events are counted. If the regions are spaced away from the telescope pointing by different amounts, the photons should have had different amounts of time to reach the detector will result in different solid angles as measured at the detector position. This would simply result in differing flux, that when unaccounted for, would not allow for a straightforward statistical significance calculation.

2.5 Part 5: Significance Calculation

From the famous Li & Ma [1] formulation of the maximum likelihood, we determined the significance of our result. In other words, using the equation:

$$S = \sqrt{-2 \ln \lambda} = \sqrt{2} \left\{ N_{\text{on}} \ln \left[\frac{1 + \alpha}{\alpha} \left(\frac{N_{\text{on}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] + N_{\text{off}} \ln \left[(1 + \alpha) \left(\frac{N_{\text{off}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] \right\}^{1/2}, \quad (2)$$

and taking α to be 1/3 since this is the ratio of the “ON” event area over the “OFF” event area, the significance was found to be 10.14 when rounded to two decimal places. From the original paper [1], the assumption about the statistical distribution of signal and background is that the random variables N_{on} and N_{off} are Poisson distributed. This is appropriate for gamma-ray physics since gamma-ray events occur consistently through time, but typically not at the same time, which is also the requirement for a Poisson random variable.

2.6 Part 6: Significance Interpretation

The statistical significance formula above (Eq. 2) is derived from a likelihood ratio test wherein the likelihood calculated under the null hypothesis that the source events N_{on} are really background events is compared to the likelihood under the alternate hypothesis. In this likelihood ratio test, the authors claim that if the null hypothesis is true, then the variable $-2 \ln \lambda$ should follow a standard normal distribution and we can take the result as the observed significance. This means that our significance of 10.14 is really a 10.14 sigma result under the null hypothesis that $\langle N_S \rangle = 0$. In other words, if we expect the significance to follow a standard normal distribution, then our result lies over 10 standard deviations from the mean under the assumption that the source events come from the background. We should therefore reject with high

significance the hypothesis that our result is purely background events under the frequentest interpretation.

3 Cut Optimization

At this stage we can recognize that since the source is photons and the background is hadrons, there should be a different and distinct shape when comparing the light produced by gamma-rays versus hadronic showers. Therefore, using the extra information about each event's detection width and length, we can be more selective about what data to use when calculating the significance.

3.1 Part 1: Mean Scaled Width and Length Distributions

The mean scaled width and mean scaled length distribution for all events are plotted in Figure 4. Clearly, most events have a MSCL and MSCW of around 1.9, however, taking the information provided in the project outline regarding the typical shapes of gamma-ray events compared to hadronic showers, it is likely these events with large length and width correspond to the background hadrons and not events from the Crab pulsar.

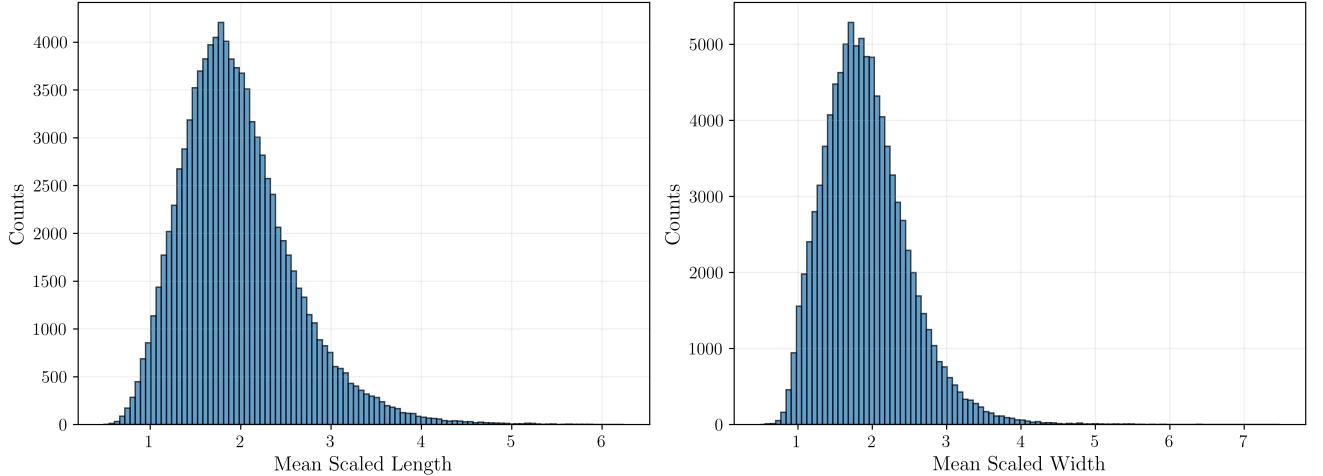


Figure 4: (Left) Mean scaled length (average length of telescope images) distribution. (Right) Analagously, the distribution of average width of the telescope images or the mean scaled width.

3.2 Part 2: Cut Optimized Significance

Since we expect gamma-ray events to appear distinctly different compared to hadronic showers, we would like to make selection cuts over the MSCLs and MSCWs and re-determine the significance of the detection at particular MSCL, MSCW cuts. The following cuts were made:

- MSCW: $0.9 \rightarrow 1.9$, steps of 0.1
- MSCL: $0.9 \rightarrow 1.9$, steps of 0.1.

Algorithmically, the first step was to creat a list of indices corresponding to each MSCL and MSCW cut. A small sanity check to be sure the function picking out the correct indices was working correctly can be seen in Figure 5. Since it near impossible to check, by hand, that

each selection cut corresponded to what it was supposed to, we could still check that there were no cuts with shared indices and that cuts corresponding to larger MSCL and MSCW values contained more indices. This latter point can be seen simply by looking at the MSCL and MSCW distributions and noticing that the distribution has an increasingly larger number of counts for values approaching 1.9 from the left. These two points are exactly what can be seen in Figure 5.

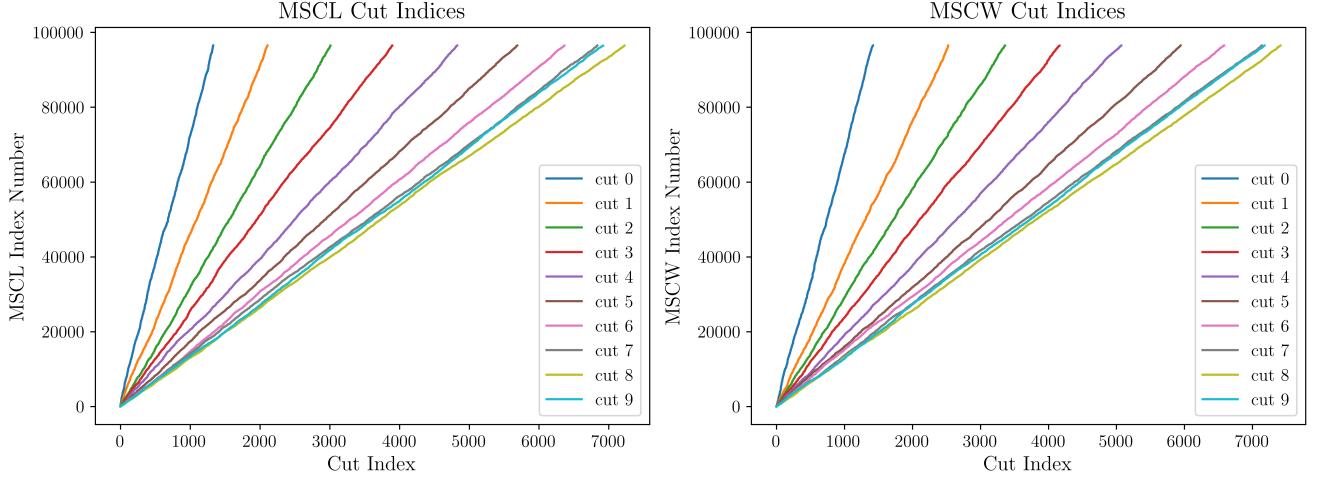


Figure 5: Indices corresponding to each of the 10 cuts. Left hand plot corresponds to the MSCLs and the right corresponds to the MSCWs. From the plot, it difficult to tell for cuts 7 - 9, though indeed each cut shares no common indices indicating the algorithm correctly picked out the unique indices each set of cut parameters.

With the correct cut indices in hand, we could simply repeat the steps to find the significance in the earlier section though this time looping through each cut combination and calculating the significance while only using events corresponding to each cut.

3.3 Part 3: Cut Significance Heatmap

The results from computing the significance for each MSCL and MSCW cut combination can be seen in Figure 6. Indeed these 2D plot clearly demonstrate that for values of MSCL and MSCW ~ 0.9 , the significance is the greatest and also at value of 13.68 exceeds the significance found earlier of 10.14. This indicates that optimizing for a set of cuts motivated by prior knowledge regarding the shape of signal versus background images can in fact lead to a more significant detection. Furthermore, the optimal set of cuts confirms the statement that the MSCLs and MSCWs around 1.9 with the highest counts in Figure 4 were likely background events, since counts corresponding to MSCL, MSCW ~ 0.9 are small compared to the distribution of lengths and widths and we know the source emission is small compared to the number of background events from examining, for example, Figure 2.

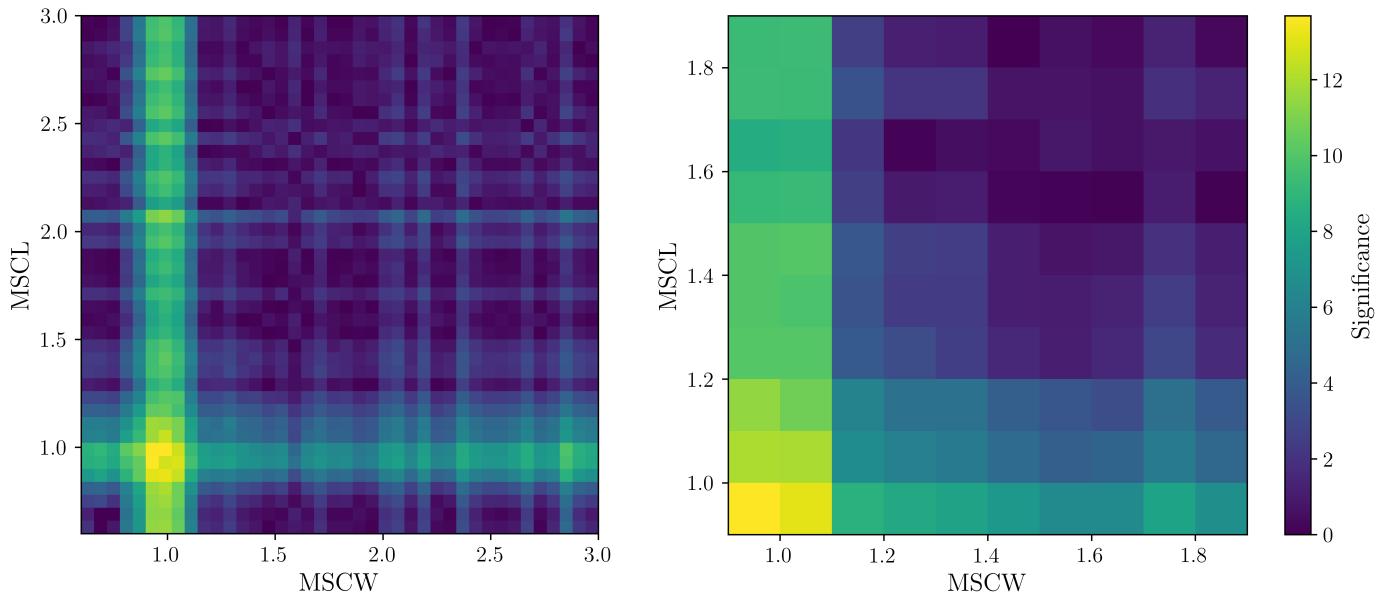


Figure 6: 2D significance array corresponding to (left) the parameters outlined earlier (both MSCL and MSCW $0.9 \rightarrow 1.9$, steps of 0.1) and (right) both MSCL and MSCW $0.6 \rightarrow 3$, steps of 0.06. The latter set of cuts indicates that what could have potentially been seen as the optimal set of cuts due to a poorly chosen span of parameter space values on the right hand side indeed does enclose the optimal MSCL and MSCW combination, as going lower than ~ 0.9 results in lower significance. The significance at optimal cut values is 13.68 when rounded to two decimal places.

3.4 Part 4: Cut Improvements

The most straightforward way to improve these cuts is to simply sample more of the parameter space and with smaller cuts as can be seen in the left plot of Figure 6. This way, we can be sure we have in fact localized the MSCW and MSCL to the highest significance values in the parameter space. The only caveat in this technique is the notion that choosing the step size to be too small may cause the significance to peak in the incorrect location since the optimal average width and length of images is physically motivated, corresponding to gamma-ray events. Instead, one could choose to define a third axis that loops over step sizes so that the optimal cut combination would not only find roughly the MSCL and MSCW combinations with highest significance, but also the optimal width of the cuts. This may be computationally costly, but seems like a highly parallelizable problem that could be computed using GPU cores.

4 References

- [1] T. P. Li and Y. Q. Ma. Analysis methods for results in gamma-ray astronomy. , 272:317–324, September 1983. [doi:10.1086/161295](https://doi.org/10.1086/161295).