CSE 572: Data Mining

Project 1

Report

Submitted to:

Prof. Ayan Banerjee

Arizona State University

Submitted by

Srinivas Shashank Mulugu

ASU ID: 1217077238

Email: smulugu@asu.edu

a) Four different kinds of features were extracted from the given CGM data arrays. These can be classified into three distinct types of features Sliding-window, Time-domain, Frequency-domain and statistical features. They are described below:

Sliding-window Features:

Moving Average (MA): The moving average is calculated by computing mean over a small window of the whole data. In the following Average vs Time graph, the blue line shows the moving average and the orange line shows the raw data. It gives us a smoother estimate of the raw data.

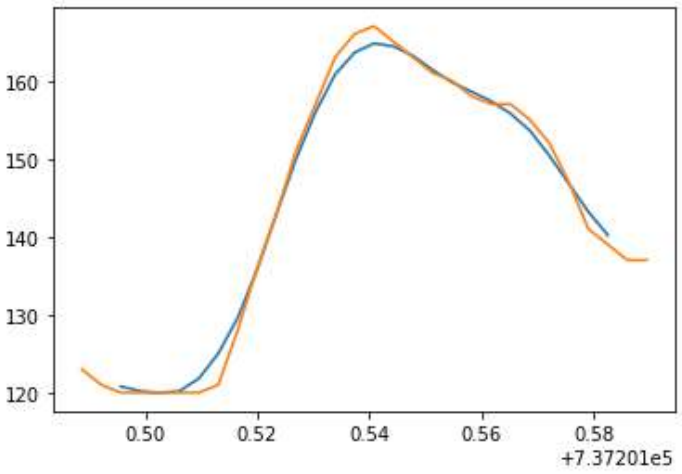


Fig. 1. Moving Average-time graph overlapped with moving average

Time-Domain Features:

Velocity: The velocity of the CGM data is calculated by estimating the rate of increase/decrease in CGM level over time for all the datapoints. Figure 2 shows an example of the velocity-time graph.

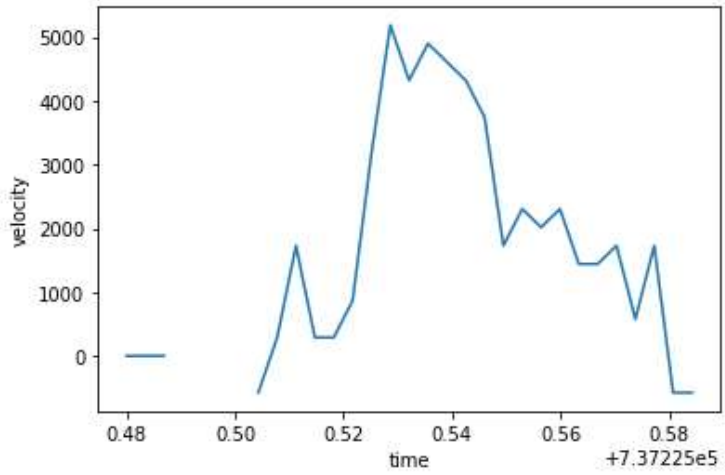


Fig. 2. Velocity-Time graph

Frequency-Domain Features:

- Fast Fourier Transform (FFT): The FFT converts our time-domain CGM data into frequency-domain data. Figure 3 shows an example of the Amplitude vs Frequency graph after performing FFT.

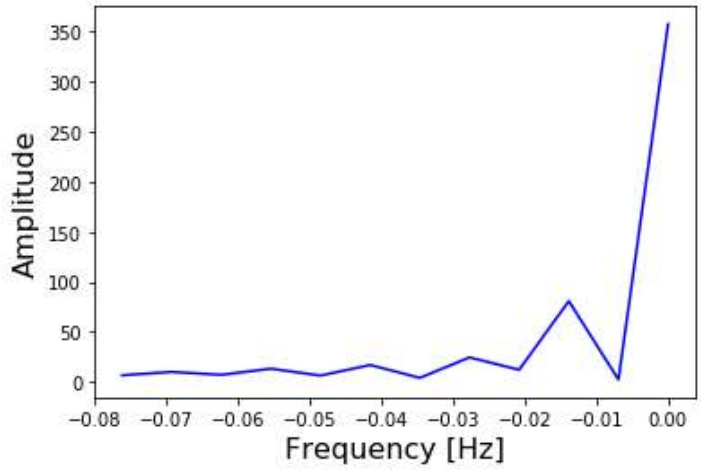


Figure 3. Amplitude vs Frequency graph

- Power Spectral Density (PSD): The PSD also converts the time-domain data into frequency-domain. Figure 4 shows the PSD vs Frequency.

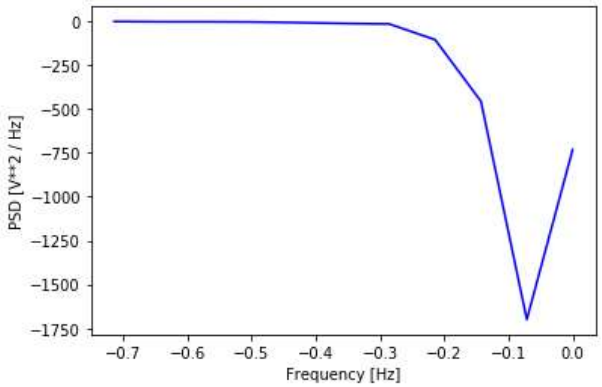


Fig. 4. PSD vs Frequency graph

Statistical Features:

- Mean: We just find average of all the rows of the CGM data. It is one of the most obvious places to start as it gives us useful information regarding the CGM data.
- Standard Deviation: In standard deviation we can identify the total amount of variation or dispersion in the values.
- Maximum Value: The maximum value of the CGM data is also calculated and included as part of the features.

b) The reasons for picking the above features are given below:

Sliding Window Features:

Moving Average: This helps in smoothing out the data by filtering out the noise as the data is approximated to its mean in every window. In this case the window size used was for every 25 minutes. Finding the MA has helped in estimating the peaks easily as the data can vary quite a lot in some windows and determining local maxima may not be very easy.

Time-Domain Features:

Velocity: The velocity can be very useful because we can identify any/all changes in the CGM levels over the progression of time. Since there can be a rapid increase or decrease in the velocity the indices where the velocity changes from positive to negative or vice versa were noted as it denotes a dramatic shift in the CGM levels.

Frequency-Domain Features:

Fast Fourier Transform (FFT): It helps in identifying the peaks in the data extremely easily which we can assume is a blood glucose spike due to intake of food.

Power Spectral Density (PSD): It helps us identify where the variance of the data is maximum or minimum against frequency. This information can be extremely useful because it tells us about the variations in the CGM data such as the peaks and the dips.

Statistical Features:

The statistical features give an idea for the overall series unlike the sliding window which calculates only the features for a certain window.

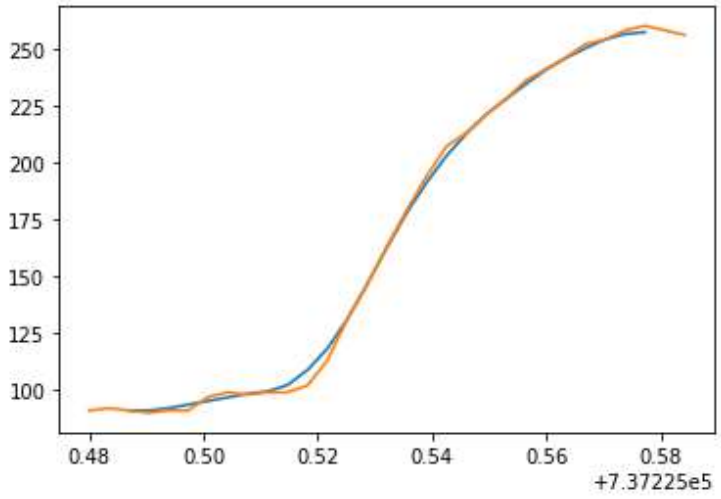
c) The values of the features observed in the feature matrix are given below:

Sliding Window Features:

Moving Average (MA): The MA values for the first time series of the first patient are given below:

[257.2, 256.4, 254.0, 250.2, 245.8, 240.6, 234.4, 227.8, 221.0, 212.6, 202.8, 191.2, 177.8, 162.6, 146.4, 131.0, 118.2, 108.8, 102.2, 99.4, 98.4, 96.8, 95.2, 93.6, 92.0, 91.0, 91.0]

The graph with the raw data, MA vs time is as shown below:

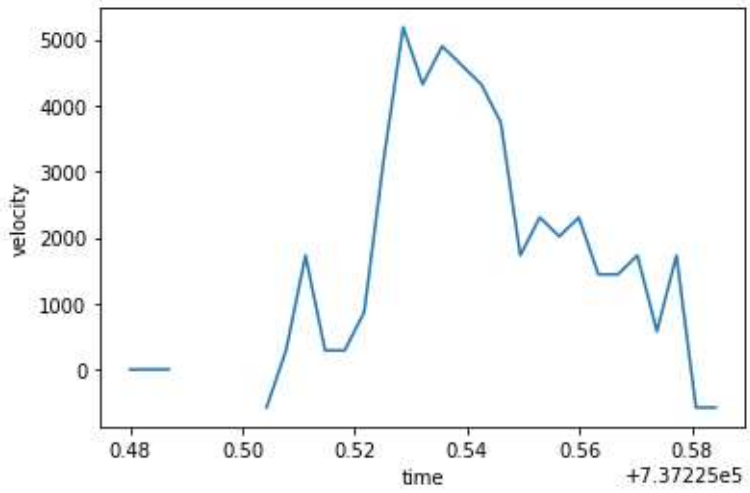


In the above graph orange represents raw data and blue represents the moving average with a window size of 25 minutes. The idea of picking this feature is that the MA would give a smoother estimate and essentially “denoise” the raw data to some extent which is apparent from the above graph.

Time-Domain Features:

Velocity: The values of velocity as well as the velocity-time graph for the first timeseries of the first patient is given below:

[180.8242775592888, -575.9998798370611, -576.0000343322774, 1728.0001029968323, 576.0000536441853, 1727.9995236398104, 1440.0001823902362, 1440.0000375509271, 2304.0001373291097, 2016.0001201629711, 2303.9995193482446, 1728.0001029968323, 3744.000223159803, 4320.0004023313895, 4607.998729706161, 4896.0006201268025, 4320.000112652781, 5184.000308990497, 3168.000188827526, 863.9998197555918, 288.0000171661387, 288.0000171661387, 1728.000160932556, 287.99992060663504, -576.0000729560944]

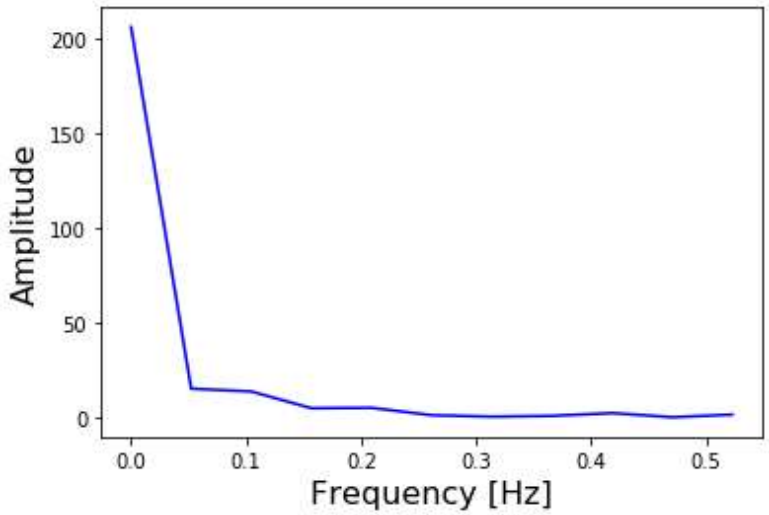


In the above graph the rapid changes in the CGM data are observed which helps us identify the dips and peaks in the data. So, velocity of the CGM data is a good feature for the data as it keeps us informed about the instantaneous changes in the data.

Frequency-Domain Features:

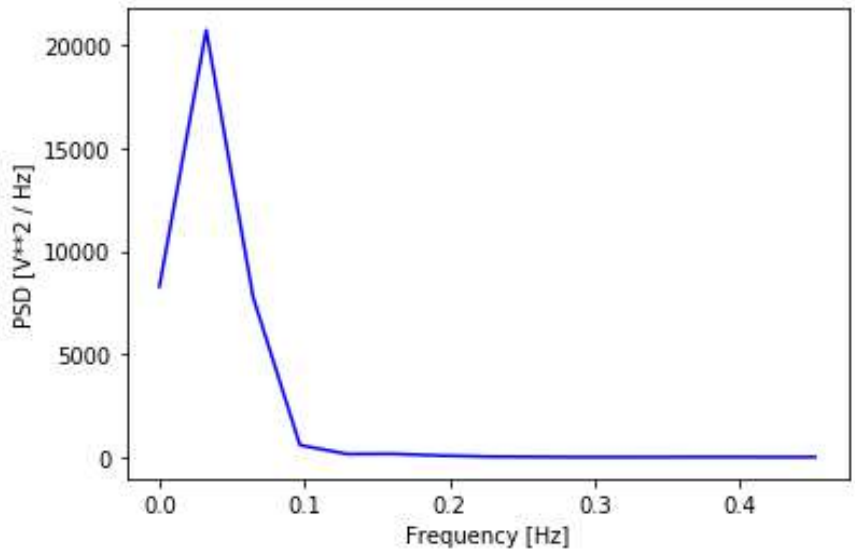
Fast Fourier Transform (FFT): The values of the FFT followed by the amplitude and frequency graph for patient 1 is given below:

205.82608695652172, 15.437033097267058, 13.956166119626115, 5.140375960788145, 5.348329779690983, 1.4979728513565869, 0.6899287569041802, 1.1024104018596685, 2.59819182556168, 0.4116612243346639, 1.7889898885702586]



From the above values and graph it is evident that FFT picks out the peaks in the data. The peak value from the data is the first value of the array i.e., 205.82. The inclusion of FFT as a feature can be crucial because it gives us information about the occurrence of data in against the frequency, which is useful because we can identify how often the peak values are achieved in the timeseries.

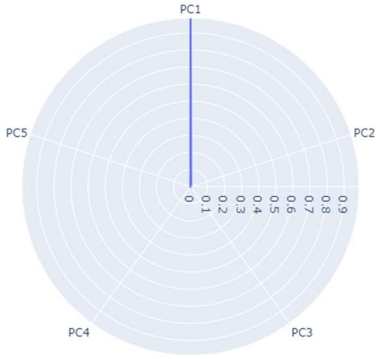
Power Spectral Density (PSD): The PSD shows the strength of variations in the data in each frequency. In the below graph we can see that there is strong variation in data in the lower frequency but as the frequency increases the PSD falls off quite dramatically. Therefore, it can be used to predict in theory, when there can be a CGM spike based on the data.



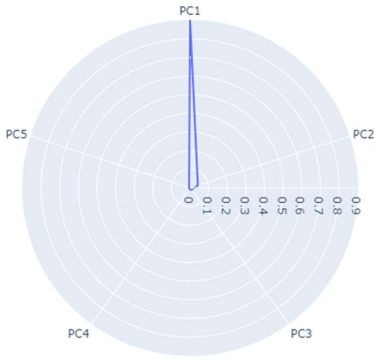
d) After the Feature Extraction process as described above, a feature matrix with 111 columns was generated for each patient. The values of these feature matrices can be found in the attached .txt files named Feature Matrix <Patient Number>.txt.

e) After getting the features from the data, these 111-140 features were passed to PCA to reduce the dimensions to 5 components. Below are the graphs for each patient:

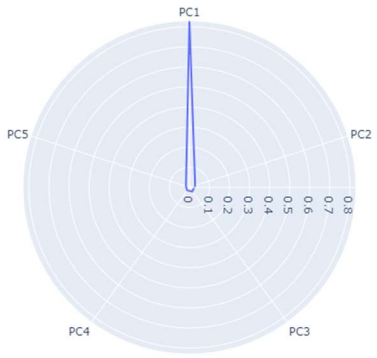
Patient 1:



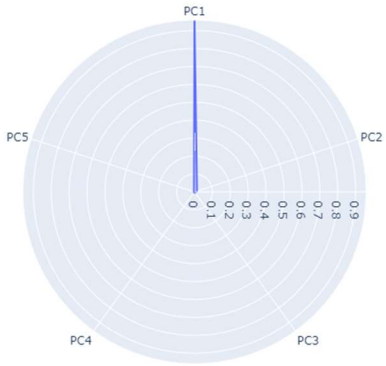
Patient 2:



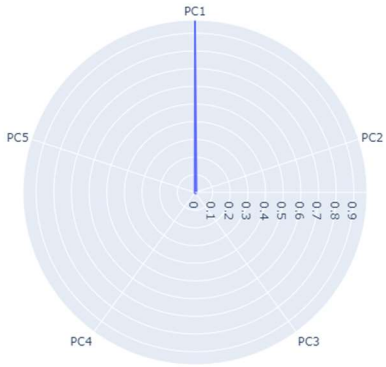
Patient 3:



Patient 4:



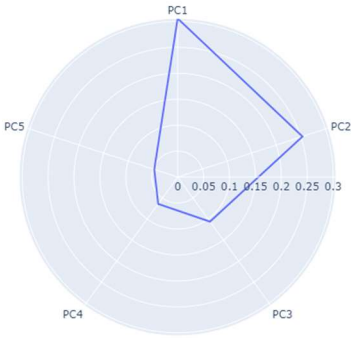
Patient 5:



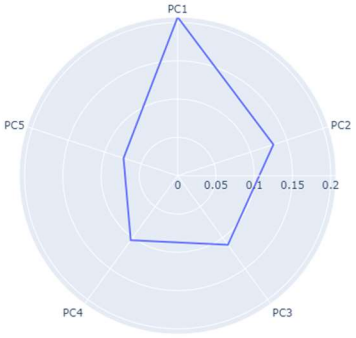
The 5 radar charts above describe the variance ratio that each of the top 5 principal components (PC) after performing PCA. From the above graphs it is apparent that there is one feature that is clearly on the top of all the rest of the features because of its high

variance. Based on testing out different combinations it was observed that this feature was the Power Spectral Density (PSD) which turned out to be the most important feature. After removing this it was observed that the variance ratio of other features increased significantly as shown below:

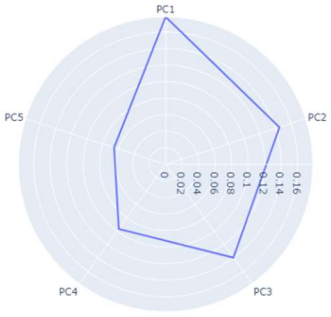
Patient 1:



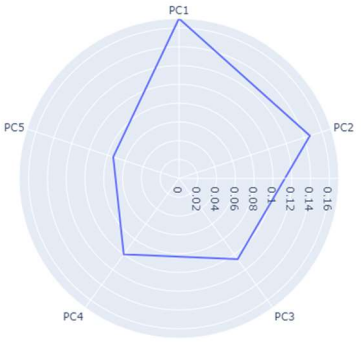
Patient 2:



Patient 3:



Patient 4:



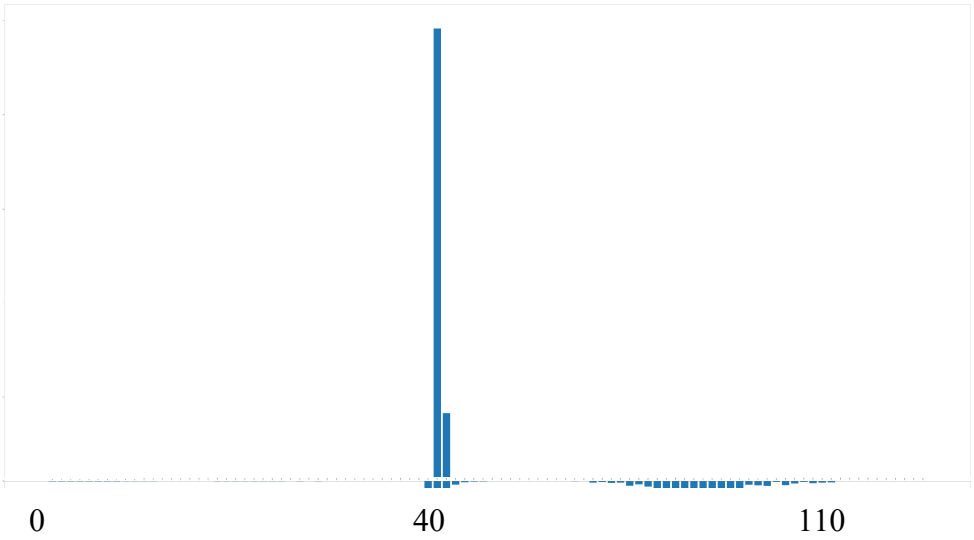
Patient 5:

Now it is clear that PSD had a very high variance ratio compared to the other features at least because the radar charts above for the 5 patients indicates a much more even split between all the features than before.

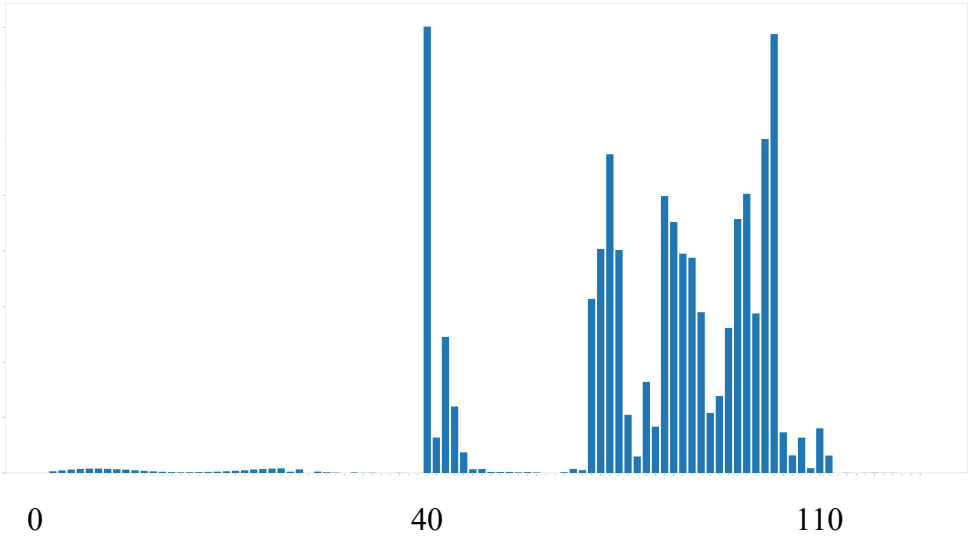
f) For all the top 5 principal components, the base component is PSD as shown in the graphs below for each patient PSD (~40 – ~70) and Velocity (70-110) is a significant part of each principal component.

Patient 1:

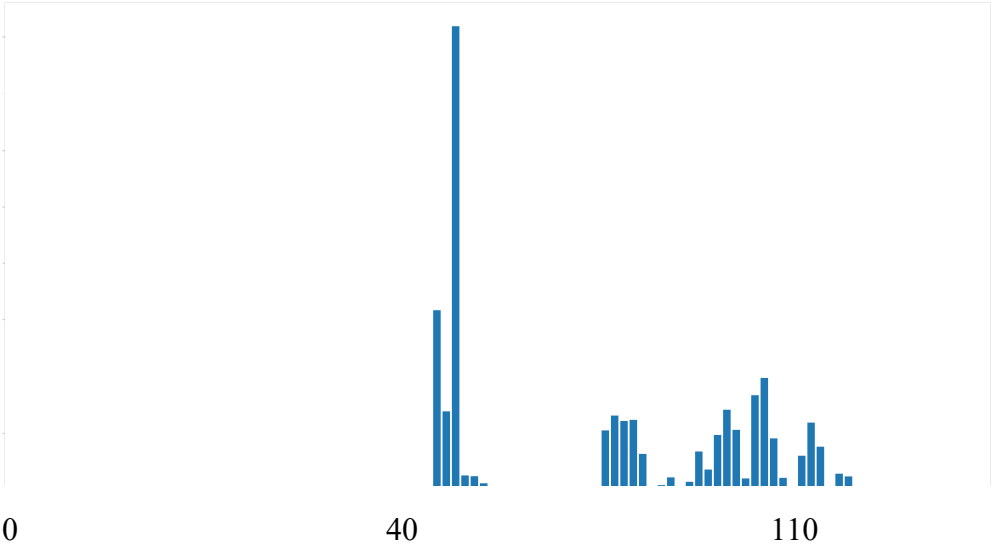
Principal Component 1



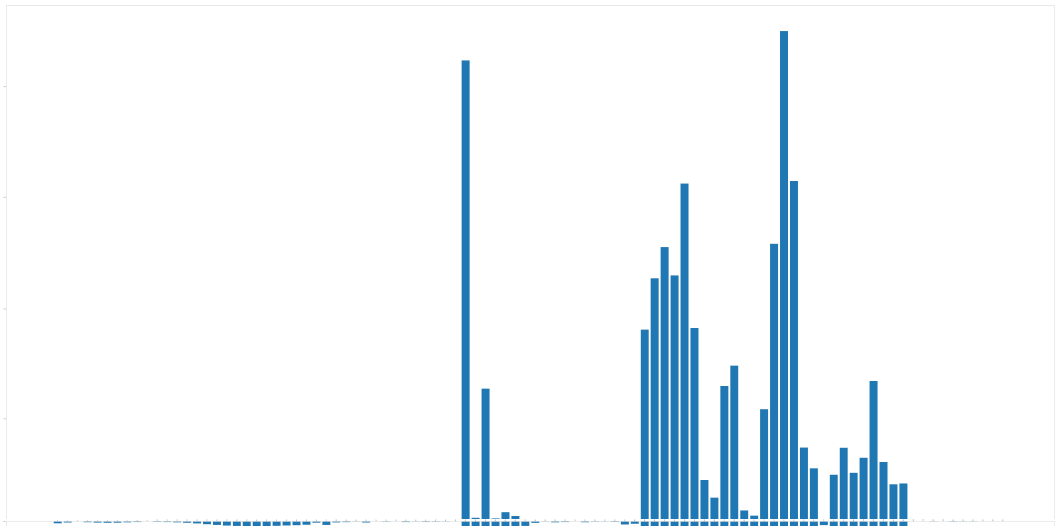
Principal Component 2:



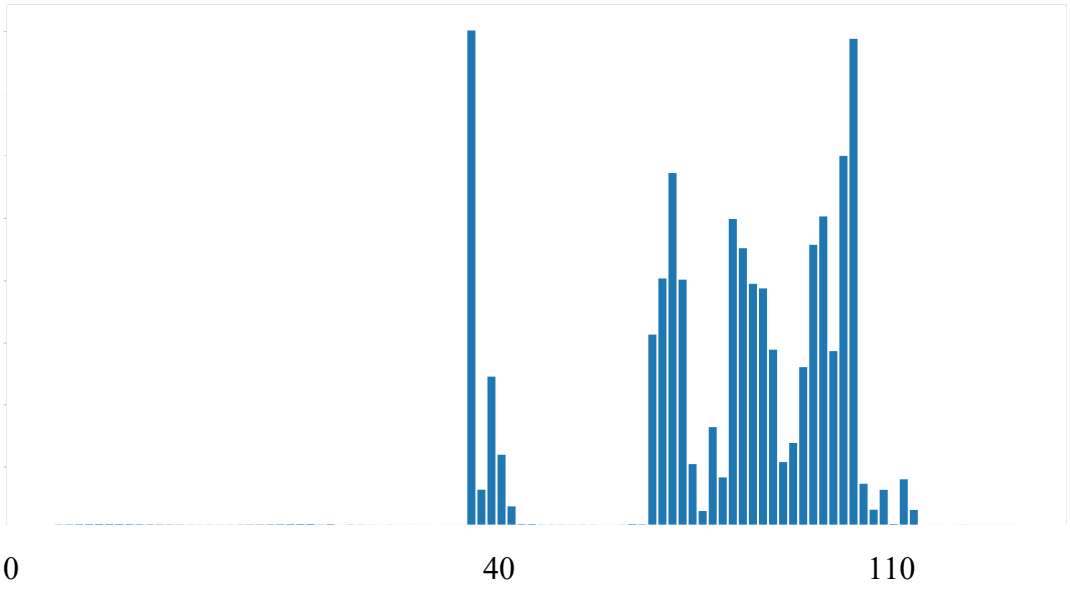
Principal Component 3:



Principal Component 4:



Principal Component 5:



From all the above graphs of the principal components the top 2 components from the data are the PSD and the velocity as they are a part of every single one of the principal components. The other three base components which are a part of the principal components are FFT and moving average to a very small extent.