**Activity Project 1 Report**

DATA MINING

CSE 572: Fall 2018

**Submitted to:**

**Professor Ayan Banerjee**

**Ira A. Fulton School of Engineering**

**Arizona State University**

**Submitted by(GROUP 20):**

**Vivek Agarwal (**[**vagarw14@asu.edu**](mailto:vagarw14@asu.edu)**)**

**Suchita Vichare (**[**ssvichar@asu.edu**](mailto:ssvichar@asu.edu)**)**

**Supriyaa Damodaraswamy (**[**sdamoda3@asu.edu**](mailto:sdamoda3@asu.edu)**)**

**Sukanya Panda (**[**spanda7@asu.edu**](mailto:spanda7@asu.edu)**)**

**Ayush Ray (**[**aray29@asu.edu**](mailto:aray29@asu.edu)**)**

**October 10, 2018**

# 1. Introduction

The Activity Recognition Project is a part of the course requirement for Data Mining (CSE 572) for the session of Fall 2018 at Arizona State University. The goal of the project is attempting to develop a computing system that can understand human activities through various components like identifying human activities, segment sequence of activities and identifying unknown activities. The data are collected from Myo Sensors which have components of accelerometer, gyroscope, orientation and EMG. There are different aspects of data collection which includes data for cooking, driving, eating, keyboard, and playing guitar.

# 2. Team Members

The team members for Group 20:

Vivek Agarwal ([vagarw14@asu.edu](mailto:vagarw14@asu.edu))

Suchita Vichare ([ssvichar@asu.edu](mailto:ssvichar@asu.edu))

Supriyaa Damodaraswamy ([sdamoda3@asu.edu](mailto:sdamoda3@asu.edu))

Sukanya Panda ([spanda7@asu.edu](mailto:spanda7@asu.edu))

Ayush Ray ([aray29@asu.edu](mailto:aray29@asu.edu))

# 3. Project Phase 1

The first phase of the project includes collection of data. The data collection is done by wearing the Myo sensor in the Right Hand and recording the data through an Android Phone. The Myorecode application helps to record the data from the Myo sensor. The data gets recorded in the Astro folder of the phone. The data for activities like eating, jumping, playing badminton, etc. have been recorded. The timestamps for the activities have also been recorded. The data is recorded with the help of myo sensors.

Myo sensors have the following hardware specifications:

SENSORS:

· Medical Grade Stainless Steel EMG sensors

· Highly sensitive nine-axis IMU containing three-axis gyroscope

· Three-axis accelerometer

· Three-axis magnetometer

PROCESSOR

* ARM Cortex M4 Processor

For the different activities, the sensor values are present in a column wise format which are transposed as with time values representing the columns and sensor values along the rows.

**Accelerometer:**

* The accelerometer sensor is used to detect movement and vibrations.
* The accelerometer measures the force along the x, y and z axes and has a unit of m/s2.

**Gyroscope:**

* The gyroscope is used to provide an additional dimension to the data of the accelerometer.
* Gyroscopes are used to measure or maintain rotational motion.
* The gyroscope was used to measure the rotation around the three different axes x, y and z, and has a unit of rad/sec.

**EMG (Electromyography) Sensor:**

* The EMG sensor is used to record the electrical activity produced by skeletal muscles, that is the electrophysical data.
* The data is measured along 8 different axis.(8-axis EMG)
* By detecting the electric potential generated by the muscle cells, the signals are recorded.

**Magnetometer(orientation):**

* The orientation along the axes x,y,z are recorded through the Azimuth angle (the angle around z-axis ), Pitch (angle around X-axis) and Roll (angle around z-axis), and is measured in degrees.
* The orientation is measured with respect to the Earth’s magnetic field, and measures the angle as the change in screen orientation.

After studying the data of the sensors, Feature Extraction and PCA on the data was performed to differentiate between different human activities.

# 4. Project Phase 2: Feature Extraction

In this phase, we have selected and implemented five feature extraction methods for the two activities chosen. Before carrying on with that we had to do some preprocessing of the data. The emg data collected had two duplicate timestamps, due to the collection of data in the left and right sensors as different tuples in the file. So now we have combined the data from both the sensors of the emg into a single tuple. The emg file now has 16 columns instead of 8 columns.

For feature extraction step, we have done sampling from which the features are extracted. The window for the emg data had been chosen as 400 rows and the window for all the other sensors’ data had been chosen as 100 for sampling. The sampling rate (every sensor has recorded the data with different frequencies) has been chosen such that all the sensors have equal number of data points. We had done this so that we have sufficient data samples in order to identify patterns while plotting both the activities.

The five feature extraction techniques that we have used are:

1. Mean
2. Root Mean Square(RMS)
3. Maximum
4. Variance
5. Variance of Fast Fourier Transform(FFT)

**Intuition behind feature selection**

Since we are extracting features for all the data we will initially have a final feature matrix with 29x5 columns (or features) in it. Since for now we have done this phase using the sample data provided by the professor the final object-feature matrix (after the sampling) has dimension of 227x145 which includes data for the both activities cumulatively. Initially, the data interpretation just by using numerical values seemed a difficult task. However careful exploration in the data gave us some understanding about the range of values assumed by certain features. Hence, we decided to get some insight into a few statistical measures, and finally made an educated guess to choose among them. We were completely aware of the fact that plotting of values for each activity along individual feature dimension would give more suitable ground to make a reliable judgement. The contribution made by each feature extraction method and how this feature are extracted is explained below.

## 4.1 Mean

The mean is also known as mathematical expectation or average. It is the central value for a discrete set of numbers. Mean was chosen as it is an appropriate option for the measure of central tendency of the data sample. It is basically the sum of all the values divided by the total number of values. We have used the MATLAB mean function to find the mean of all the 29 columns of sensor data. A scatter plot of the mean of both the activities was done to know how well they differentiate between the two activities. They were plotted on an overlap scatter plot. We are getting the most distinguished results for accelerometer and gyro sensor along the ‘z’ axis which is shown in the below graphs.



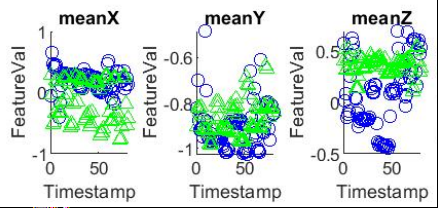


Fig4.1.1: Mean of the accelerometer data

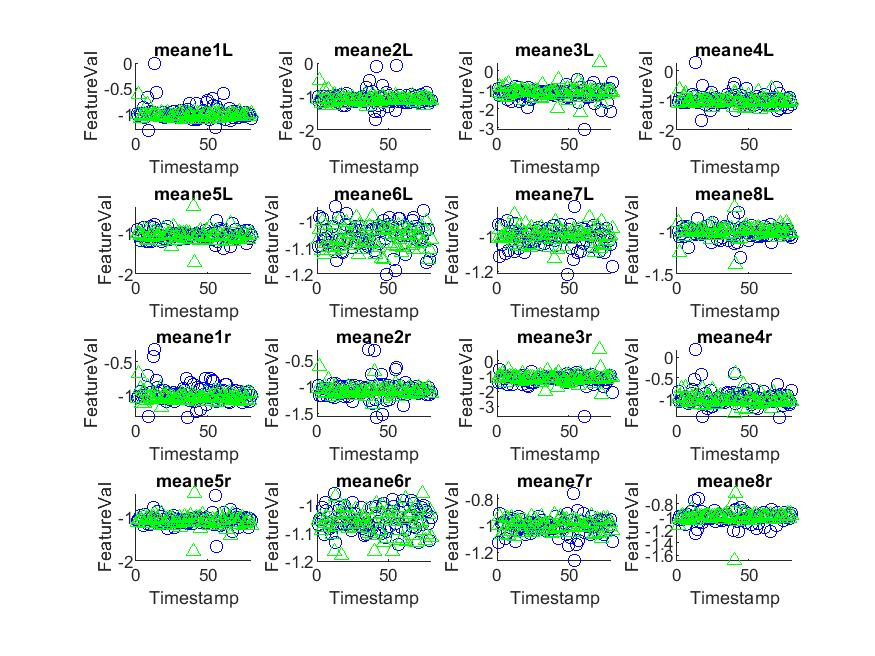


Fig4.1.2: Mean of the emg data

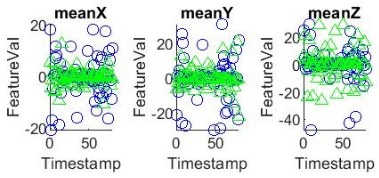


Fig4.1.3: Mean of the gyro data

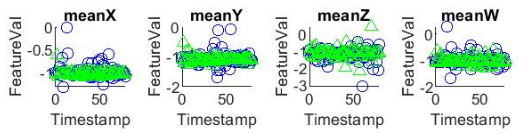


Fig4.1.4: Mean of the orientation data

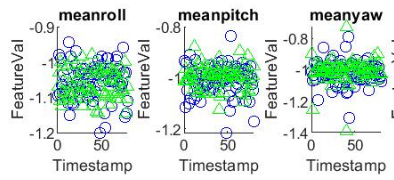


Fig4.1.5: Mean of the orientation\_Euler data

## 4.2 Root Mean Square

RMS is the root mean square of the values, ie., it is the square root of the arithmetic mean of the squares of a set of values. It is also called as the quadratic mean. RMS feature was used to get the sense of the magnitude of the data, as the data set contains both positive and negative values. We have used the MATLAB rms function to find the rms of all the 29 columns of sensor data. A scatter plot of the mean of both the activities was done to know how well they differentiate between the two activities. They were plotted on an overlap scatter plot. We get the most distinguished results for accelerometer sensor along ‘z’ axis and emg sensor along ‘emg8’ axis which is shown in the below graphs.



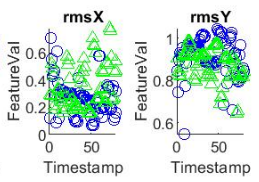
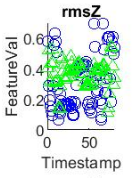
 

Fig4.2.1: RMS of accelerometer data



Fig4.2.2: RMS of emg data

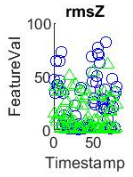
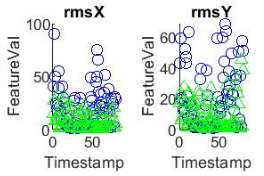


Fig4.2.3: RMS of gyro data

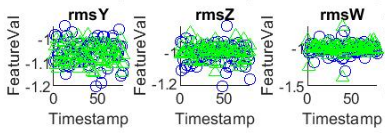
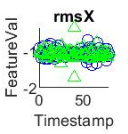


Fig4.2.4: RMS of orientation data

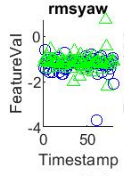
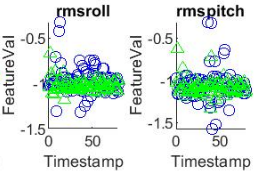


Fig4.2.5: RMS of orientation\_Euler data

## 4.3 Maximum

Maximum is the highest value in a set of values. We have used the MATLAB max function to find the maximum of all the 29 columns of sensor data. A scatter plot of the mean of both the activities was done to know how well they differentiate between the two activities. They were plotted on an overlap scatter plot. We are getting the most distinguished results for emg sensor along ‘emg8’ axis and accelerometer sensor along ‘z’ axis, which is shown in the below graphs.



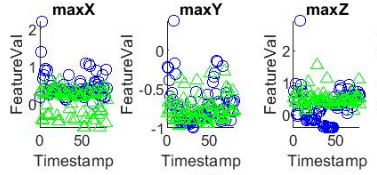


Fig4.3.1: Maximum of accelerometer data

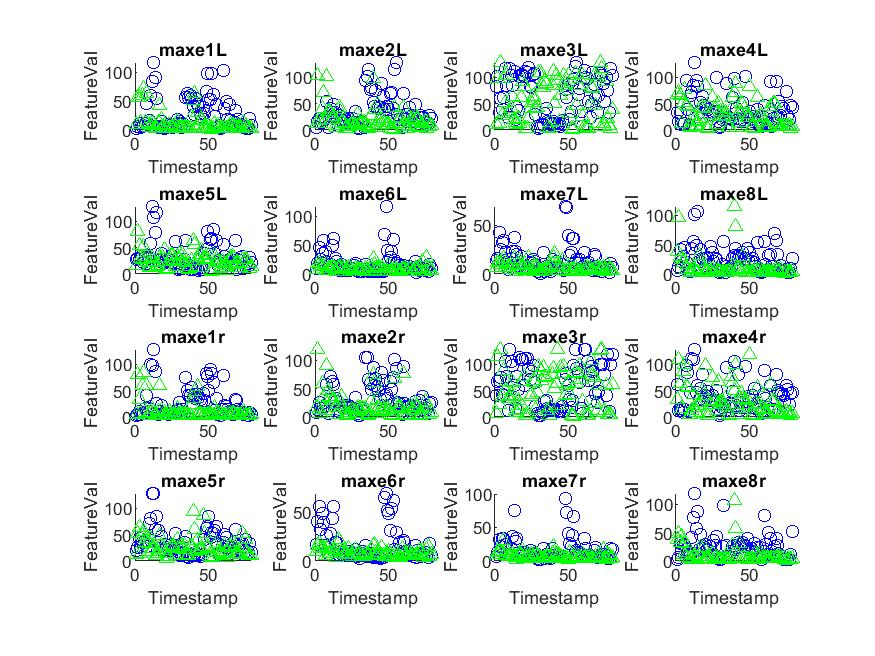


Fig4.3.2: Maximum of emg data

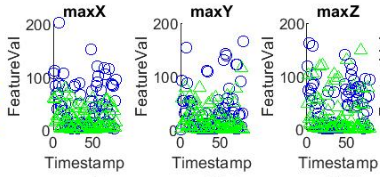


Fig4.3.3: Maximum of gyro data

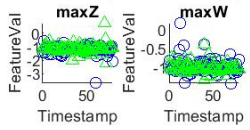
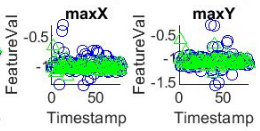


Fig4.3.4: Maximum of orientation data

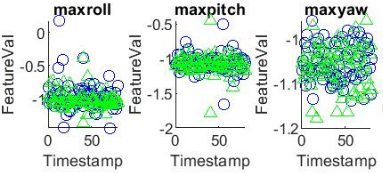


Fig 4.3.5: Maximum of orientation\_Euler data

## 4.4 Variance

Variance represents the spread of values of a data set. It is calculated by calculating the average of difference of a value from its mean. Variance is used because it gives us an idea of how spreaded the data set is. We have used the MATLAB var function to find the variance of all the 29 columns of sensor data. A scatter plot of the mean of both the activities was done to know how well they differentiate between the two activities. They were plotted on an overlap scatter plot. We initially believed that variance could be a potential feature to get different results, but none of the sensors showed considerably different results for mean variance.



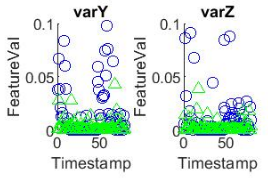
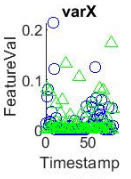


Fig4.4.1: Variance of accelerometer data

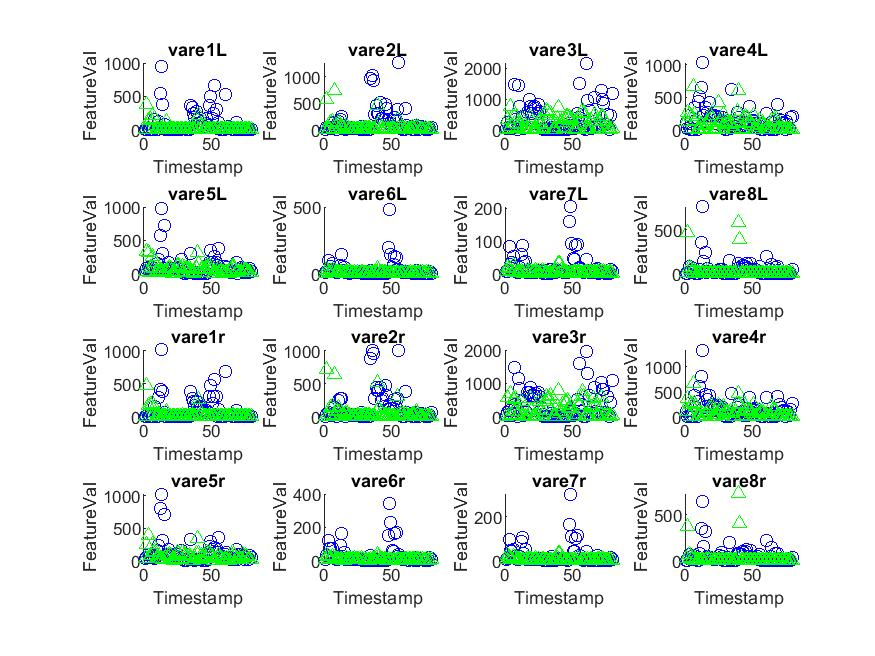


Fig4.4.2: Variance from emg data

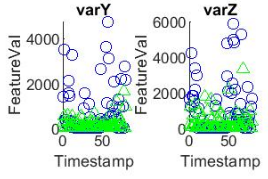
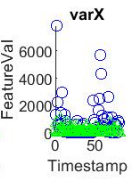


Fig4.4.3: Variance of gyro data

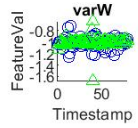
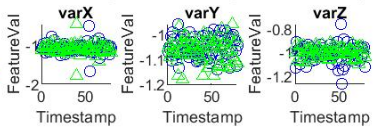


Fig4.4.4: Variance of orientation data

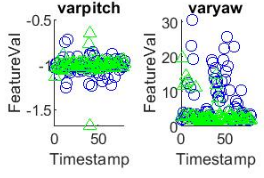
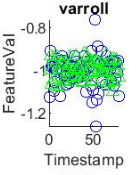


Fig4.4.5: Variance of orientation\_Euler data

## 4.5 Variance of Fast Fourier Transform

Fast fourier transform is the feature extraction technique that samples a signal over a period of time and divides it into its frequency components. Here the variance of fft is used as it gives more distinction between the two activities. FFT feature was used to segregate the data based on its frequency. We have used the MATLAB FFT function to find the Fast Fourier Transform of all the 29 columns of sensor data. A scatter plot of the mean of both the activities was done to know how well they differentiate between the two activities. They were plotted on an overlap scatter plot. We get the most distinguished results for emg sensor along ‘emg7’ axis and orientation sensor along ‘x’ axis and euler angle ‘pitch’, which is shown in the below graphs.



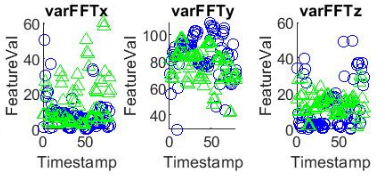


Fig4.5.1: Variance of FFT of accelerometer data

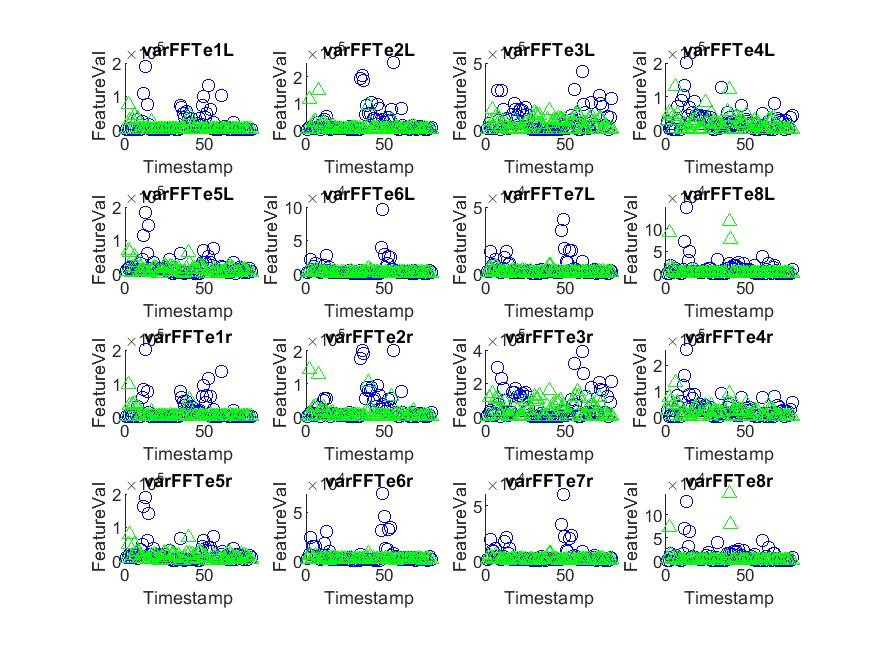


Fig4.5.2: Variance of FFT of emg data

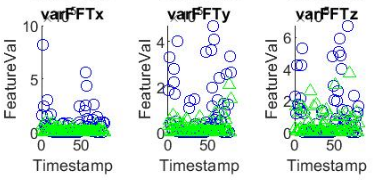


Fig4.5.3: Variance of FFT of gyro data

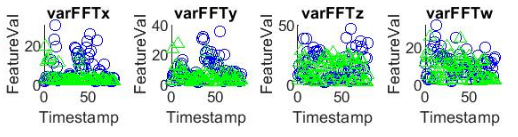


Fig4.5.4: Variance of FFT of orientation data

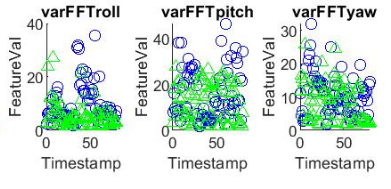


Fig4.5.5: Variance of FFT of orientation\_Euler data

## 4.6 Intuition about each feature and their outcomes:

We performed this five different feature extraction techniques namely: 1) Mean, 2) Root mean square(RMS), 3) Maximum, 4)Variance, 5)Variance of Fast Fourier Transform(FFT). Our intuition was that all these features would provide a proper distinction between various activities. And as we observe all the outputs we can conclude that we have maximum distinction in Mean and Variance of Fast Fourier Transform(FFT) out of all the feature extraction techniques.

# 5. Project Phase 3: Feature Selection

**5.1. Subtask 1: Arranging the feature matrix**

Principal Component Analysis (PCA) takes only one matrix, so, we merge the results obtained in Phase 2 in a single matrix. Hence, the feature matrix will have 29 x 5 features corresponding to each action and the rows corresponding to the sampled timestamps in Phase 2. We were able to find few useful features having higher discrimination power by plotting graphs against every feature for the selected two activities. Hence, we decided to proceed further with selected features so that we can perform PCA on the resulting matrix to find best latent semantics which have the highest discrimination power, even among the ones selected during Phase 2 feature selection process. We have performed PCA on each activity individually first and then applied PCA on using the data for the two activities simultaneously.

**5.2. Subtask 2 : Execution of PCA**

PCA decomposes a correlation matrix into a matrix with Principal Components and the resulting matrix contains the Principal Components in decreasing order of their variance.

We pass the matrix obtained in Subtask 1 - 160 rows (80 rows corresponding to each activity) and 10 columns (selected features) - to PCA function of MATLAB.

PCA returns the following:

Coeff - a 10 x 10 matrix, representing the coefficients for Principal Components a.k.a Eigenvectors and the columns are in decreasing order of their variance (or eigenvalues).

Score - Principal Component scores are the transformed representation of the input matrix in the Principal Component Space.

Explained - The percentage of variance depicted by each Principal Component.

Latent - The Principal Variances i.e. eigenvalues of the input matrix.

The code for PCA has been included in Assignment1\_Group20.m file. We have used PCA function of MATLAB to perform PCA.

**5.3. Subtask 3: Make sense of the eigenvectors**

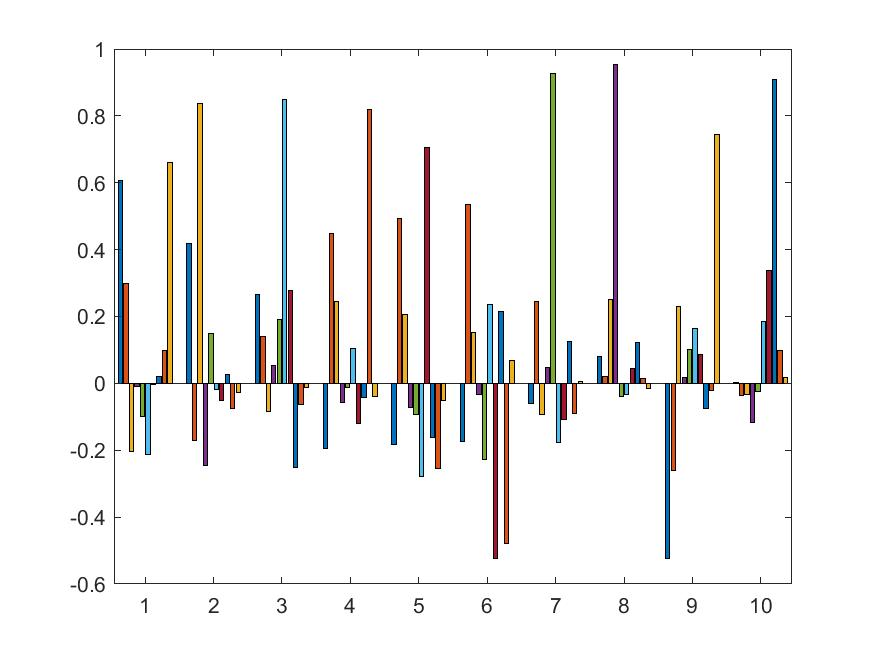
Finding eigenvectors and eigenvalues of the covariance matrix basically means fitting the principal components along the variance of the data.

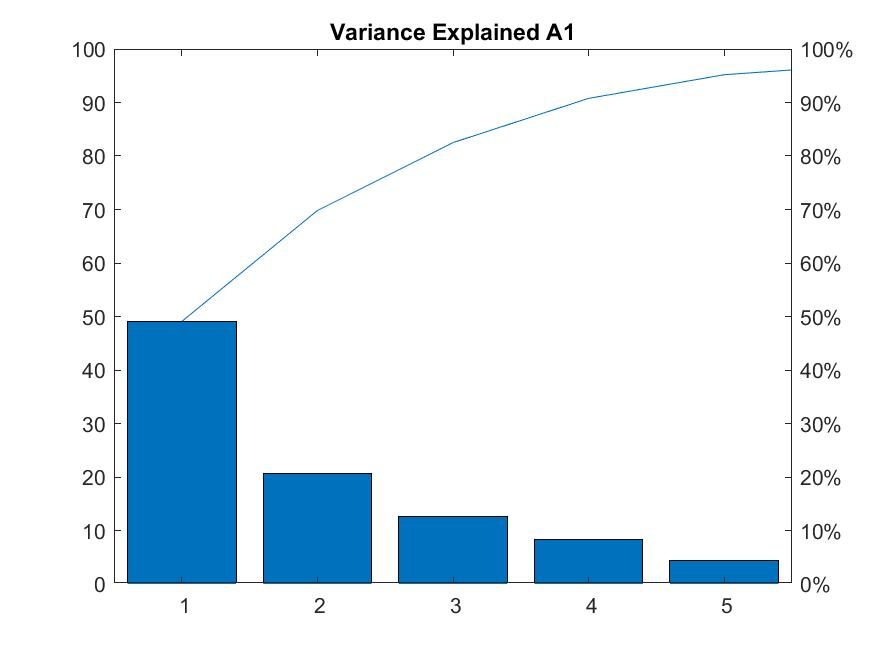
The ‘coeff’ matrix returned by the PCA function, denotes the eigenvectors i.e the vectors denoting the Principal Components with decreasing variance (i.e. Principal Components with the most variance first and the least variance last). The features along with the data is more spread out have higher weightage for their components in eigenvectors along those directions.

The eigenvectors with highest variance are most representative of the data while eigenvectors with lowest variance can be chosen to be dropped depending on the application.

We have plotted the Eigenvectors below:

1. **EigenVectors after PCA performed on Activity 1 (Cooking 1 folder)**

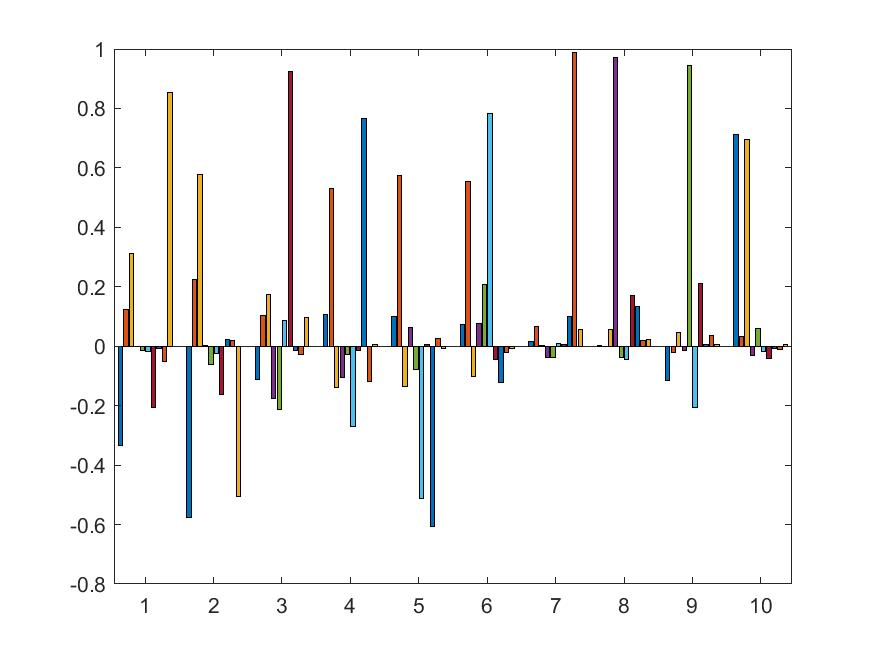


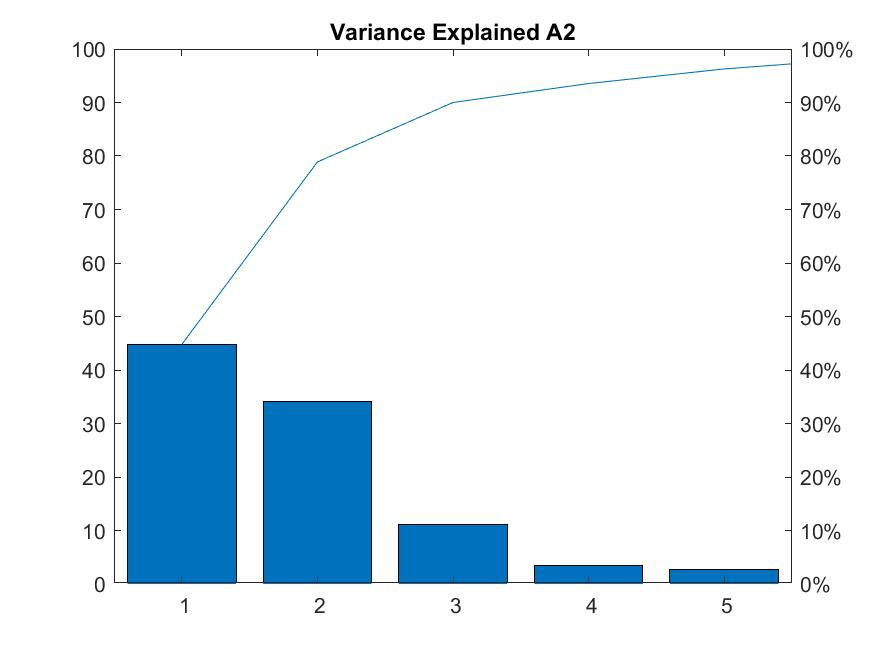


We tried to figure out how many vectors in this latent space would essentially be required to represent the data capturing most of the variance. For this we have plotted the percentage of variance attributed to each of the eigenvector. This is also known as Variance Explained by the eigenvectors.

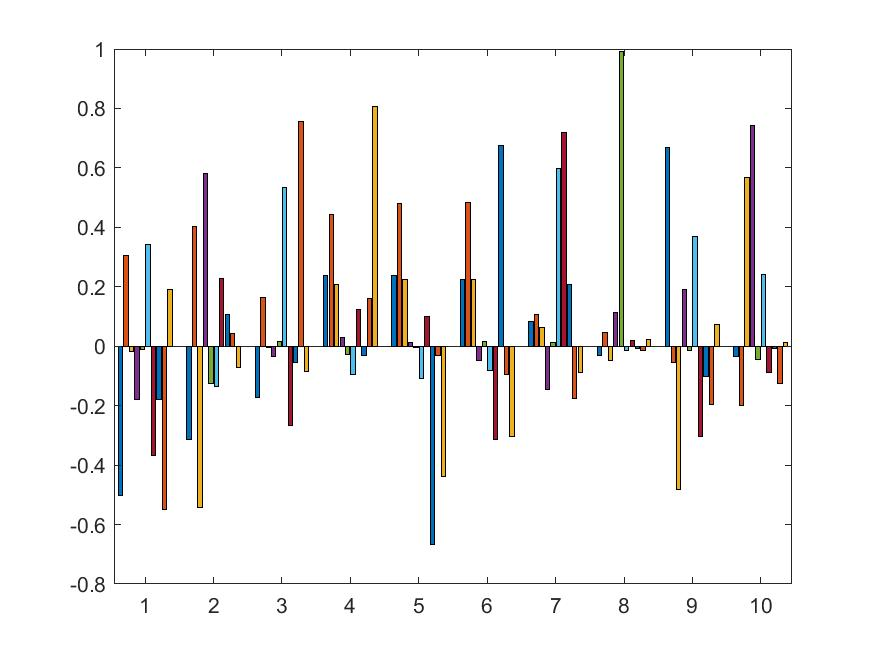
Through the above plots we can construe that first five eigenvectors are good enough to represent most (95%) of the information. The First Principal Component alone contributes 50% of the information. The bar graph plot of eigenvectors represents the individual weights of original features in each eigenvector. Every eigenvector has a unique color in the bar plot and each bar represents the weights (components) along original features. The higher value of the weight or component signifies that most of the data lies along that feature component of the eigenvector. In principal component 1, first three values have higher values which means these corresponding features are reasonable enough to represent the data along this principal component. In this case PC1 has most prominent features like Mean, Rms and Max values along ‘Z’ axis of Accelerometer sensor and orientation sensor along ‘x’ axis.

1. **EigenVectors after PCA performed on Activity 2 (EatFood 1 folder)**





Similar to Activity 1, only 5 EigenVectors are reasonable enough to represent 95% of the information with PC1 contributing about 45%. In this case PC1 has most prominent features like Variance of FFT of Euler Orientation along Pitch axis and Max values along ‘Z’ axis of Accelerometer sensor.

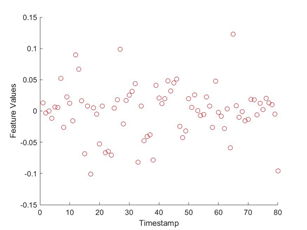
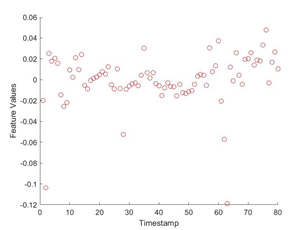
1. **EigenVectors after PCA performed on data for both above activities together**

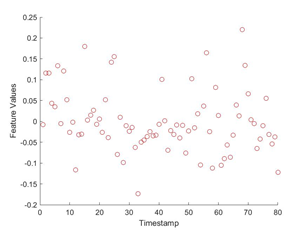
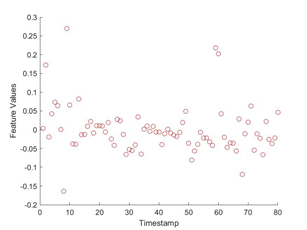
Again, we were able to find most representative features along each Principal components which five Eigenvectors contributing to 95% of the information. It will not surprise us that in PC1 most contributing components were along features - Orientation sensor along ‘x’ axis Accelerometer, mean and max along Z axis. This make sense as movement during eating is mostly along z-axis while cooking has more horizontal hand motion in x-axis.

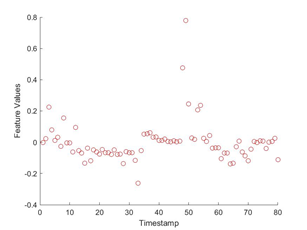
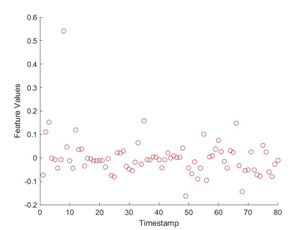
**5.4. Subtask 4: Results of PCA**

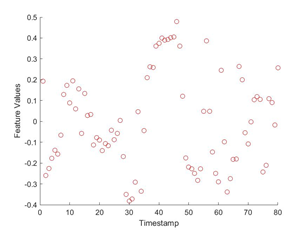
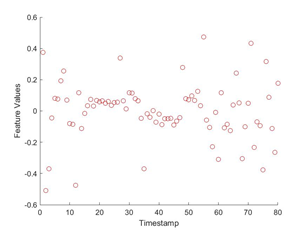
We have plotted the score - the parameter showing the data in transformed space using principal components - for cooking and eating actions and plotted them. The below plots shows how each Principal component vectors represents data when data is projected along these vectors individually.

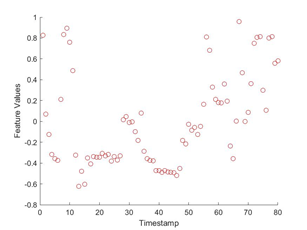
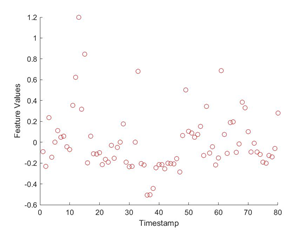
**PCA results for the Cooking Activity:**



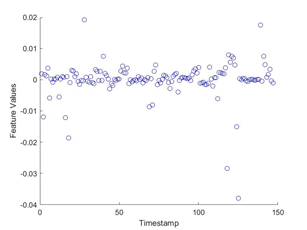
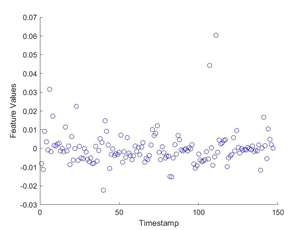
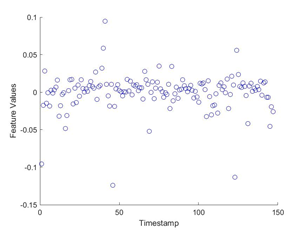
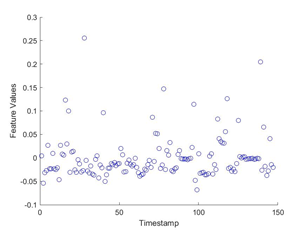
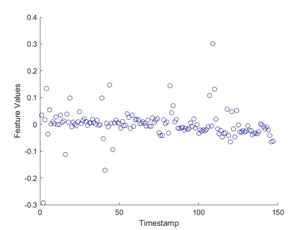
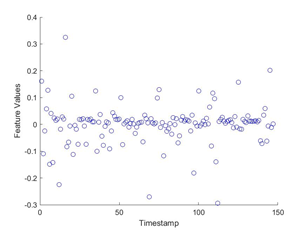
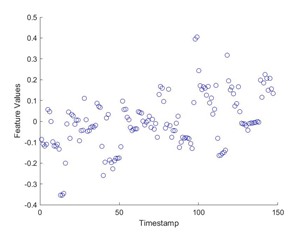
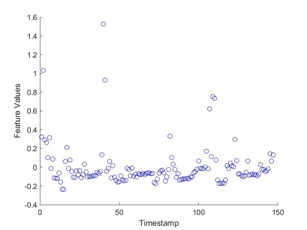
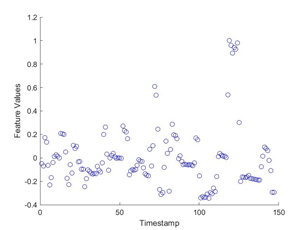






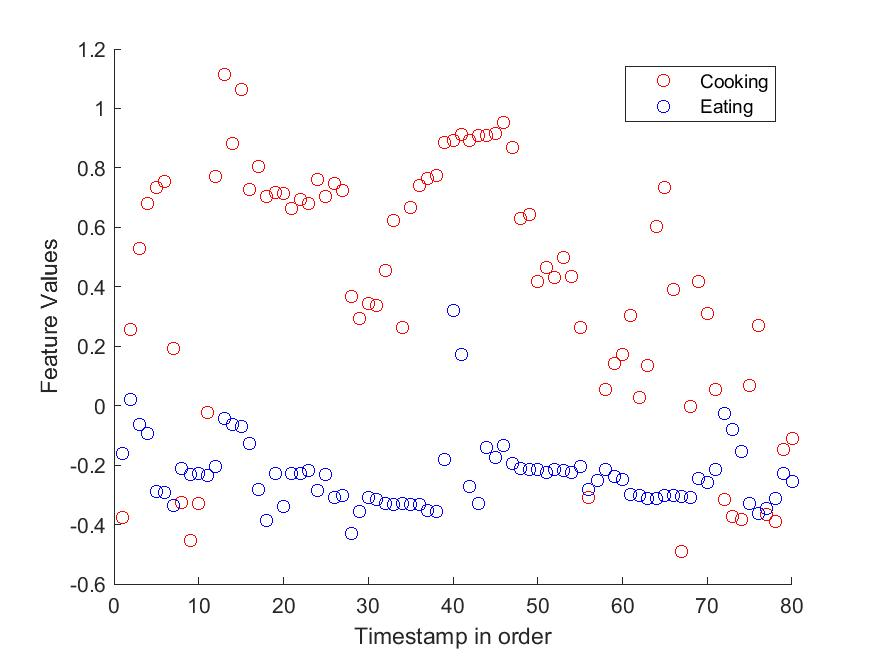


**PCA for the Eating food Activity:**

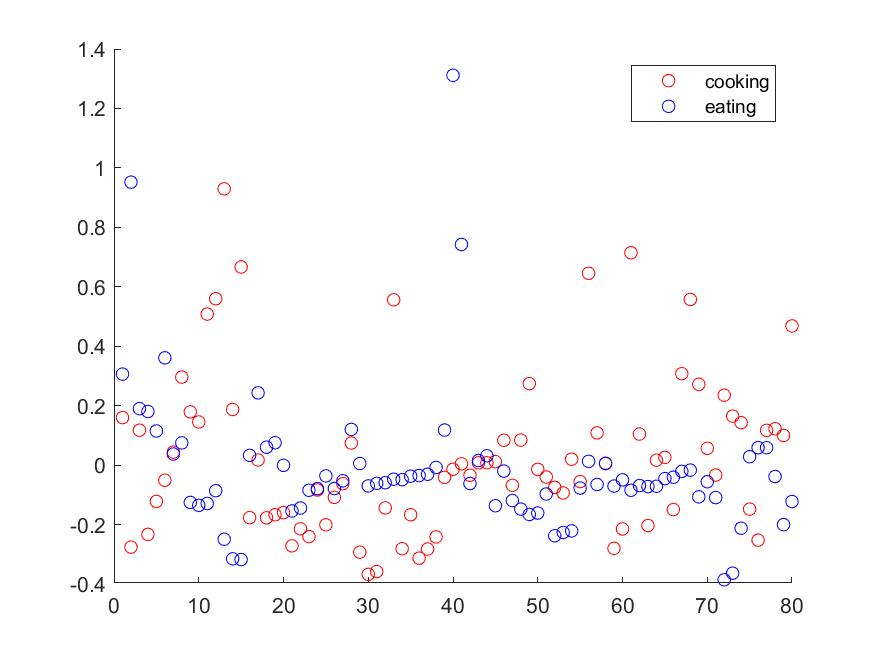
****

**PCA results for both Eating and Cooking activities:**

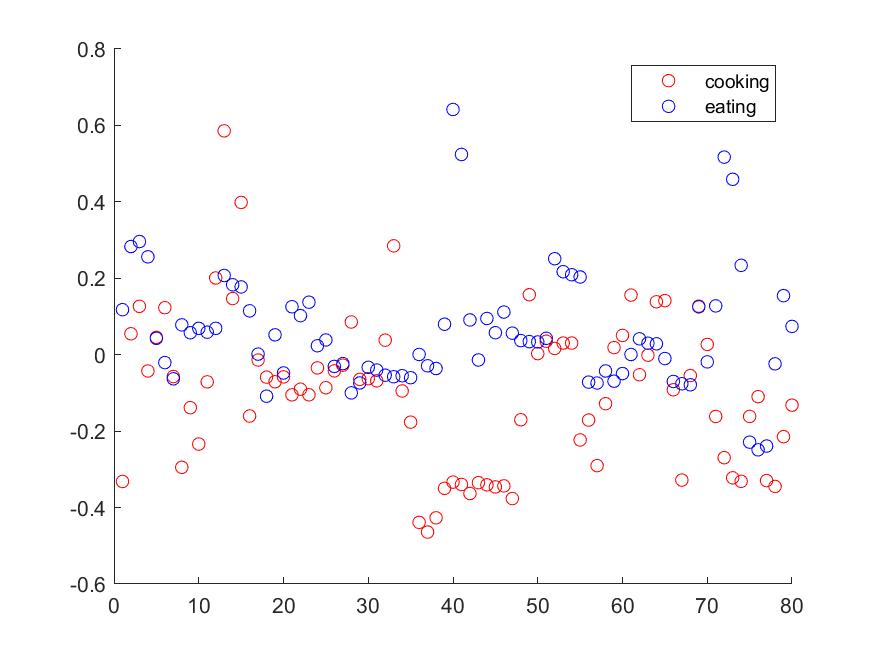
The graphs allow us to understand the degree of variance for the two activities and how they differ. The eigenvectors contributing to higher variance has in fact helped us to segregate the data points corresponding activities in our case.



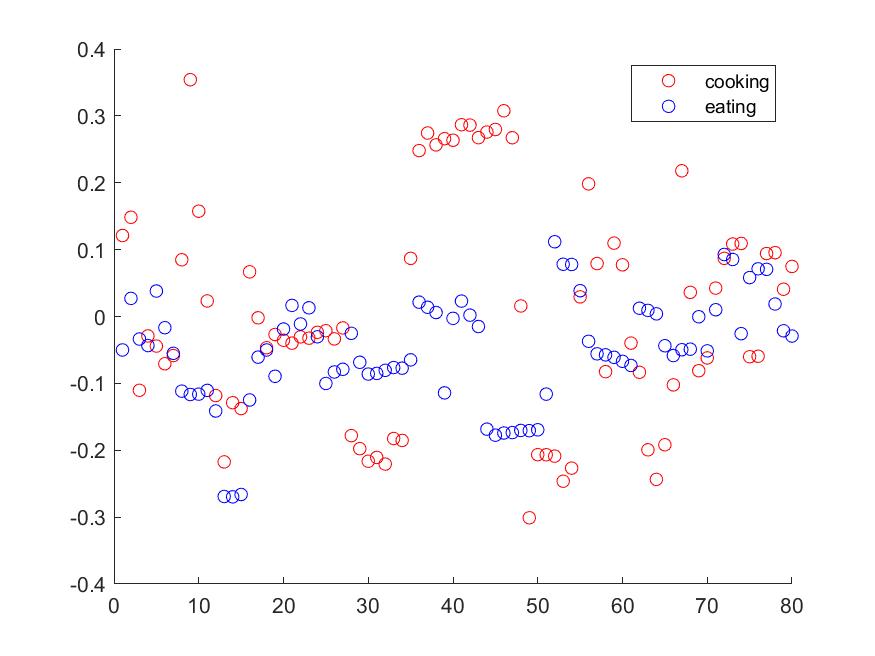
Plot\_Feature\_1



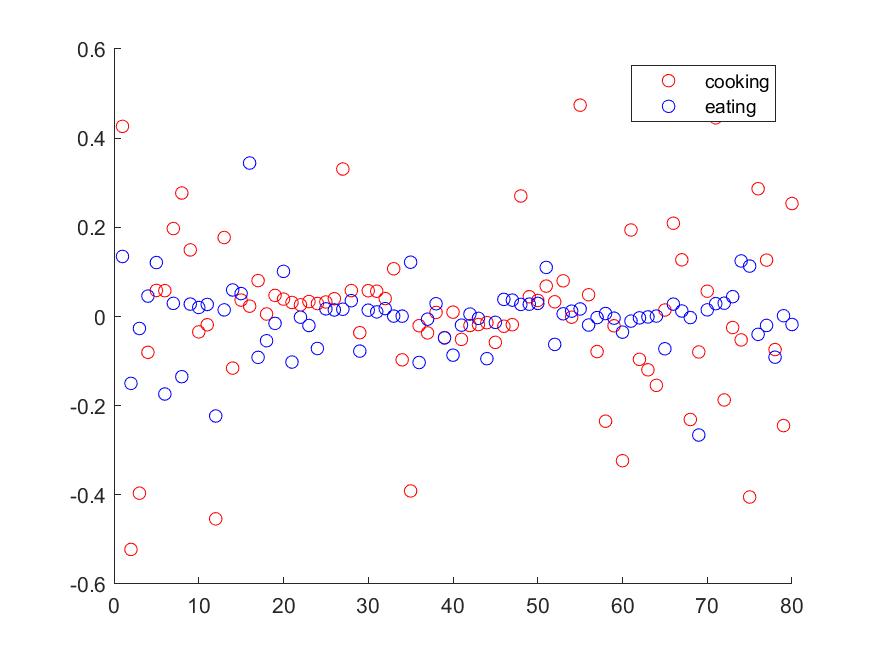
Plot\_Feature\_2



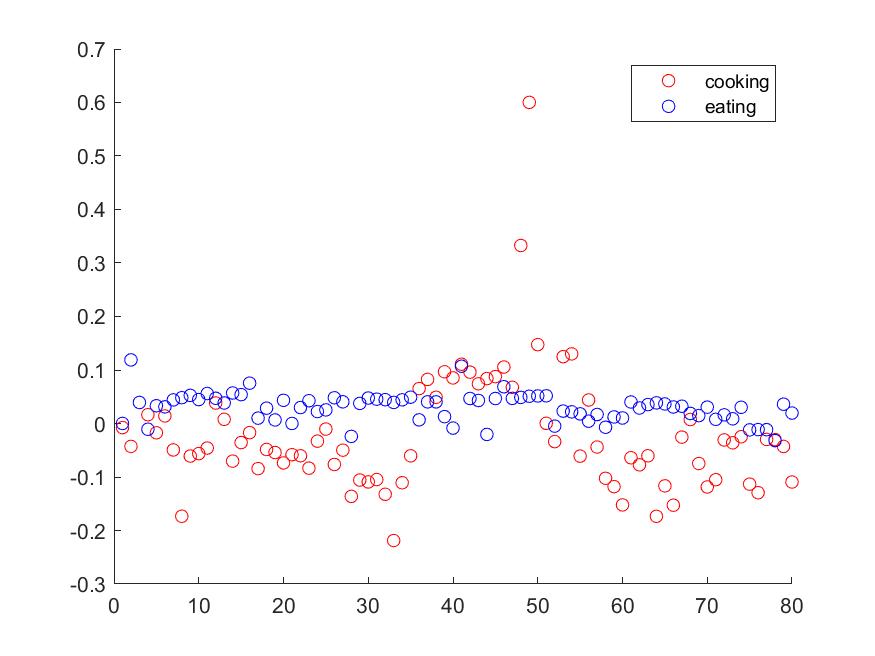
Plot\_Feature\_3



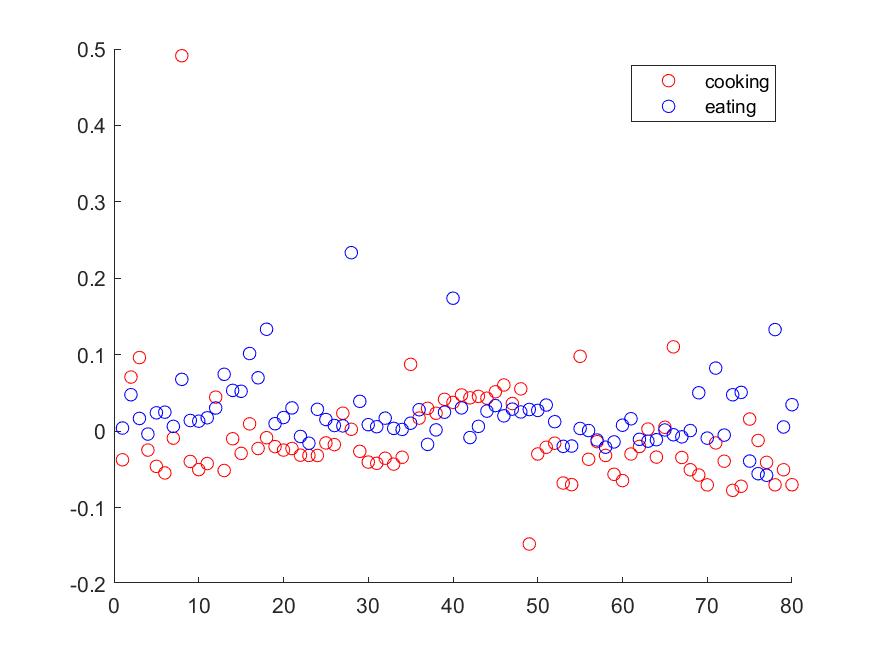
Plot\_Feature\_4



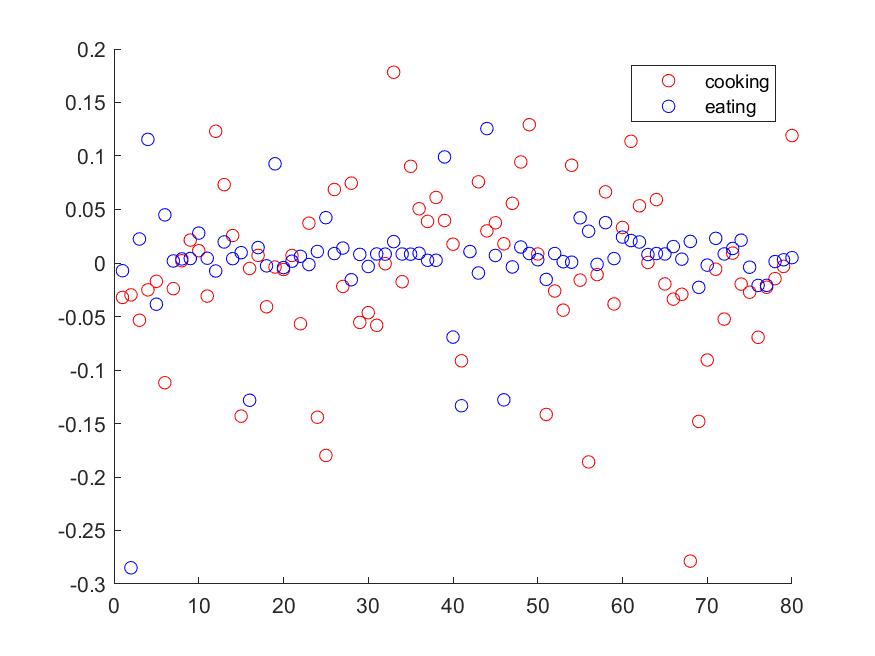
Plot\_Feature\_5



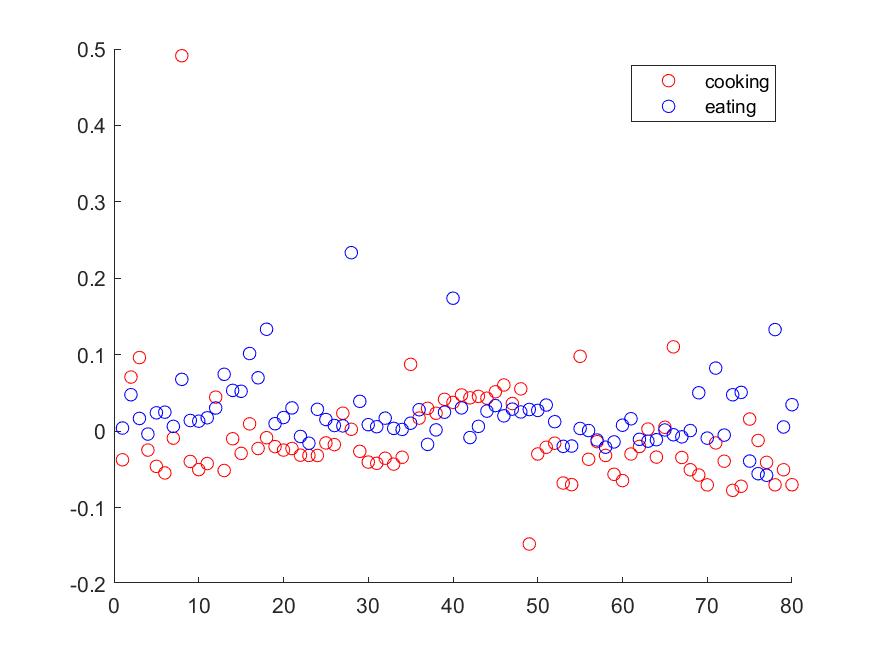
Plot\_Feature\_6



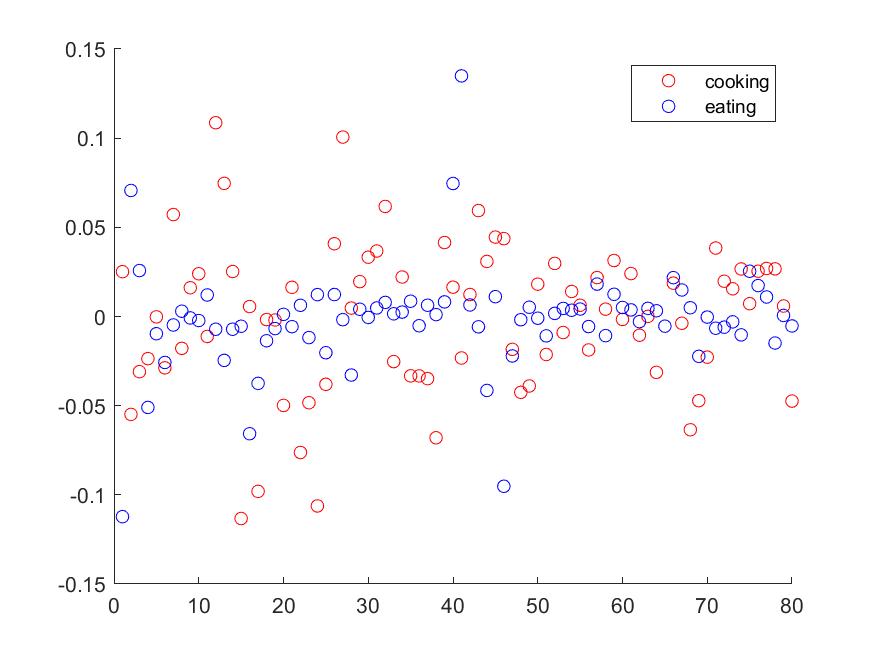
Plot\_Feature\_7



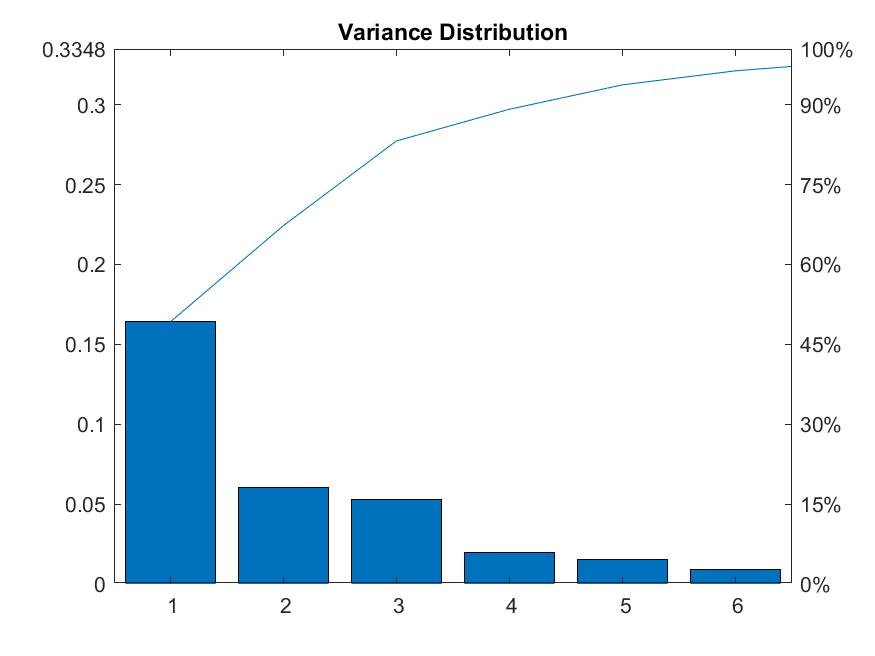
Plot\_Feature\_8

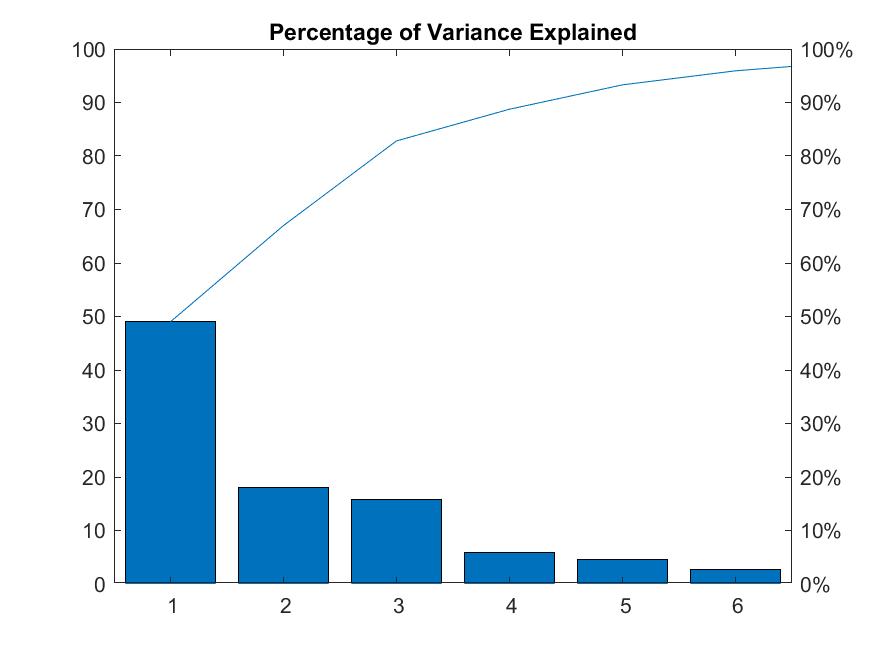


Plot\_Feature\_9



Plot\_Feature\_10





It is quite evident the first eigenvector has displayed a major contribution in representing about ~50% of information. We can also conclude that first five eigenvectors can represent more than 90% of the information. Thus, reducing the feature space to 5 or 6 latent features would help in representing data in this transformed space.

**5.5 Subtask 5: Was doing PCA helpful?**

Yes, doing PCA was helpful as it helped us reduce dimension of the data and helped us make sense of what features are really helpful in distinguishing the two activities with less dimensional data.

We believe that our intuitions that we had assumed in task two were also correct to a great extent.

**References :**

1. **Accelerometer and rate gyroscope measurement of kinematics: an inexpensive alternative to optical motion analysis systems**

**Ruth E.Mayagoitia, Anand V.Nene, Peter H.Veltink**

1. **Techniques of EMG signal analysis: detection, processing, classification and applications**

**M. B. I. Reaz, M. S. Hussain and F. Mohd-Yasin**

1. **ETHOS: Miniature orientation sensor for wearable human motion analysis**

**Holger Harms, Oliver Amft, Rene Winkler, Johannes Schumm, Martin Kusserow, Gerhard Troester**

1. **“X.” Principal Component Analysis of Raw Data - MATLAB, www.mathworks.com/help/stats/pca.html.**