Assignment 1 Report

CSE 572 Data Mining

Fall 2018

Submitted to:

Professor Ayan Banerjee

Ira A. Fulton Schools of Engineering

Arizona State University

Submitted by:

Ashm Walia

Harshit Laddha

Manav Bagai

Raj Buddhadev

Shubham Mittal

(Group - 2)

Table of Contents

1. Introduction

The aim of this project is to develop a computing system that can extract the most prominent features of sensor data collected while performing various human activities. The process involves collection of the data while performing activities such as eating, cooking, cycling, sleeping, etc., performing five aggregation tasks on this collected data and finally extracting the most variating features using PCA and plotting these features. The extracted data and corresponding features can be used by classifiers such as decision trees to predict different activities.

2. Team Members

- Ashm Walia - awalia6@asu.edu
- Harshit Laddha - hladdha@asu.edu
- Manav Bagai - mbagai@asu.edu
- Raj Buddhadev - rbuddhad@asu.edu
- Shubham Mittal - smitta21@asu.edu

3. Phase 1 - Data Collection

The first phase of this project consisted of collecting data of two routine human activities using Myo band. We have collected the data for the activities of eating and cycling. However, for this assignment, we are using the data provided by Professor Ayan Banerjee for the activities of **eating, cooking** and **no movement**.

Member	Activity	Date	Time
Ashm Walia	Eating	9/17/2018	8:14 pm to 8:22 pm
Ashm Walia	Cycling	9/18/2018	8:56 am to 9:05 am
Ashm Walia	Cycling	9/18/2018	12:15 pm to 12:20 pm
Manav Bagai	Cycling	9/18/2018	5:51 pm to 6:09 pm
Manav Bagai	Eating	9/18/2018	7:05 pm to 7:40 pm

4. Phase 2 - Feature Extraction

The aim of this phase was to apply five features extraction techniques on the data available for the three activities chosen. For this, we combined all the data available into a single table data structure of MATLAB. We have also added two columns storing the class of these data points i.e the activity name.

Feature Extraction Techniques that we have applied in this task are:

1. Fast Fourier Transform(FFT)

Fast fourier transform is the feature extraction technique that samples a signal over a period of time and divides it into its frequency components.
In our implementation of FFT, we have used MATLAB function of fft on the data.

Justification: Ours is a temporal data, varying over time. We are using FFT because it gives the best option for generating frequencies related to data points, which will help us differentiate between activities in the future.
We select top six frequencies for every attribute of the sensors.

2. Discrete Wavelength Transform(DWT)

Discrete Wavelength Transform is any wavelet transform for which the wavelets are discretely sampled. As with fft, it captures frequency information. We are using MATLAB dwt function on the data to get discrete wavelength transformation.

Justification: Use-case of DFT in our project is to get the most interesting frequencies in terms of variation of data and hence we select top six frequencies for every attribute of the sensors.

3. Standard Deviation

Standard deviation is used to quantify the amount of variation or dispersion in the entire dataset. We are using the MATLAB function of std for getting the standard deviation of the data.

Justification: Standard deviation is used mainly because it helps in knowing the deviation and variance in the data, which is very helpful aspect while doing the PCA.

4. Z-score

According to wikipedia, Z Score is signed number of standard deviations a datapoint is from mean.

$$z = \frac{(x-\bar{x})}{\sigma}$$

where x is the data point, $\bar{x}$ is the mean of the data and σ is the standard deviation of the data.

Justification: For example, a z-score of 1 is 1 standard deviation above the mean. The main purpose of this technique is that it makes the data in a way that makes the mean 0 with standard deviation of 1.

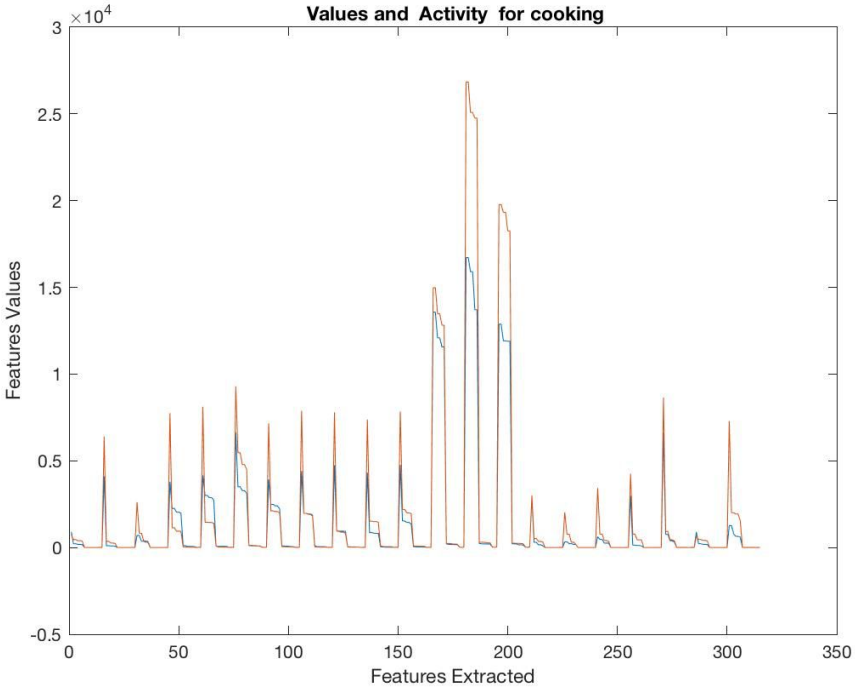
5. Root Mean Square

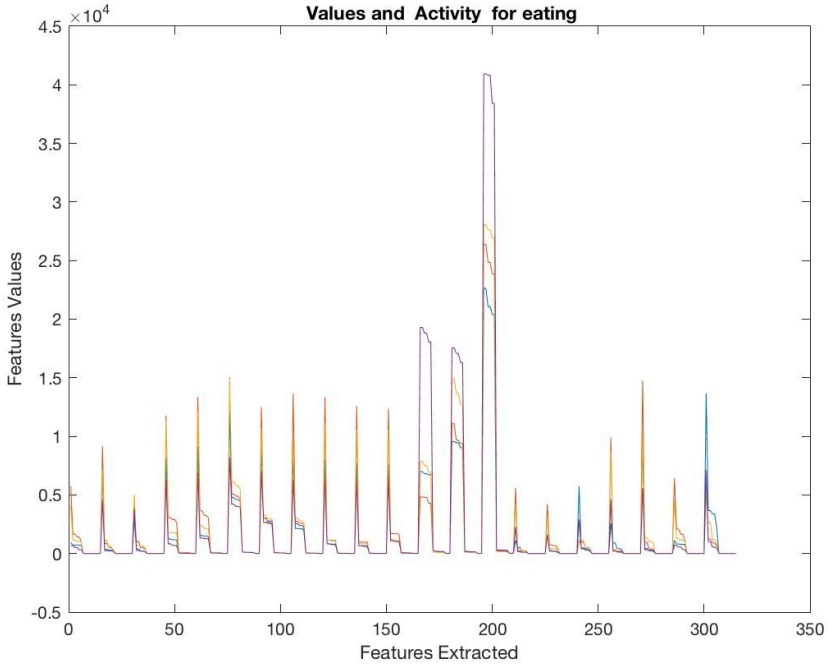
Root mean square means square root of the arithmetic mean of the squares of the set of values.

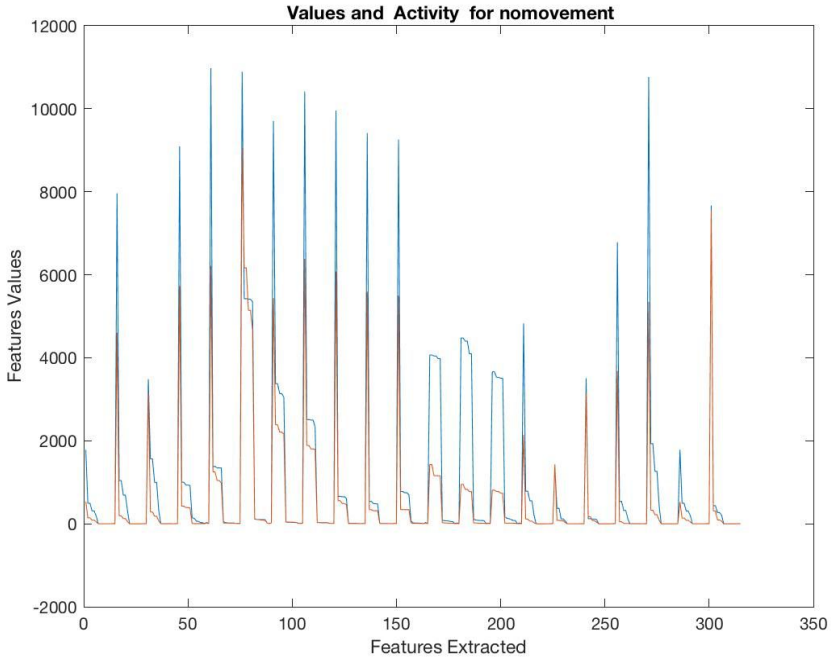
$$RMS = \sqrt{\frac{x_1^2+x_2^2+....+x_n^2}{n}}$$

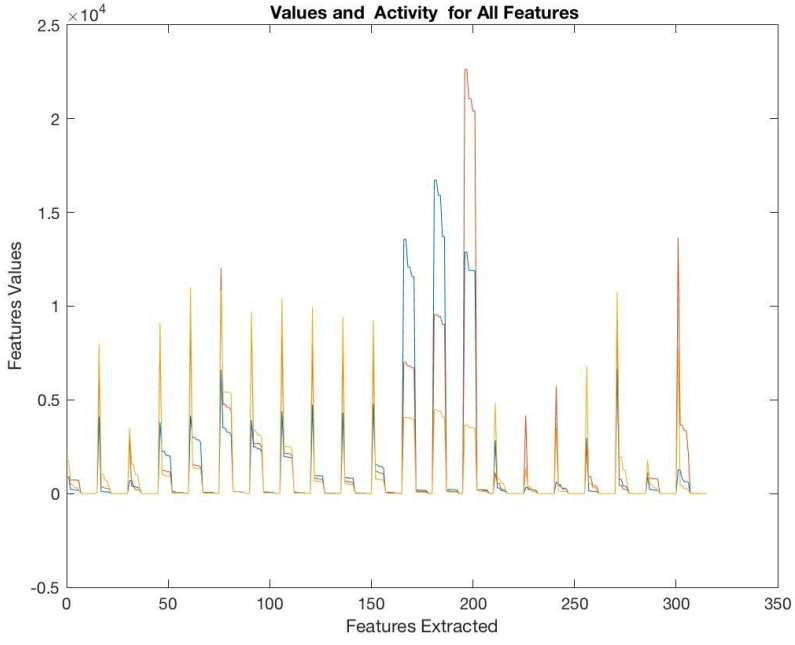
Justification: We use RMS in our project to be able to compare the squared means of different sensor data so that it could make sense as to which sensor data is actually meaning full.

Below are the plots for the three activities after application of the five feature extraction measures:






Values and Activity for nomovement
Features Values
Features Extracted
12000
10000
8000
6000
4000
2000
0
-2000
0
50
100
150
200
250
300
350


Values and Activity for All Features
×10^4
Features Values
Features Extracted
2.5
2
1.5
1
0.5
0
-0.5
0
50
100
150
200
250
300
350

Out of all the features extraction techniques we have used, FFT produces the best frequency domains by transforming a continuous stream of data. Because in this phase we combined the data of all the activities in a single table. Use of FFT reduces our work significantly, which makes enormous difference in time required to transform such large sequence of data in a single table.

5. Phase 3 - Feature Selection using PCA

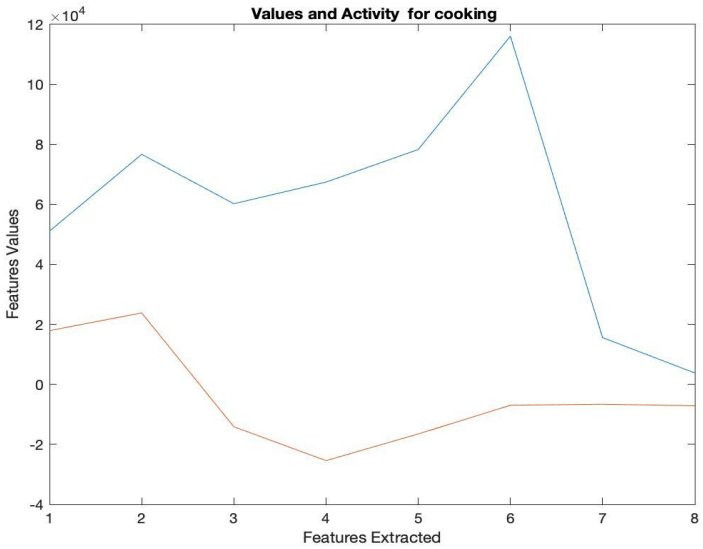
The goal of Feature Selection is to reduce the features extracted in the previous phase, so that the resulting features can be efficiently used by classifiers such as Decision Trees, Rule Based Classifier etc.

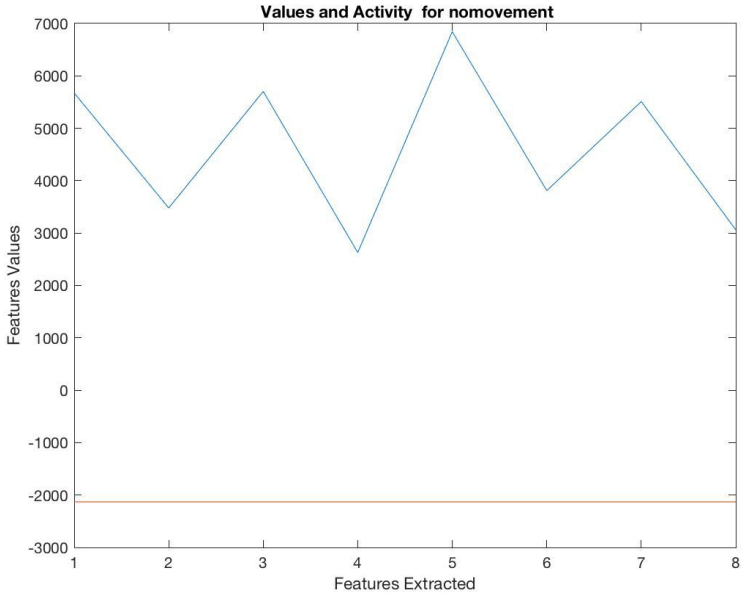
Principal Component Analysis: PCA is a Dimensionality Reduction technique that brings out the features that vary the most. We apply PCA to our dataset to bring out the most sensible and variating attributes. PCA treats the dimension with large variance as important.

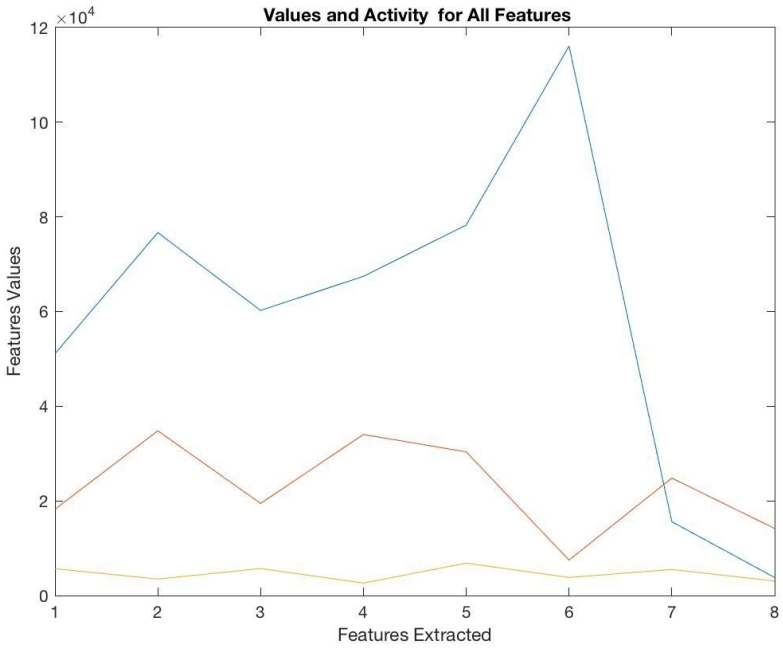
Steps to Find PCA of the data:

1. First step is to find Covariance matrix 'R' of the attributes. Let's say there are k attributes in the data. So a covariance matrix of k*k dimension is calculated.
2. Second step is to find eigenvalues of the Covariance Matrix 'R'.
$$R - \lambda I = 0$$
3. Third step is to find eigenvectors corresponding to each eigenvalue.
$$R\overline{v} = \lambda\overline{v}$$
We represent all the eigenvectors in a matrix 'V' of k*k dimensions in decreasing order of eigenvalues. This matrix V is orthogonal i.e. $V * V^T = I$. Also, a Principle Matrix S is created where, every diagonal element is the eigenvalue corresponding to the eigenvector in V.
4. Fourth step is to represent the PCA in the form
$$R' = V * S * V^T$$
We reduce the dimension by setting lower valued eigenvalues as 0. The new R' has variance preserved.

Explanation: After the feature extraction, we get a 8 x 315 matrix where each of the 8 rows represent an activity (Cooking1, Cooking2, Eat1, Eat2, Eat3, Eat4, and NoMovement1, NoMovement2) and the 315 columns represent features extracted for each activity. We apply PCA to this matrix, which contains data for all the sensors and get 315 eigenvalues. Further we plot the top 5 eigenvalues for each Activity as shown below.

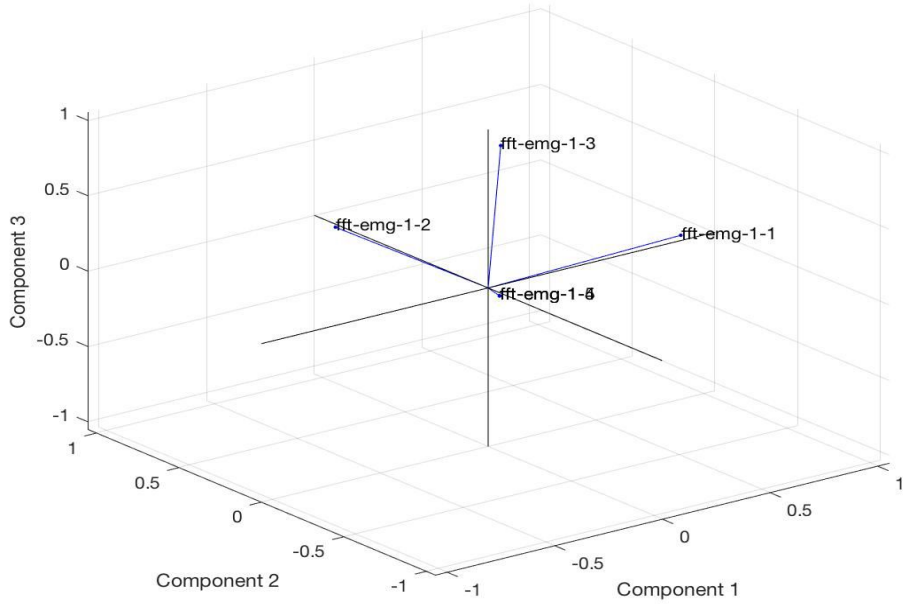


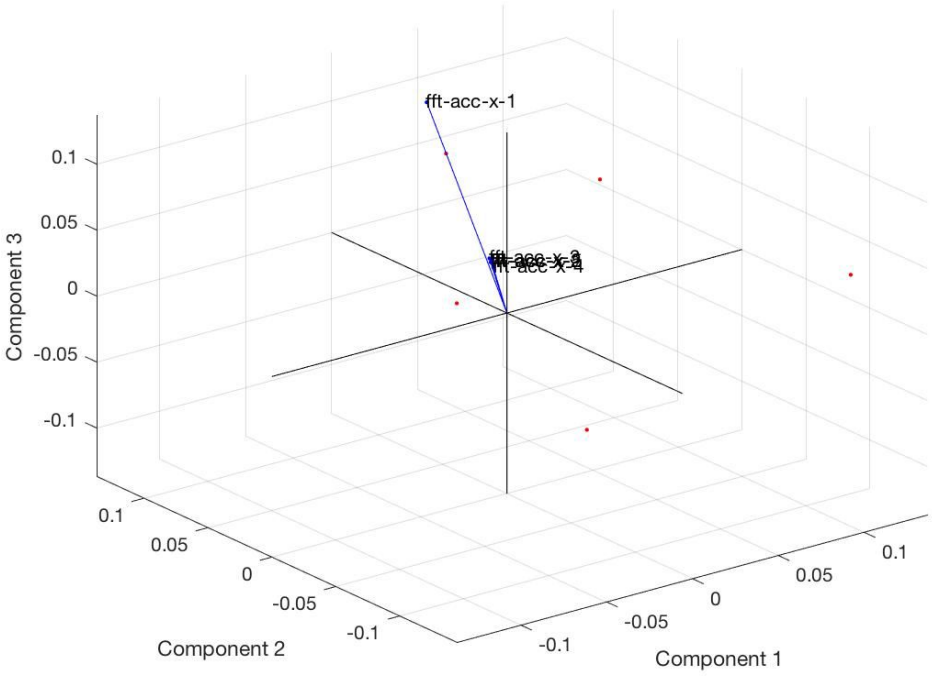


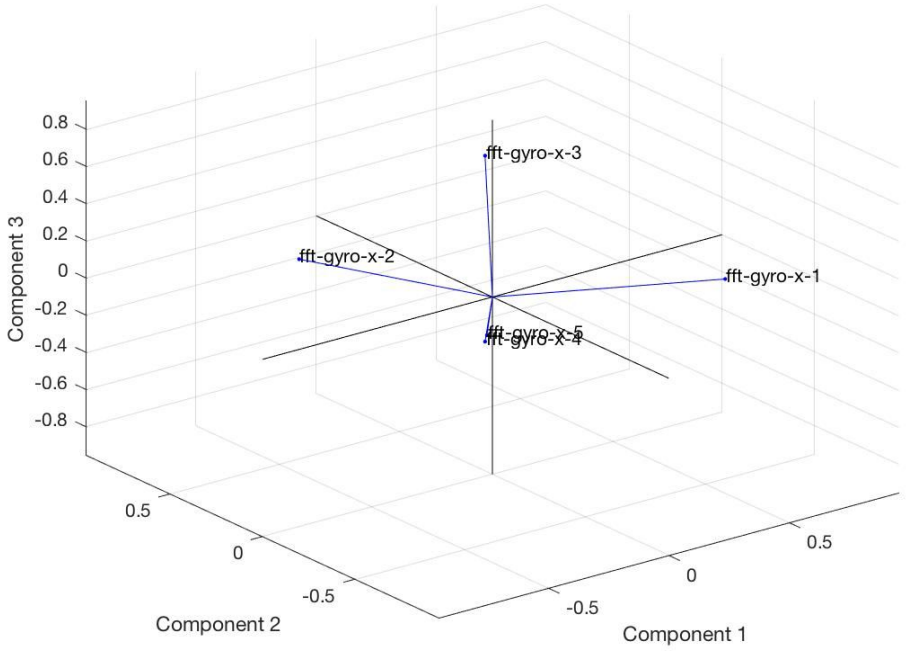


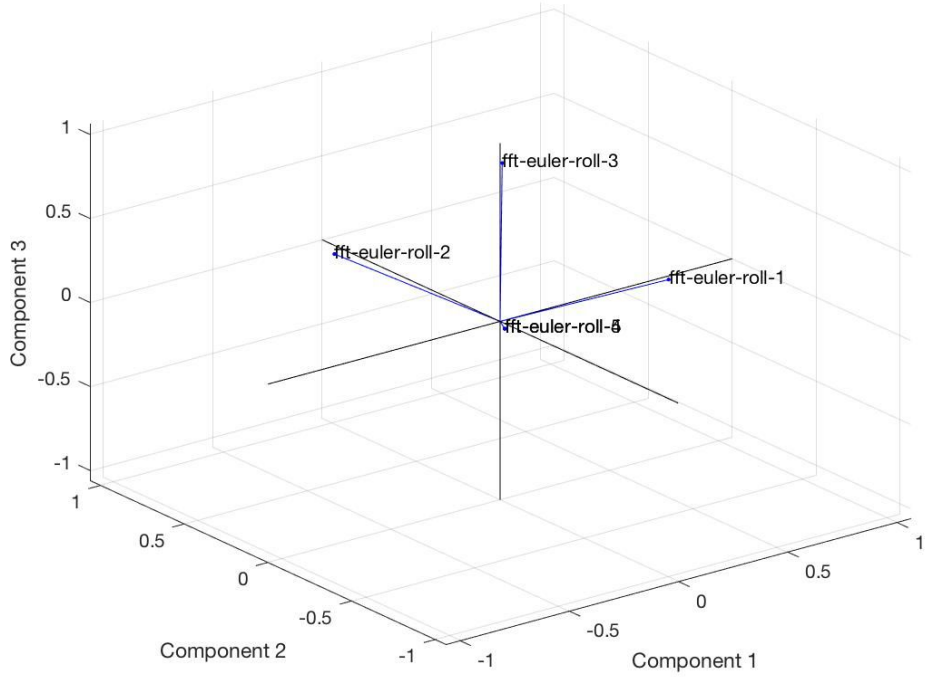
Further, we plot graphs that show transformation of these top 5 features for each sensor used in the 3 dimensional space. This gives us a good sense of how these features vary with respect to the sensor data.

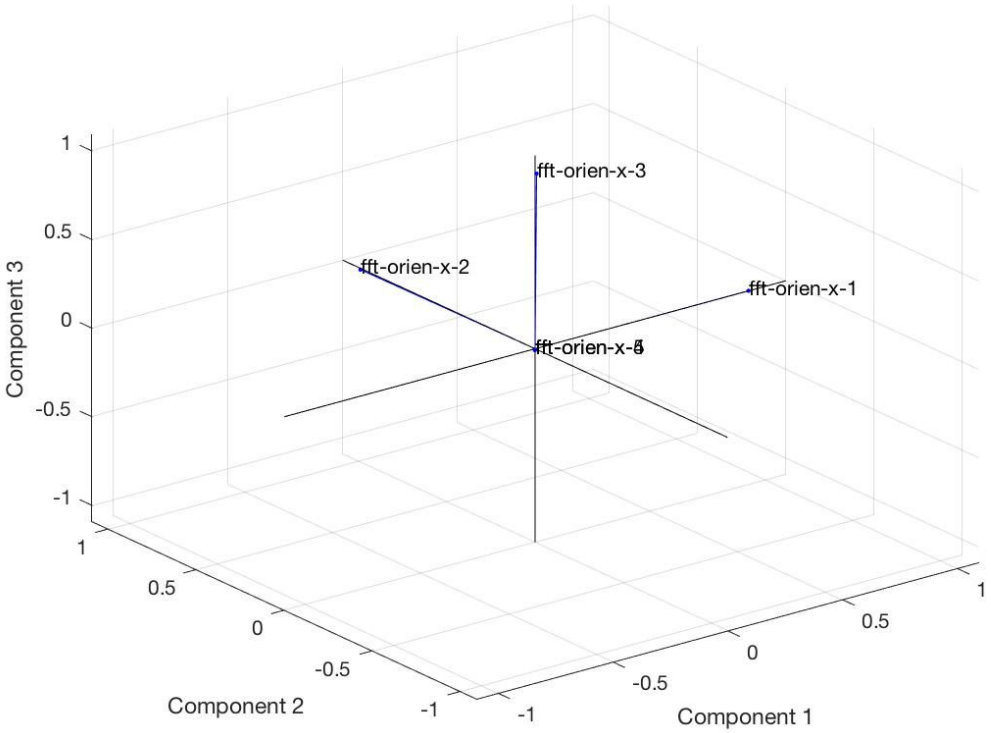
Below are the plots:




fft-acc-x-1
0.1
0.05
0
-0.05
-0.1
Component 3
0.1
0.05
0
-0.05
-0.1
Component 2
0.1
0.05
0
-0.05
-0.1
Component 1


fft-gyro-x-3
fft-gyro-x-2
fft-gyro-x-1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Component 3
0.5
0
-0.5
Component 2
0.5
0
-0.5
Component 1


fft-euler-roll-3
fft-euler-roll-2
fft-euler-roll-1
fft-euler-roll-5
Component 3
Component 2
Component 1
1
0.5
0
-0.5
-1


fft-orien-x-3
fft-orien-x-2
fft-orien-x-1
fft-orien-x-5
Component 3
Component 2
Component 1
1
0.5
0
-0.5
-1

By calculating, the first three principal components, and visualizing the samples in this three-dimensional space, we can create a visualization containing more of the variance in the original data than any other trio of linear combinations, so in this sense PCA provides the optimal three-dimensional sample representation. One of the keys behind the success of PCA is that in addition to the low-dimensional sample representation, it provides a synchronized low-dimensional representation of the variables.

6. Summary

In Phase 1, we performed collecting data of two routine activities using Myo band. We have collected the data for the activities of eating and cycling.

In Phase 2, we applied five feature techniques, i.e. FFT, DWT, Standard Deviation, Z Score and RMS. Out these five, we analysed that FFT is the best technique for feature extraction.

In Phase 3, we performed feature selection using PCA. PCA is an unsupervised method, meaning that no information about groups is used in the dimension reduction. This means that PCA shows a visual representation of the most important features in the data set. By most important we mean is only those attributes are chosen which are having high eigenvalues, i.e. high variance. So in this sense PCA provides the optimal feature selection and sample representation in lower dimension.