CSE 572 Data mining Project 1

A)

The 4 different feature-type used to extract the features from each data cell array are:

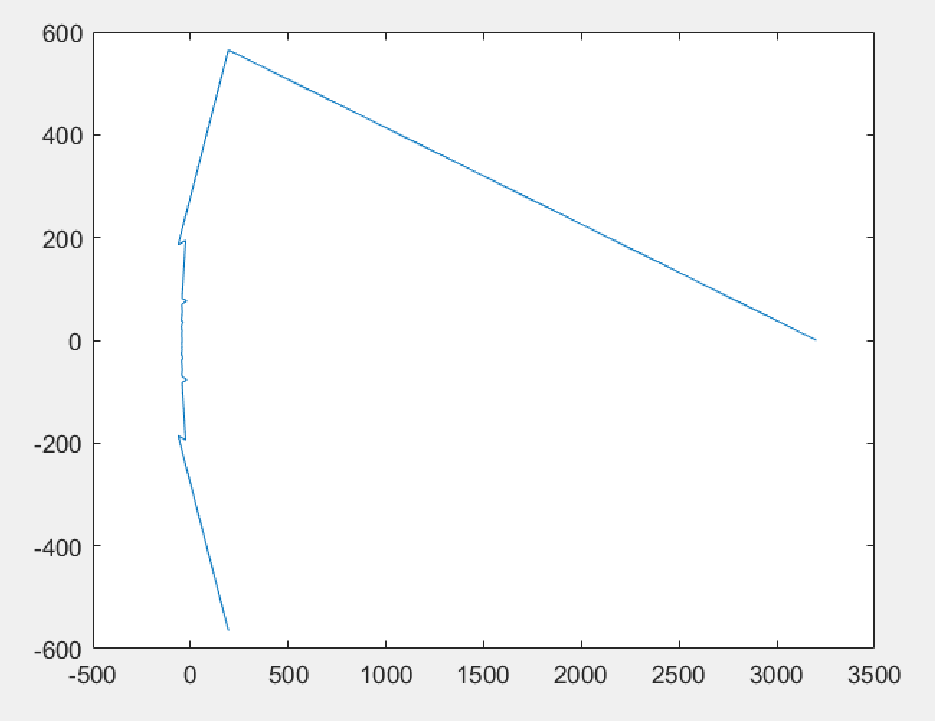
1) Fast Fourier transform
2) Discrete Wavelet Transform
3) Moving Standard Deviation
4) Moving Root Mean Square velocity

1) Fast Fourier transform (FFT):

A fast Fourier transform (FFT) computes the discrete Fourier transform (DFT) of a sequence decomposing a sequence of values into components of different frequencies. For each row of the CGMSeries1 matrix we calculate 'FFT'. Avoid the imaginary numbers by considering only the absolute values. Sort the row in descending order and select only the top 4 features. These top 4 features are considered as best representation features as per FFT section.

The values of top 4 'fft' features are:

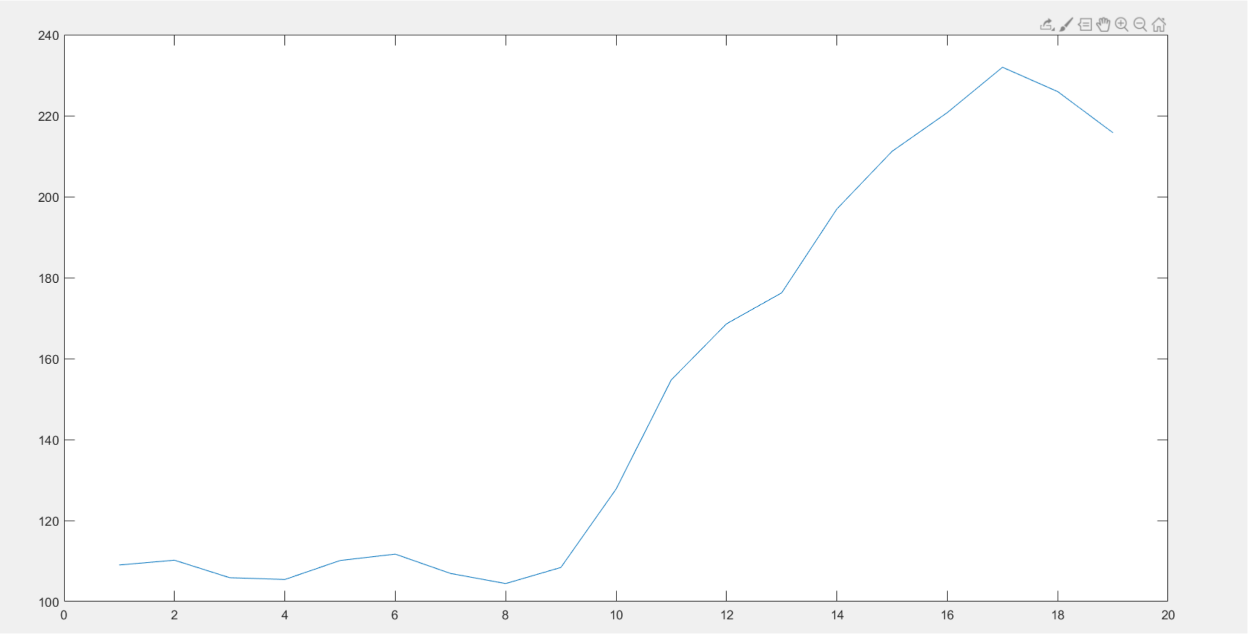
CGMSeries_p1_fft_value = 3.2020   0.5973   0.5973   0.1960



2) Discrete Wavelet Transform

Discrete Wavelet Transform decomposes a signal into a group of constituent signals, known as wavelets. They are useful to remove noise from the signals, detect abrupt discontinuities, and compress data. For each row of the CGMSeries1 matrix calculate 'dwt'. Sort the row in descending order and select only the top 4 features.

The values of top 4 'fft' feature values are:
CGMSeries_p1_dwt_values = 231.9581, 225.9441, 220.7593, 215.7907

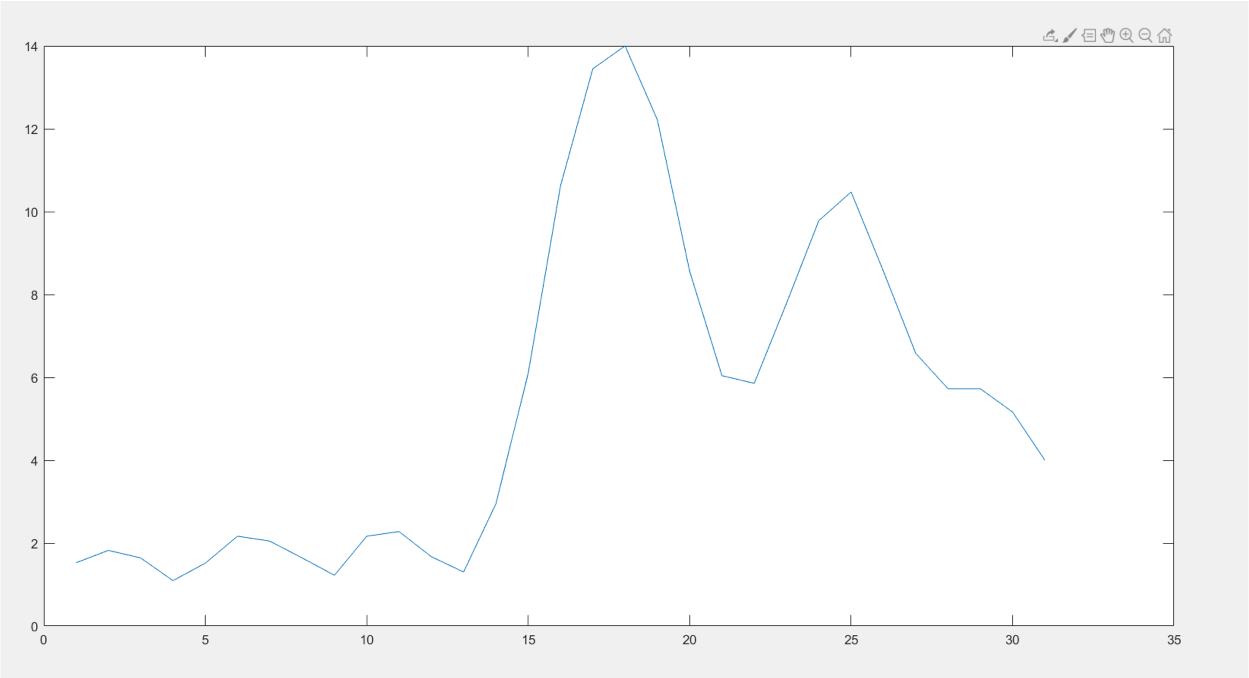


3) Moving Standard Deviation

Standard Deviation (SD) is the square root of the Variance which informs how spread out the values are. The moving standard deviation calculates SD over a sliding window of length 5 from each row. For each row in CGMSeries1 matrix calculate the moving standard deviation and sort the row in descending order. We select only the top 4 values from each row.

The values of top 4 'fft' feature values are:

CGMSeries_p1_std_values = 14.0000   13.4462   12.2147   10.6207

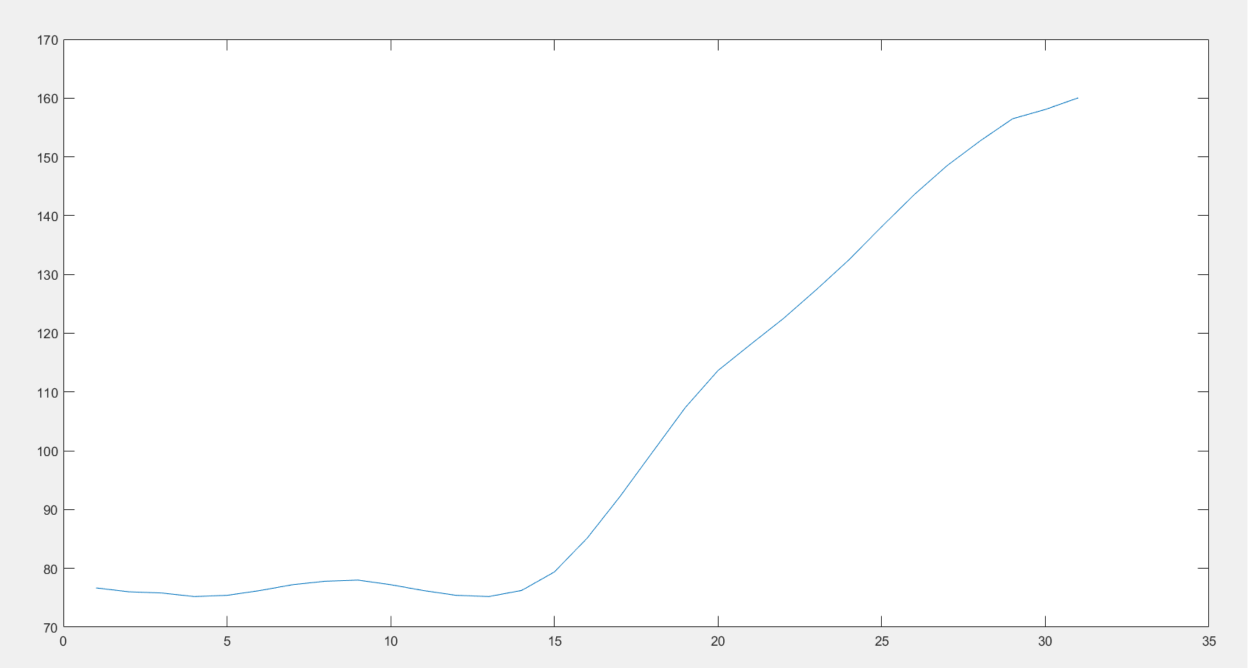


4) Moving Root Mean Square

RMS of a set of values is the square root of the arithmetic mean of the squares of the values. RMS is generally used to measure the magnitude of a data set. For each row, of the CGMSeries1 matrix we calculate Moving Root Mean Square. The RMS is calculated with a sliding window of 5. Sort the calculated values in descending order and select only the top 4 features.

The values of top 4 Moving Root Mean Square Velocity Feature values are:

CGMSeries_p1_rms_values = 160.0333 158.0633 156.4839 152.6860



B)

Fast Fourier transform

Fast Fourier transform algorithm is efficient in finding magnitude and location of the points that make up the signal of interest. In CGM data, FFT helps us to find the time period when the blood sugar level will be maximum for a patient.

Discrete Wavelet Transform

Discrete Wavelet Transform is useful in removing any noises in the timeseries. Moreover, it provides intuition of both frequency and time information of our data. Thus, it will provide information on the most frequent occurrences along with the information on when the event occurred.

Moving Standard Deviation

Standard Deviation is the square root of the Variance. We know that a higher variance feature is important for model. So, Moving Standard Deviation will help to find the features with high variance.

Moving Root Mean Square

The root mean square is a measure of the magnitude of the data set. So, the feature with highest magnitude will inform us when the patient had food and the raise in sugar level after the food consumption.

C)

We performed these four different feature extraction techniques namely

1) Fast Fourier transform
2) Discrete Wavelet Transform
3) Moving Standard Deviation
4) Moving Root Mean Square velocity

For Fast Fourier transform, in graph we can see a sudden rise in value and then stops at a certain magnitude on top and values start to decrease slowly. So, the point where the FFT is maximum may be the time, the person had meals.

For moving DWT, as the glucose levels increases, the moving DWT values will also increase gradually and therefore the maxima in the graph indicates that person has taken the meal. That is the reason why we took top 4 values for moving RWT feed as input to feature vector for each row of the patient data.

For Moving Standard Deviation, we can see that when there is rise in blood sugar level, then deviations are the high. So, the point when the deviation is highest, may tentatively be the time he had his meal.

For moving RMS, we can see after some time, the values start to rise suddenly and there is a continuous rise then onwards. This sudden rise may correspond to maximum change in the glucose level for a given time interval and is an indicator for the meal consumed by the person.

D)

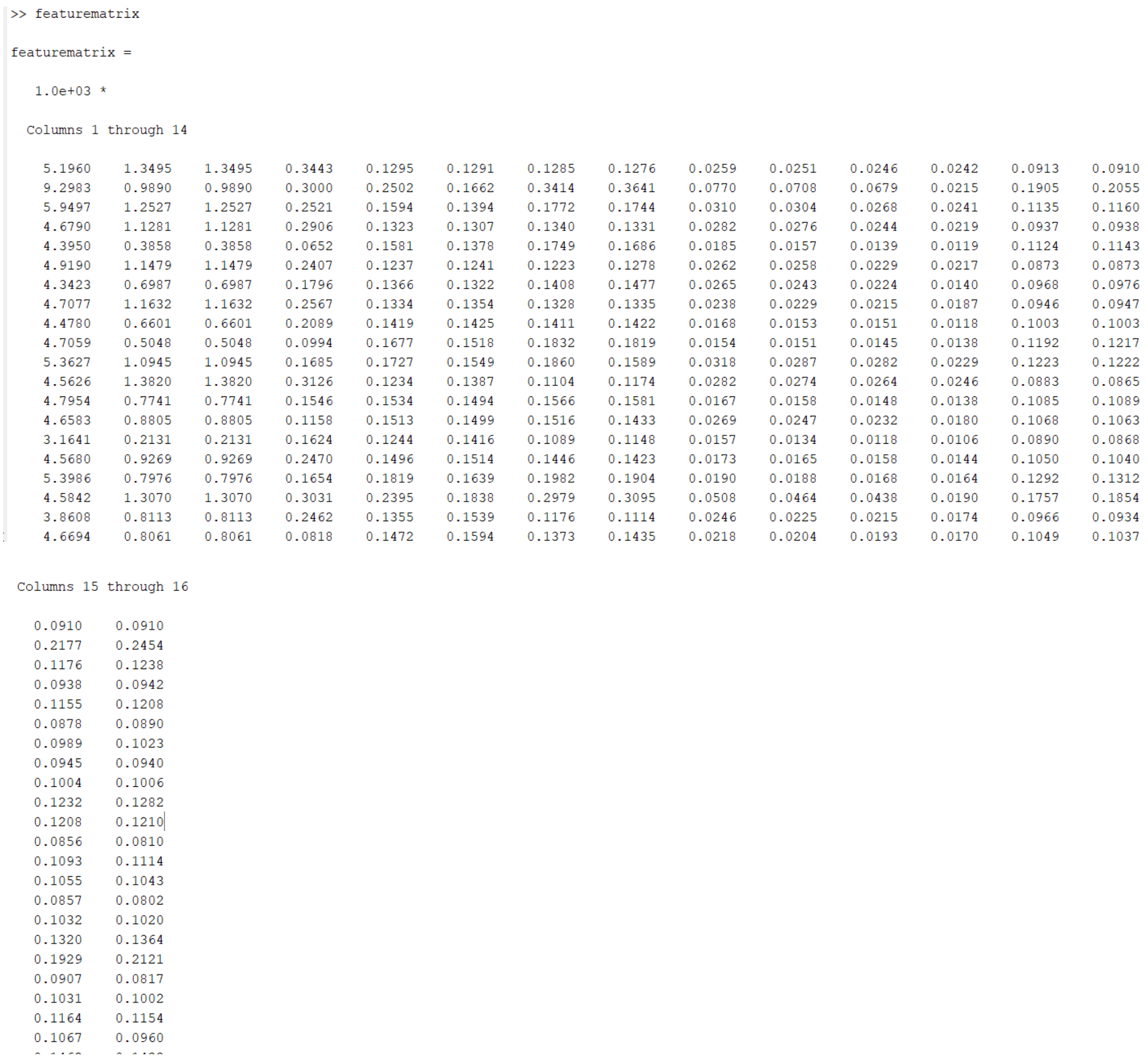
In Feature matrix each column represents a feature and contains values of a given feature across all the data points. Each row represents a set of data points. For each row in CGM series, we find top 4 features for each feature-type (4*4) and add them into a vector. The feature matrix is formed by concatenation of 33 vectors (all rows of CGM series).

A shorthand version of feature matrix code is given below:

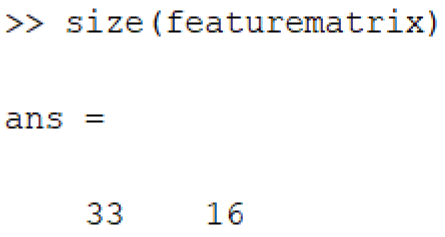
```
for rows= 1: size (CGMSeries1,1)
```

```
        vector = [CGMSeries_p1_fft_value, CGMSeries_p1_dwt_values,
        CGMSeries_p1_std_values, CGMSeries_p1_rms_values];
        feature_matrix = [feature_matrix; vector];
end
```

The resultant feature_matrix is 33*16 matrix as shows below:



```
>> featurematrix

featurematrix =

   1.0e+03 *

  Columns 1 through 14

    5.1960    1.3495    1.3495    0.3443    0.1295    0.1291    0.1285    0.1276    0.0259    0.0251    0.0246    0.0242    0.0913    0.0910
    9.2983    0.9890    0.9890    0.3000    0.2502    0.1662    0.3414    0.3641    0.0770    0.0708    0.0679    0.0215    0.1905    0.2055
    5.9497    1.2527    1.2527    0.2521    0.1594    0.1394    0.1772    0.1744    0.0310    0.0304    0.0268    0.0241    0.1135    0.1160
    4.6790    1.1281    1.1281    0.2906    0.1323    0.1307    0.1340    0.1331    0.0282    0.0276    0.0244    0.0219    0.0937    0.0938
    4.3950    0.3858    0.3858    0.0652    0.1581    0.1378    0.1749    0.1686    0.0185    0.0157    0.0139    0.0119    0.1124    0.1143
    4.9190    1.1479    1.1479    0.2407    0.1237    0.1241    0.1223    0.1278    0.0262    0.0258    0.0229    0.0217    0.0873    0.0873
    4.3423    0.6987    0.6987    0.1796    0.1366    0.1322    0.1408    0.1477    0.0265    0.0243    0.0224    0.0140    0.0968    0.0976
    4.7077    1.1632    1.1632    0.2567    0.1334    0.1354    0.1328    0.1335    0.0238    0.0229    0.0215    0.0187    0.0946    0.0947
    4.4780    0.6601    0.6601    0.2089    0.1419    0.1425    0.1411    0.1422    0.0168    0.0153    0.0151    0.0118    0.1003    0.1003
    4.7059    0.5048    0.5048    0.0994    0.1677    0.1518    0.1832    0.1819    0.0154    0.0151    0.0145    0.0138    0.1192    0.1217
    5.3627    1.0945    1.0945    0.1685    0.1727    0.1549    0.1860    0.1589    0.0318    0.0287    0.0282    0.0229    0.1223    0.1222
    4.5626    1.3820    1.3820    0.3126    0.1234    0.1387    0.1104    0.1174    0.0282    0.0274    0.0264    0.0246    0.0883    0.0865
    4.7954    0.7741    0.7741    0.1546    0.1534    0.1494    0.1566    0.1581    0.0167    0.0158    0.0148    0.0138    0.1085    0.1089
    4.6583    0.8805    0.8805    0.1158    0.1513    0.1499    0.1516    0.1433    0.0269    0.0247    0.0232    0.0180    0.1068    0.1063
    3.1641    0.2131    0.2131    0.1624    0.1244    0.1416    0.1089    0.1148    0.0157    0.0134    0.0118    0.0106    0.0890    0.0868
    4.5680    0.9269    0.9269    0.2470    0.1496    0.1514    0.1446    0.1423    0.0173    0.0165    0.0158    0.0144    0.1050    0.1040
    5.3986    0.7976    0.7976    0.1654    0.1819    0.1639    0.1982    0.1904    0.0190    0.0188    0.0168    0.0164    0.1292    0.1312
    4.5842    1.3070    1.3070    0.3031    0.2395    0.1838    0.2979    0.3095    0.0508    0.0464    0.0438    0.0190    0.1757    0.1854
    3.8608    0.8113    0.8113    0.2462    0.1355    0.1539    0.1176    0.1114    0.0246    0.0225    0.0215    0.0174    0.0966    0.0934
    4.6694    0.8061    0.8061    0.0818    0.1472    0.1594    0.1373    0.1435    0.0218    0.0204    0.0193    0.0170    0.1049    0.1037

 Columns 15 through 16

   0.0910    0.0910
   0.2177    0.2454
   0.1176    0.1238
   0.0938    0.0942
   0.1155    0.1208
   0.0878    0.0890
   0.0989    0.1023
   0.0945    0.0940
   0.1004    0.1006
   0.1232    0.1282
   0.1208    0.1210
   0.0856    0.0810
   0.1093    0.1114
   0.1055    0.1043
   0.0857    0.0802
   0.1032    0.1020
   0.1320    0.1364
   0.1929    0.2121
   0.0907    0.0817
   0.1031    0.1002
   0.1164    0.1154
   0.1067    0.0960
```

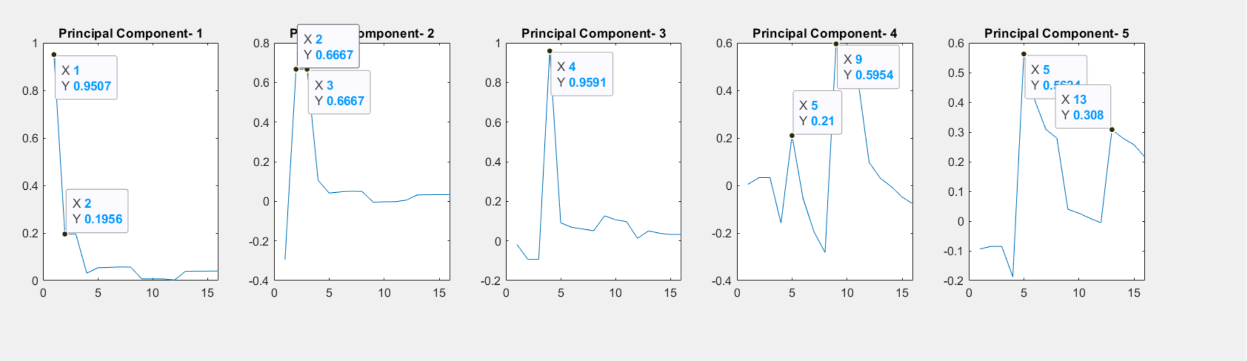


```
>> size(featurematrix)

ans =

    33    16
```

E) Provide this feature matrix to PCA and derive the new feature matrix. Chose the top 5 features and plot them for each time series. (5 points)

Before the PCA operation, we will normalize the feature matrix because the features have different range. The matrix is normalized by the Euclidean norm (2-norm). Then, find the co-efficient of the feature matrix using PCA. The 5 higher importance features are chosen. These top 5 features are selected and multiplied with normalized_feature_matrix to get an updated_feature_matrix

The Principal Component graphs are shown below:



F) For each feature in the top 5 argue why it is chosen as a top five feature in PCA? (3 points each) total 15.

The top 5 Principal components chosen as PCA returns the Eigen vectors. The Eigen vectors are in decreasing order. In each graph, we see that in each component there are certain features which have a greater importance.

For example, the principal component plots 1,2,3 feature one which are the 1st,2nd, 3rd and 4 th maximum value of FFT. This suggest that "FFT" is the most important feature to consider. And thus FFT states when the person had the meal.

Also, in the fourth component, we have features 2 which are maximum values Discrete Wavelet Transform time series and Moving Standard Deviation.