Data analysis of Epileptic seizure data

Suhas Shastry

November 24, 2017

Background

Epileptic seizure is the abnormal activity in the brain signals for few moments resulting in strange sensations, emotions, and behavior. The data was captured by a technique called Electroencephalography (EEG). EEG is a noninvasive method involves placing metal electrodes on scalp which measure voltage fluctuation in brain signals.

Data was collected on 100 patients and 400 healthy test subjects for a duration of 23.6 seconds. This time series data was sampled into 4097 data points. This dataset after integration has 500 rows and 4097 columns. Original data being high-dimensional in nature, it poses a challenge for classical methods in data analysis as sample co-variance matrix is not a good estimator of population co variance matrix.

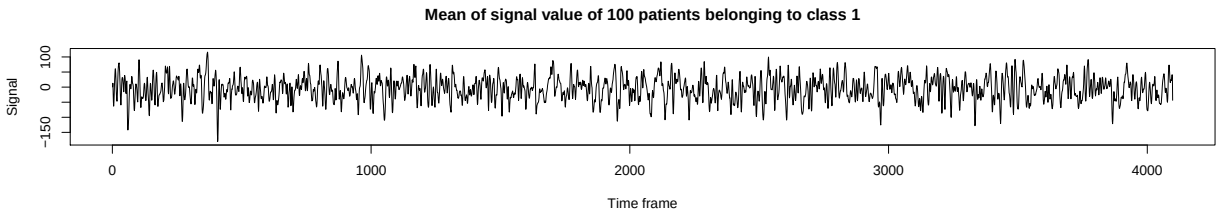
Hence data was then divided and shuffled every 4097 data points into 23 chunks by data owner, each chunk contains 178 data points for 1 second, and each data point is the value of the EEG recording at a different point in time. So now data has 23 x 500 = 11500 pieces of information(row), each information contains 178 data points for 1 second(column), the last column represents the label y $\{1, 2, 3, 4, 5\}$.
The response variable is y in column 179, the Explanatory variables $X_1, X_2, \ldots, X_{178}$
y contains the category of the 178-dimensional input vector. Specifically, y in $\{1, 2, 3, 4, 5\}$:

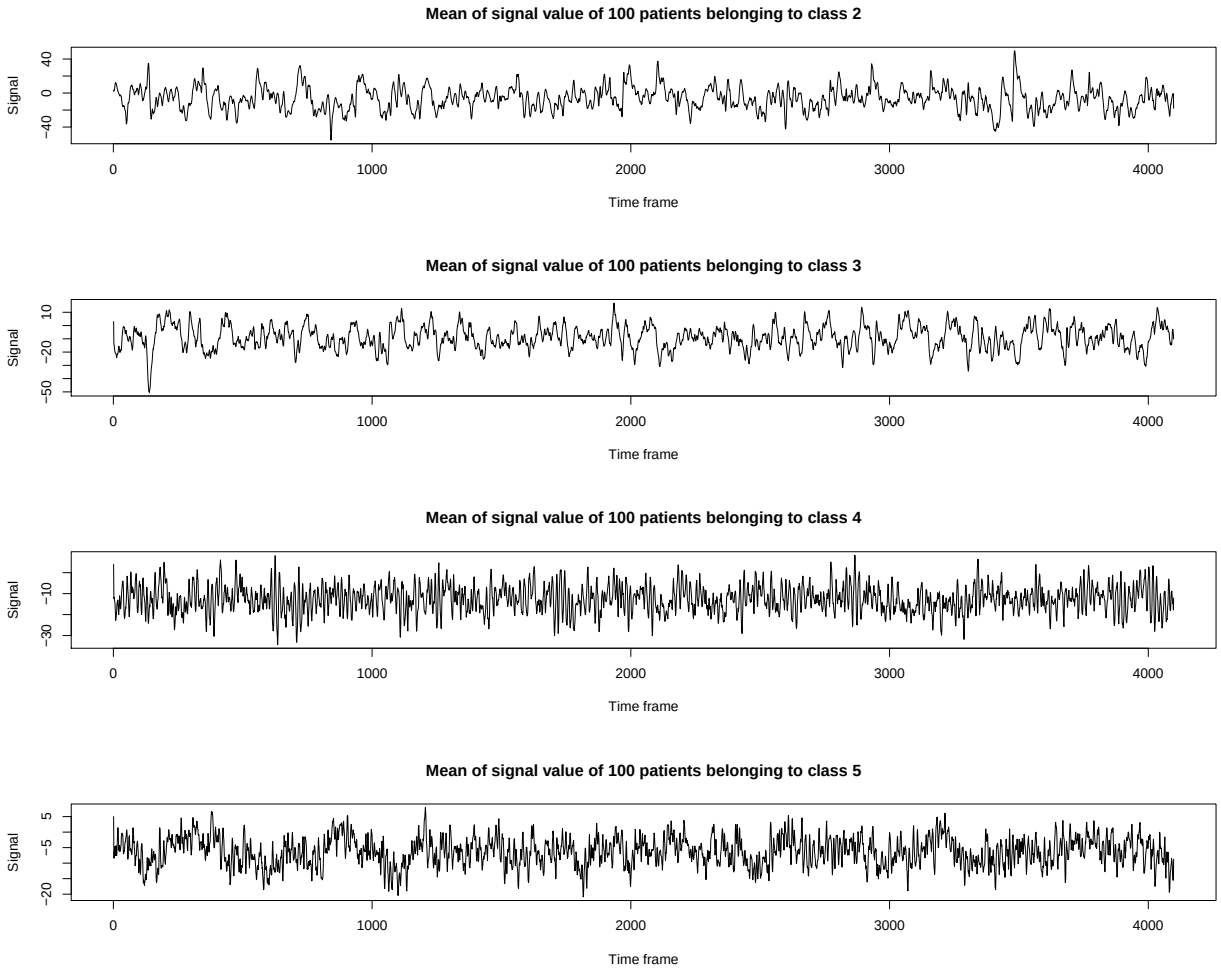
1 - Recording of seizure activity
2 - They recorder the EEG from the area where the tumor was located
3 - They identify where the region of the tumor was in the brain and recording the EEG activity from the healthy brain area
4 - eyes closed, means when they were recording the EEG signal the patient had their eyes closed
5 - eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open
The goal of this project is to distinguish seizure data and non-seizure data.

Analysis

Time series data

To picture the original data, column means were plotted for 4097 time-frames for 5 different classes.





Even though nothing can from be inferred from these graphs, it gives the basic idea about the brain signals corresponding to five different classes.

Manova

Idea here is to check whether all five classes have the same mean vector. For this lets first check the equality of variance as Manova requires five datasets to be from population having same co-variance matrix.

```
#Testing cov.x1 = cov.x2 = cov.x3 = cov.x4 = cov.x5
pool.cov <- ((n-1)/N)*(cov.x1 + cov.x2 + cov.x3 + cov.x4 + cov.x5)
test.stat1 <- N*logofdetP(pool.cov) - n*(logofdetP(cov.x1) +
            logofdetN(cov.x2) + logofdetN(cov.x3) +
              logofdetN(cov.x4) +logofdetN(cov.x5))
m <- (k-1)*p*(p+1)/2
thr1.chisq <- qchisq(0.95, m)
test.stat1>thr1.chisq
```

```
## [1] TRUE
```

As test statistic is greater than critical value, we are rejecting null hypothesis at level $\alpha = 0.05$. As all datasets are not from population of same co-variance matrix, we cannot continue further to conduct Manova test.

Hotelling T^2

Since means of five datasets cannot be compared to equality, motivation here is to test pairwise using two sample Hotelling T^2 test.

Class 1 and rest

Mean μ_1 of class 1 = Mean of class 2 to 5 μ_{2-5}.
$H_0 : \mu_1 = \mu_{2-5}$ vs $H_A : \mu_1 \neq \mu_{2-5}$

```
Hotelling(mean.x1,mean.x25,cov.x1,cov.x25,n,n*4)
```

```
##              [,1]
## [1,] 0.001660352
```

There is a strong evidence ($p - value = 0.0017$) to conclude that means are not equal. This indicates that there is a possibility of data classification into two groups, one belonging to class 1 and the other belonging to rest of the classes. Even though there is no guarantee that a classification should exist, rejecting null hypothesis is a motivation to attempt classification on data set which is done in later section.

Class 1 and class 3

Mean μ_1 of class 1 = Mean μ_3 of class 3
$H_0 : \mu_1 = \mu_3$ vs $H_A : \mu_1 \neq \mu_3$
As datasets of class 1 is collected from seizure activity and class 3 is collected from healthy brain in the same region, there should be difference in signal activities. Else the entire data analysis makes very little sense. Hence their means are tested for equality

```
Hotelling(mean.x1,mean.x3,cov.x1,cov.x3,n,n)
```

```
##              [,1]
## [1,] 0.003398835
```

Clearly, these means are not equal ($p - value = 0.0034$). Classification is done on these two groups in later section.

Class 4 and class 5

Mean μ_4 of class 4 = Mean μ_5 of class 5
$H_0 : \mu_4 = \mu_5$ vs $H_A : \mu_4 \neq \mu_5$
Class 4 indicates the signals of healthy brain when the subject closes his/her eyes. Class 5 indicates the signals of healthy brain when the subject opens his/her eyes. Expecting these means to be different.

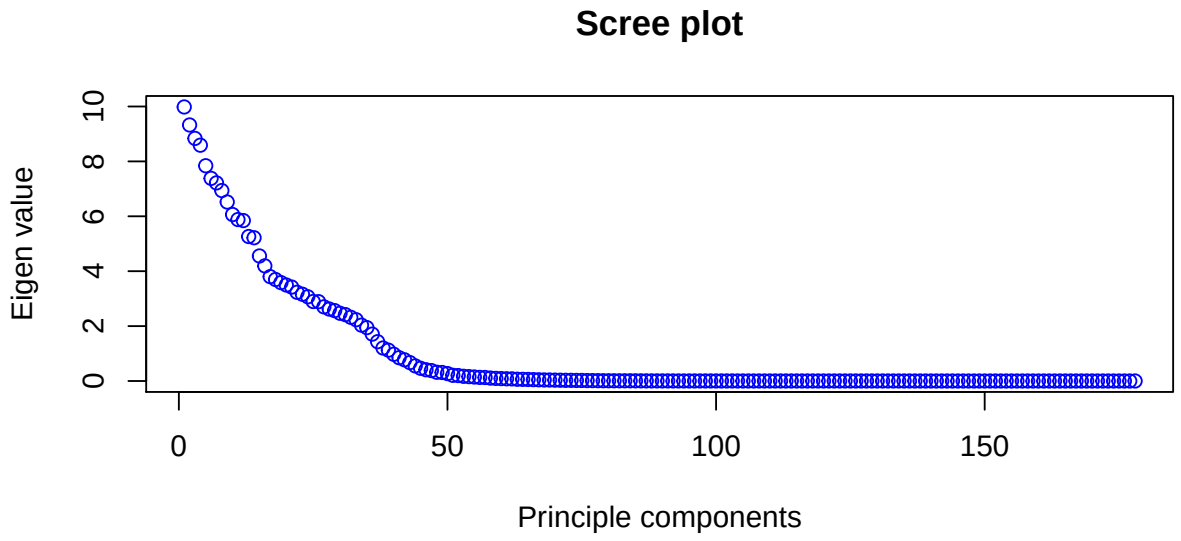
```
Hotelling(mean.x4,mean.x5,cov.x4,cov.x5,n,n)
```

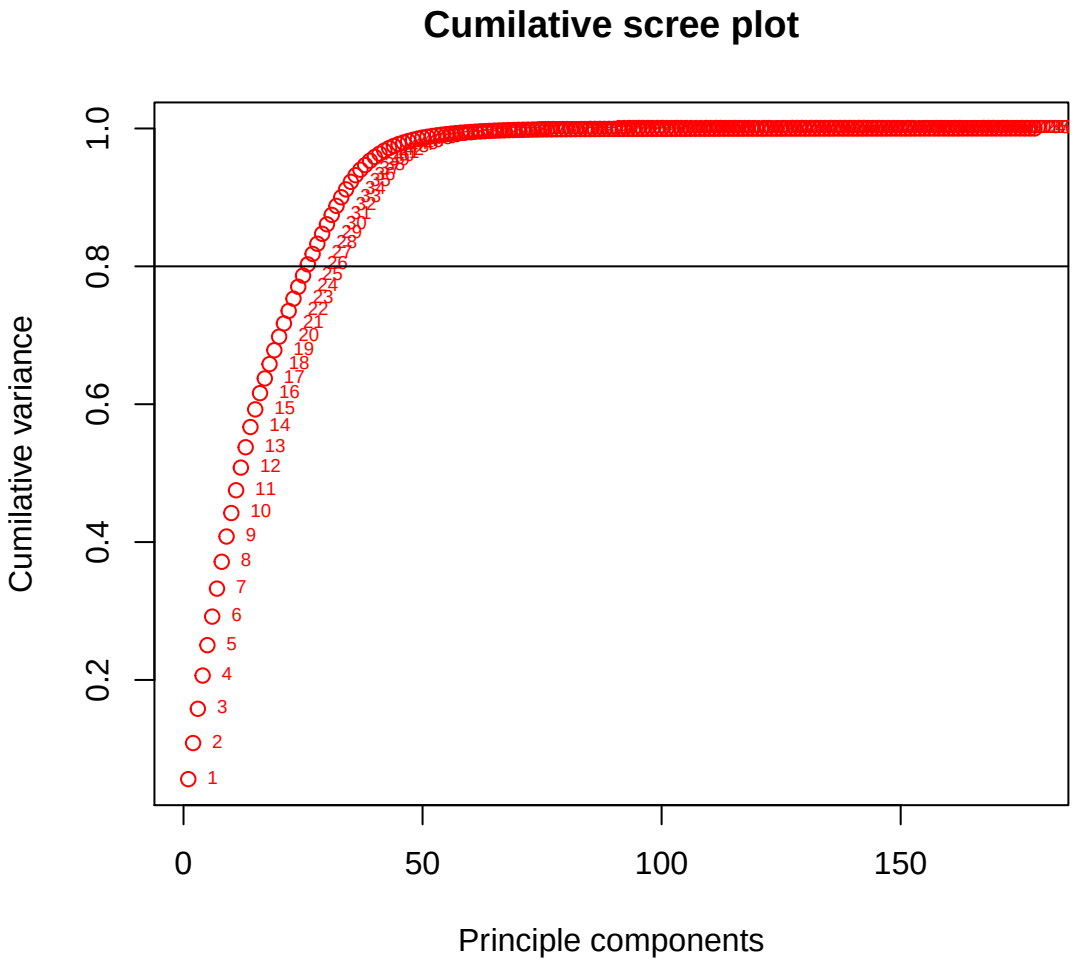
```
##              [,1]
## [1,] 0.003714651
```

These means are different ($p - value = 0.0037$).

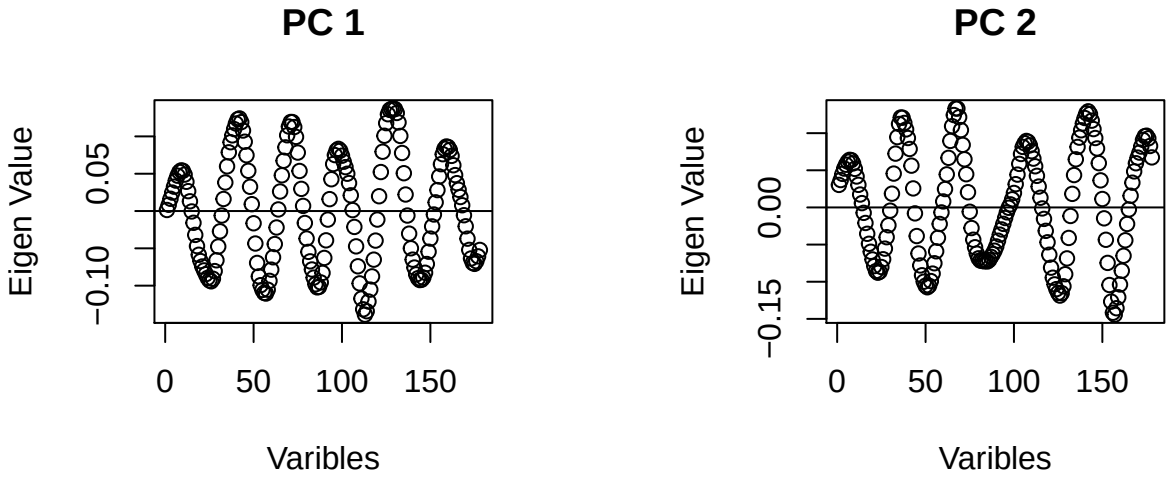
Principal Component Analysis (PCA)

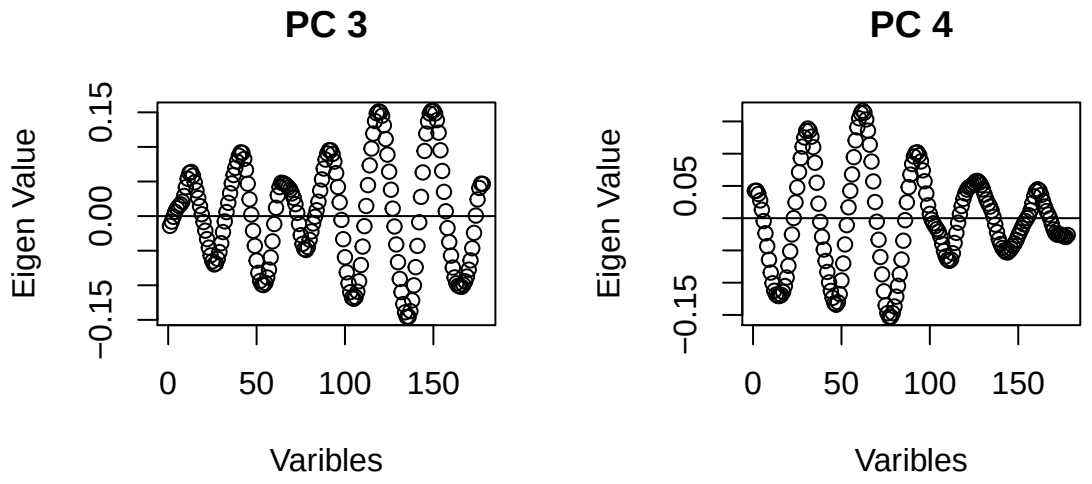
Analysis on the co-variance matrix of the data is carried out using PCA. PCA converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs). As the data has 178 columns, we get a 178X178 farinaceous matrix and hence 178 PCs. Here are the scree plot of PCs and cumulative scree plot of PCs. Scree plot gives the idea about the variance explained by each PC and cumulative scree plot explains percentage of total variance explained by PCs.



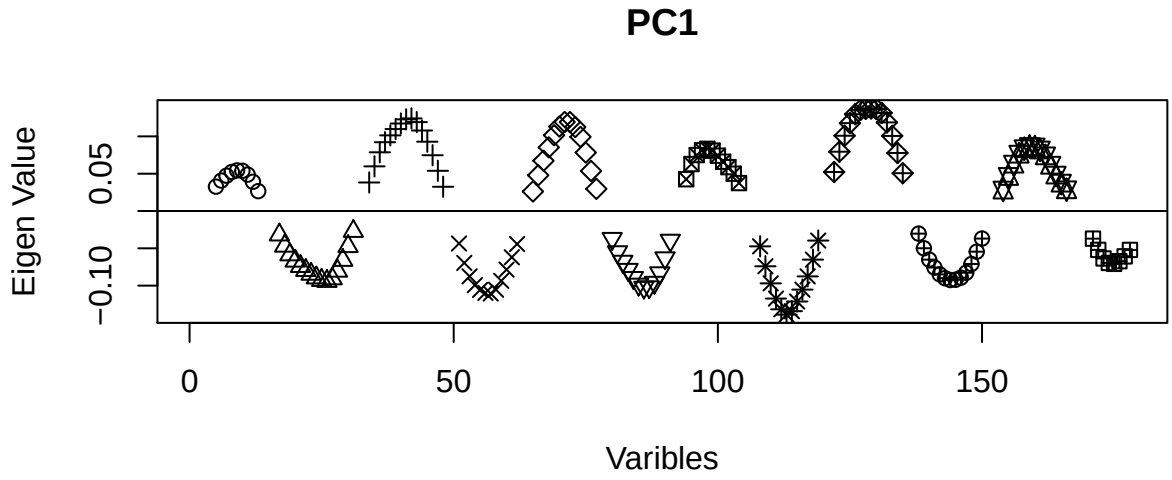


Since it takes 25 PCs to explain 80% of the total variance, only loading plots of first 4 PCs are drawn.





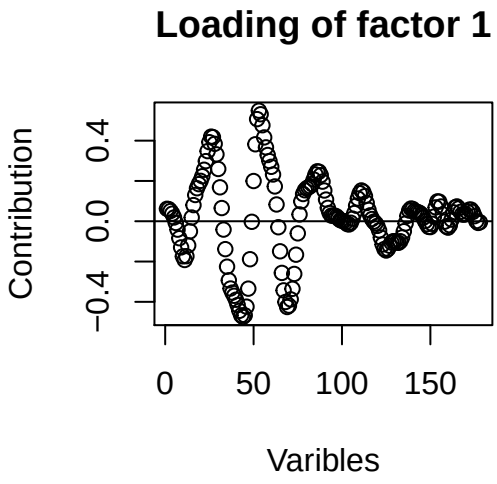
These PCs are used later for dimension reduction and clustering.

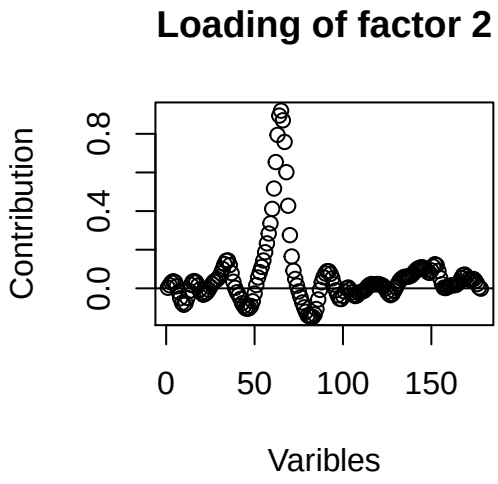


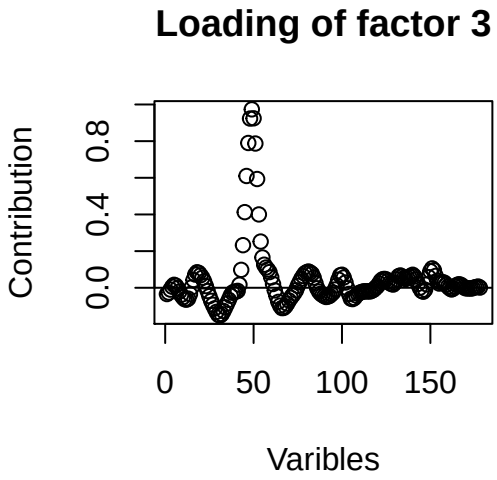
PC1 is majorly dependent on these 12 sets of variables. Linear combination of these 12 sets of variables has the highest variance.

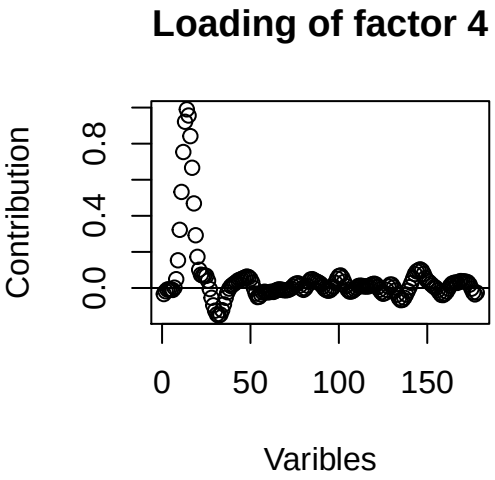
Factor Analysis

Aim is to find factors loadings or group of variables which have the highest co variances among them. Latent factors driving the data can be obtained using factor analysis.





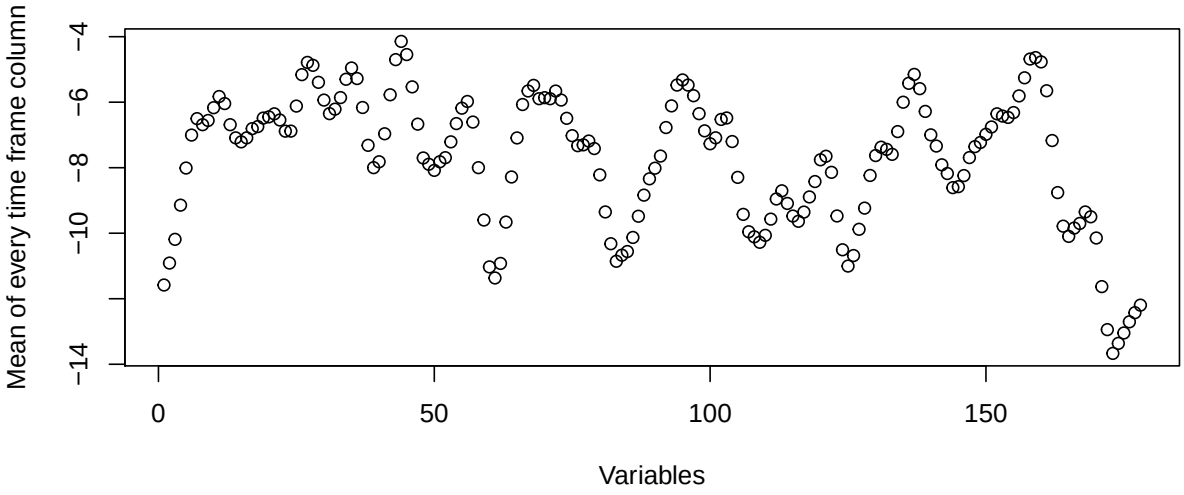




Few spikes can be seen in loadings of each factor. Contribution from rest of the variables is close to zero. There is a clear factor representation in variables which are represented by spikes. These variables in spikes have the highest co variance among themselves. In the data point of view, brain signals of these time frames are closely dependent on each other.

Canonical Correlation Analysis (CCA)

Variables are divided into two groups such that linear combination of these variables will have highest correlation between the groups. To find the two group of variables mean of each column in the data was plot expecting that there will be some spike of disruption in the graph.



Since there are more than one spikes, data could not be divided into two meaningful groups and hence CCA could not be conducted on the data.

Cluster Analysis

The main ambition of the project is to distinguish seizure data and non seizure data. Clustering is a basic to find intuitive guidance to separate data into classes. Classical clustering algorithms are applied on whole data. For each algorithm, a table is formed. Table rows are the actual classes data belong and columns are the classes after clustering. Ideally all the entries are expected on diagonals and non-diagonal entries are expected to be zero.

Complete linkage hierarchical clustering

```
##
##        1    2    3    4    5
##   1 1474  629   60  136    1
##   2 2257   39    1    3    0
##   3 2300    0    0    0    0
##   4 2300    0    0    0    0
##   5 2300    0    0    0    0
```

Average linkage hierarchical clustering

```
##
##        1    2    3    4    5
##   1 2294    1    1    3    1
##   2 2300    0    0    0    0
##   3 2300    0    0    0    0
##   4 2300    0    0    0    0
##   5 2300    0    0    0    0
```

Single linkage hierarchical clustering

```
##
##        1    2    3    4    5
##   1 2296    1    1    1    1
##   2 2300    0    0    0    0
##   3 2300    0    0    0    0
##   4 2300    0    0    0    0
##   5 2300    0    0    0    0
```

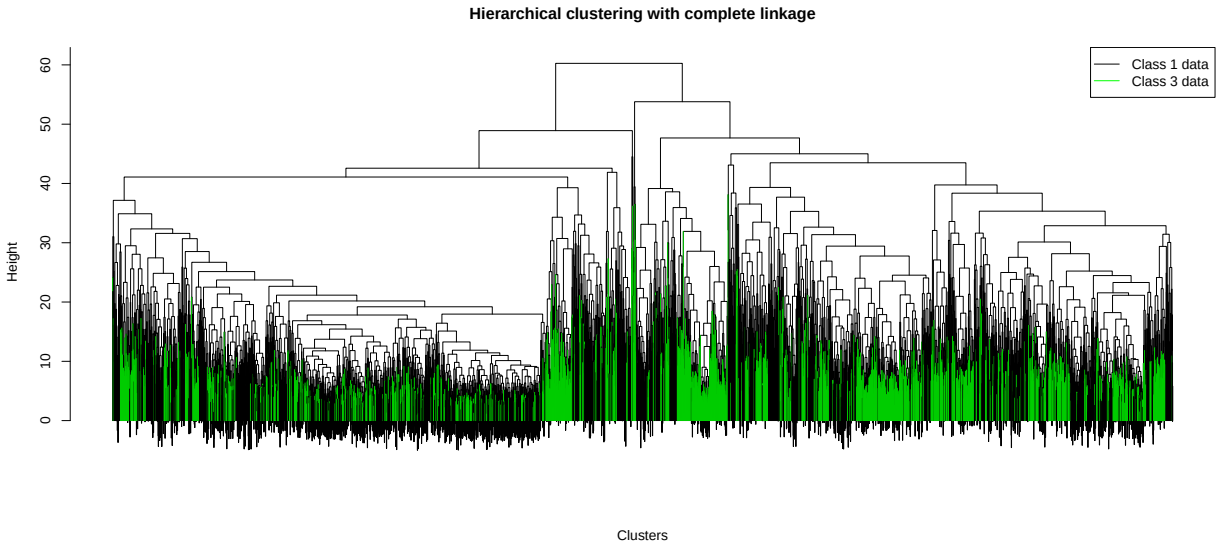
K-Means clustering

```
##
##        1    2    3    4    5
##   1  774  589  330  333  274
##   2 1497  787    1    9    6
##   3 1503  797    0    0    0
##   4 1456  844    0    0    0
##   5 1658  642    0    0    0
```

As we can see, none of the clustering methods are good. No algorithm could distinguish seizure data and non-seizure data. Another approach involving data reduction from 178 dimensions to 25 dimensions using PCA was employed. Clustering was tried on reduced dimension space between seizure data (Class 1) and healthy brain's data collected from the seizure region (Class 3).

Complete linkage for two classes

Dendrogram for two classes is plotted to check the behavior of the algorithm.



As the data points of two different colors are deeply mixed among each other (indicated by different color), simple clustering algorithms cannot distinguish the data.

```
##          cl_complete
```

```
##           Class 1 Class 3
##   Class 1     976    1324
##   Class 3    1364     936
```

K-means for two classes

```
##
##           Class 1 Class 3
##   Class 1    1686     614
##   Class 3    1264    1036
```

Classical clustering algorithm fail to distinguish seizure and non seizure data even in reduced space. Hence classification which employs learning from know responses is a way out to distinguish the data.

Classification

K-nearest neighbor (knn)

5-nearest neighbor classification algorithm was applied. Whole data set with 5 classes is used as training data. Then classification rules are applied on the same dataset.

```
##            cl
##             Class 1 Class 2-5
##   Class 1      1650       650
##   Class 2-5       1      9199
```

knn does mediocre job in establishing classification rules to distinguish seizure data and normal data. Same algorithm is applied to class 1 and class 3 data but with a slight modification. First half of the data is used as training data to classify the rest half of the data.

```
##          cl
##           Class 1 Class 3
##   Class 1     777     373
##   Class 3       0    1150
```

Classification seems reasonably better than clustering. But knn is a very basic algorithm and the classification can be improved using random forest and naive Bayes algorithm.

Random forest

Random Forrest is applied to class 1 and class 3 data in which first half of the data is used as training data to classify the rest half of the data.

```
##
##           Class 1 Class 3
##   Class 1    1127      23
##   Class 3      39    1111
```

Prediction of random forrest is worse than knn. The same algorithm was applied to distinguish the state eyes. Class 4 represents the brain signal when eyes are closed and class 5 represents brain signals when eyes are open.

```
##
##           Class 4 Class 5
##   Class 4     905     245
##   Class 5     245     905
```

Random forest is not so bad in classifying class 4 and class 5 data.

Comments and Future work

In the begining of the analysis, Manova was tried on the data set to compare mean vectors belonging to each class. As it could not be applied, pair-wise Hotelling T^2 was applied on specific data. Even though Hotelling T^2 could significantly conclude that means are not equal, this was not helpful in clustering or classification. Later PCA, FA and CCA was applied so as to get the insights about the variance-covarinace matrix of the population. There were sets of variable clearly driving components in PCA. But using them for dimension reduction and do clustering could not give better results.

Original time series data was morphed into small dimensional data. Much of the information was lost in this transformation. In future, time series analysis can be done on the original dataset. Even random forests and knn are very unsophisticated learning classification algorithms. It can be improved with Naive Bayes, SVM or neural networks. Here is a classification example of SVM on original time series data trying to classify seizure data and rest.

```
##            pred
##             Class 1 Class 2-5
##   Class 1        40         10
##   Class 2-5       2        198
```