

Modelling EEG Signals Using Polynomial Regression

Introduction to Statistical Methods for Data Science



February 3, 2023

Softwarica College

Contents

[Introduction to EEG signal 2](#_Toc127390720)

[Objective of assignment 3](#_Toc127390721)

[Methodology 3](#_Toc127390722)

[Result 3](#_Toc127390723)

[Task 1: Show the Time series plots of input and output EEG signal 3](#_Toc127390724)

[Task 1.2: Distribution of each EEG signal 4](#_Toc127390725)

[6](#_Toc127390726)

[Task 1.3.: Show Correlation and scatter plots between different input and output signals of EEG 6](#_Toc127390727)

[Correlation shows the relation or degree of association and measures the direction and strength of the linear relationship between two variables.It ranges from -1 to +1. -1 refer to perfect negative correlation, +1 refers to perfect positive correlation and 0 refers to no correlation. Scatter plot plots two variables from which we can see the one variables affected by another. It can show the linear and curvilinear in the data present. In this task, I use the cor() function to calculate the correlation and plot() command to plot the scatter plot of signals. 6](#_Toc127390728)

[X1 signal and output signal 6](#_Toc127390729)

[X2 signal and output signal 7](#_Toc127390730)

[X3 signal and output signal 8](#_Toc127390731)

[X4 signal and output signal 9](#_Toc127390732)

[Task 2: Regression modeling the relationship between EEG signals 10](#_Toc127390733)

[Task 2.1: Estimating the parameter of the model using linear regression 11](#_Toc127390734)

[Task 2.2: Calculating RSS (model Residual Sum of Squared error) for each model 12](#_Toc127390735)

[Task 2.3 & 2.4 : Calculation AIC, BIC & log likelihood function 12](#_Toc127390736)

[For model 1 13](#_Toc127390737)

[For model 2 13](#_Toc127390738)

[For model 3 13](#_Toc127390739)

[For model 4 14](#_Toc127390740)

[For model 5 14](#_Toc127390741)

[For model 1 14](#_Toc127390742)

[For model 2 15](#_Toc127390743)

[For model 3 15](#_Toc127390744)

[Task 2.6: Choosing the best model 17](#_Toc127390745)

[Task 2.7: Training and testing the Model 2 from the 17](#_Toc127390746)

[17](#_Toc127390747)

[Task 3: Approximate Bayesian Computation (ABC 18](#_Toc127390748)

[Conclusion 21](#_Toc127390749)

[References 21](#_Toc127390750)

# Introduction to EEG signal

EEG, which stands for "Electroencephalography," is a medical test that measures the electrical activity of a brain. It's used to diagnose and monitor various neurological conditions, such as epilepsy, sleep disorders, and brain damage, and to study brain functions. It is also used in animals for the same research and diagnosis as humans.

The test is performed by attaching electrodes to the scalp. The electrodes pick up the electrical signals produced by the brain and send them to a machine that records and displays the results as a series of waves. The waves represent the electrical activity in different parts of the brain.

EEG is a non-invasive test, meaning it does not involve incisions or injections and is generally considered safe. The test typically takes about 30 minutes to an hour and does not cause pain or discomfort.

It is one diagnostic tool in neurology and is commonly used in combination with other tests, such as imaging studies, to get a complete picture of the brain's function and health.

# Objective of assignment

1. Plotting the signal to find a pattern.
2. Finding the best model using the regression model.
3. Training and testing the best model using the data given.

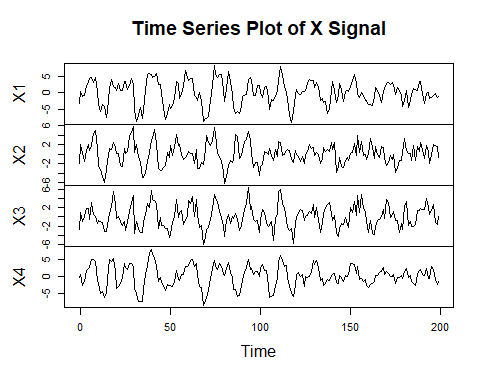
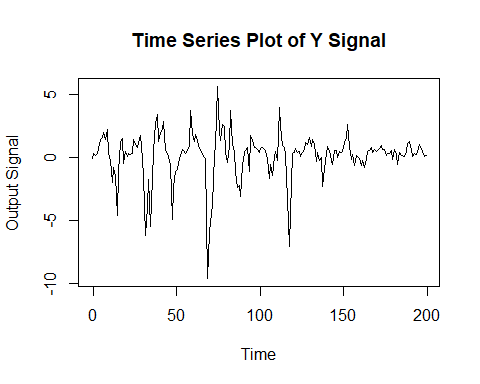
# Methodology

The tasks were completed using R-Lab. The data set and libraries required in R-Lab are included.

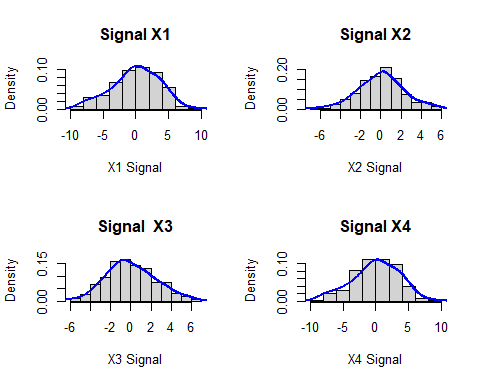
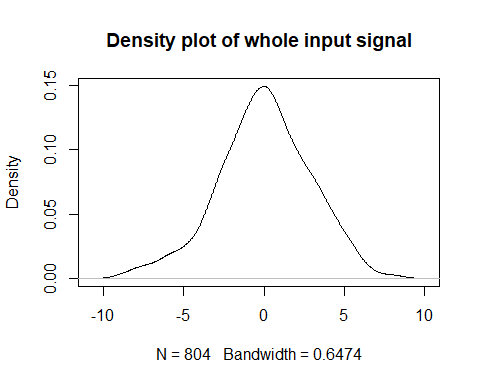
# Result

## Task 1: Show the Time series plots of input and output EEG signal

A time series plot is a data visualization tool that illustrates data points at successive intervals of time.For the time series plot, I used the ts() command to define X and Y variables against time. The start point is minimum of X variables and end point is maximum of time data and frequency is set to 1. Then I used the plot() command to plot the time series command

Task 1.2: Distribution of each EEG signal

A histogram is a graphical tool which shows the frequency of the data. A density plot is based on a histogram which shows the line, smooths out the curve and, reduces the noise in the data. Histogram and density can be a plot in the same graph using plot() and abline() functions. To plot the density of the signals, I first calculated the density of the required signal and plotted it using the plot() command.



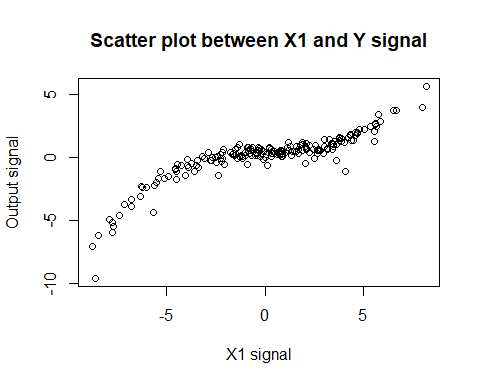
## 

## Task 1.3.: Show Correlation and scatter plots between different input and output signals of EEG

### Correlation shows the relation or degree of association and measures the direction and strength of the linear relationship between two variables.It ranges from -1 to +1. -1 refer to perfect negative correlation, +1 refers to perfect positive correlation and 0 refers to no correlation. Scatter plot plots two variables from which we can see the one variables affected by another. It can show the linear and curvilinear in the data present. In this task, I use the cor() function to calculate the correlation and plot() command to plot the scatter plot of signals.

### X1 signal and output signal

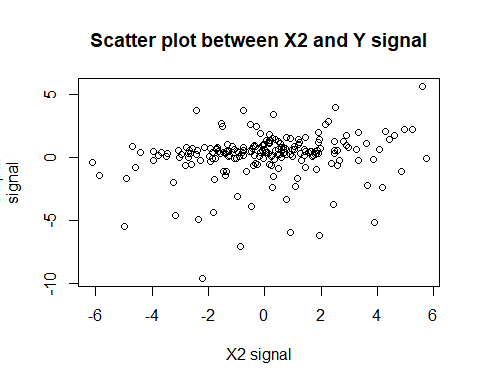
0.8569935



The correlation (.856) show fairly strong positive relationship and scatter plot show the curvilinear relation between X1 and output signal.

### X2 signal and output signal

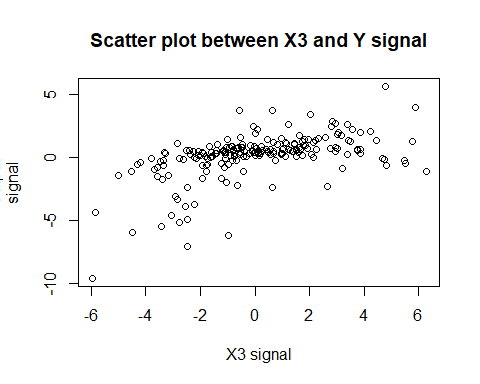
0.1941002



The correlation (.194) show the very weak positive relationship and scatter plot show the linear relation between X2 and output signal.

### X3 signal and output signal

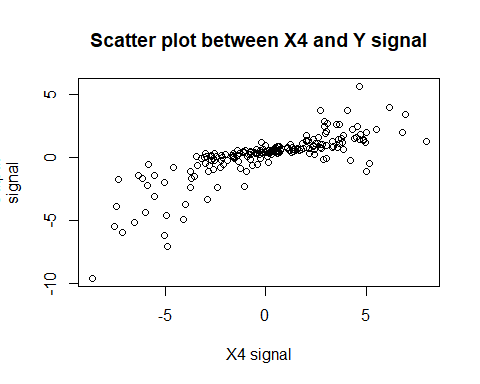
0.5255484



The correlation (.525) show the weak positive relationship and scatter plot show the linear relation between X3 and output signal.

### X4 signal and output signal

0.7746722



The correlation (.77) show the strong positive relationship and scatter plot show the linear relation between X4 and output signal.

## Task 2: Regression modeling the relationship between EEG signals

Regression modelling can be done with various methods, but for this assignment, we chose the least square method (linear regression). Linear regression can solve a nonlinear polynomial equation. We were given five nonlinear polynomials models that could best fit the data. From those models, we have to choose the best model. We will look for the RSS, AIC (Akaike information criterion), BIC (Bayesian information criterion) and log-likelihood functions. The model given were:

Model 1: 𝑦 = 𝜃1 𝑥4 + 𝜃2 𝑥12 + 𝜃3 𝑥13 + 𝜃4 𝑥24 + 𝜃5 𝑥14 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

Model 2: 𝑦 = 𝜃1 𝑥4 + 𝜃2 𝑥13 + 𝜃3 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

Model 3: 𝑦 = 𝜃1 𝑥33 + 𝜃2 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

Model 4: 𝑦 = 𝜃1 𝑥2 + 𝜃2 𝑥13 + 𝜃4 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

Model 5: 𝑦 = 𝜃1 𝑥4 + 𝜃2 𝑥12 + 𝜃3 𝑥13 + 𝜃4 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

### Task 2.1: Estimating the parameter of the model using linear regression

In this task I used lm() command for least square method and used its attribute coefficients to look for the coefficients (parameter of required model).

#### For model 1

Model 1: 𝑦 = 𝜃1 𝑥4 + 𝜃2 𝑥12 + 𝜃3 𝑥13 + 𝜃4 𝑥24 + 𝜃5 𝑥14 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

(Intercept) I(X4) I(X1^2) I(X1^3) I(X2^4)   
 0.4015807794 0.1277101097 -0.0002902170 0.0096688129 -0.0004098917   
 I(X1^4)   
 -0.0001543367

#### For model 2

Model 2: 𝑦 = 𝜃1 𝑥4 + 𝜃2 𝑥13 + 𝜃3 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

(Intercept) I(X4) I(X1^3) I(X3^4)   
 0.483065688 0.143578928 0.010038614 -0.001912836

#### For model 3

Model 3: 𝑦 = 𝜃1 𝑥33 + 𝜃2 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

(Intercept) I(X3^3) I(X3^4)   
 0.340561975 0.021330543 -0.002857744

#### For model 4

Model 4: 𝑦 = 𝜃1 𝑥2 + 𝜃2 𝑥13 + 𝜃4 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

(Intercept) I(X2) I(X1^3) I(X3^4)   
 0.509013488 0.053322048 0.012067145 -0.001855997

#### For model 5

Model 5: 𝑦 = 𝜃1 𝑥4 + 𝜃2 𝑥12 + 𝜃3 𝑥13 + 𝜃4 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

(Intercept) X4 I(X1^2) I(X1^3) I(X3^4)   
 0.4798284629 0.1434460704 0.0003254641 0.0100562759 -0.0019198749

### Task 2.2: Calculating RSS (model Residual Sum of Squared error) for each model

RSS (Residual Sum of Square) is square sum of error of predict line and data in linear regression. Error is calculated by the difference in the distance of data and the regression line in a graph. In this task I again used the attribute of lm() command to calculate RSS. For this I squared model residuals and sum all the result to get RSS of the model.

#### For model 1

35.39663

#### For model 2

2.139762

#### For model 3

463.3124

#### For model 4

20.259

#### For model 5

2.135503

### Task 2.3 & 2.4 : Calculation AIC, BIC & log likelihood function

Log–likelihood function is defined as the logarithm of the likelihood function, which is a function that measures how well a given set of parameters of the statistical model fits the observed data. It ranges from infinity to negative infinity. The higher the value of the log-likelihood functions better a model fits the data. The log-likelihood function is typically denoted by the symbol "L" and is defined as:

L(θ | x) = log f(x | θ)

The AIC (Akaike information criterion) is a statistical function used to compare the fit of different models to a set of data. AIC is defined as follows:

AIC = -2 ln(L) + 2k

Where, L is the likelihood function of the model and k is the number of parameters in the model

AIC takes into account the number of parameters in the model (k) and the goodness of fit (ln(L)). Model having lower AIC value indicates that the model fit best to the data and is a good model complexity whereas the model having higher AIC value have a worse fit to the data and less likely to represent the true relationship between the variables in the data.

The BIC (Bayesian Information Criterion) is like a AIC which also compare the fit of different models to a set of data. BIC is defined as follows:

BIC = -2 ln(L) + k ln(n)

Where, L is the likelihood function of the model, k is the number of parameters in the model, and n is the number of observations in the data set.

BIC put more weight in the complexity of the model then AIC. The interpretation of BIC is similar to that of AIC. A lower BIC value indicates a better model fit, as it indicates that the model has a good balance between the fit to the data and model complexity. BIC values are always larger than AIC values for the same model, but the difference in values decreases as the sample size (n) increases.

For this task I used the glance() command to look for the AIC, BIC and log with is degree of freedom. This glance() command come from broom library. It shows different values such as AIC, BIC, r.squared, p.value. of linear regression models.

### For model 1

AIC = 235

BIC= 285

Log = -111

### For model 2

AIC = -333

BIC= -316

Log = 171

### For model 3

AIC = 746

BIC= 759

Log = -369

### For model 4

AIC = 119

BIC= 136

Log = -54.6

### For model 5

AIC = -331

BIC= -311

Log = 172

#### Task 2.5: Q-Q Plot of the model

For Q-Q plot of the model, I used plot() command. Plot command plot the different attributes of other function. In this plot command I plot the Q-Q by using “which=2” syntax.

### For model 1

### For model 2

#### 

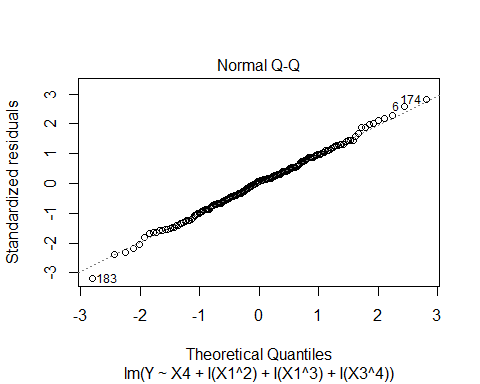
### For model 3

#### 

#### For model 4

#### 

#### For model 5



### Task 2.6: Choosing the best model

From the result data we came to find the best model is 2. It’s AIC, BIC are the lowest among all.

#### AIC = -333

#### BIC= -316

The lowest AIC, BIC means that the model has least amount of residual error. The Q-Q Plot also show the best fit of the data in the linear line. This model 2 represent the least error and bet fit the data set so we choose model 2 as the best model for further analyses.

Model 2: 𝑦 = 𝜃1 𝑥4 + 𝜃2 𝑥13 + 𝜃3 𝑥34 + 𝜃𝑏𝑖𝑎𝑠 + 𝜀

### Task 2.7: Training and testing the Model 2 from the

The selected model 2 was train for the data set given to test models. Dataset was split in 70-30% to train and to test the model. After training and testing the model the predicted data and real data have MSE (Mean Square Error) .0106601. Predicted data and actual test data show the linear positive relation.

0.0106601

### 

#### Calculation 95% confident level

From the predicted and actual dataset we calculate a two tailed test. Our null hypothesis state that “There is no significant different in the mean of both dataset”. The mean of the both data set are equal. The alternative hypothesis state that there is significant different in the mean of both dataset.

We choose the confident level of this t test is 95%.

Welch Two Sample t-test

data: dfstest$Y and dfstest$P

t = -0.045125, df = 119.96, p-value =

0.9641

alternative hypothesis: true difference in means is not equal to 0

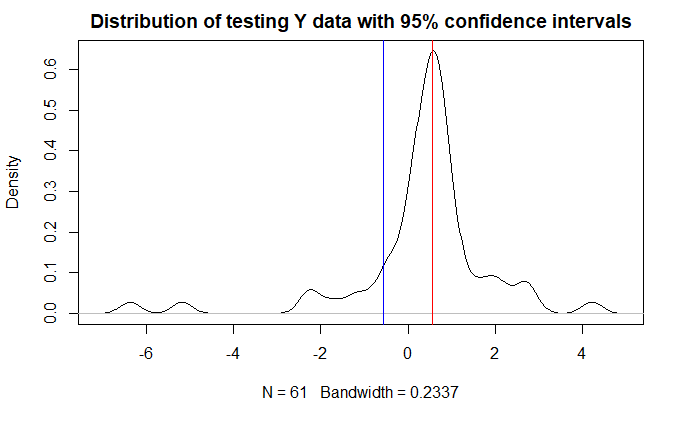
95 percent confidence interval:

-0.5748152 0.5491975

sample estimates:

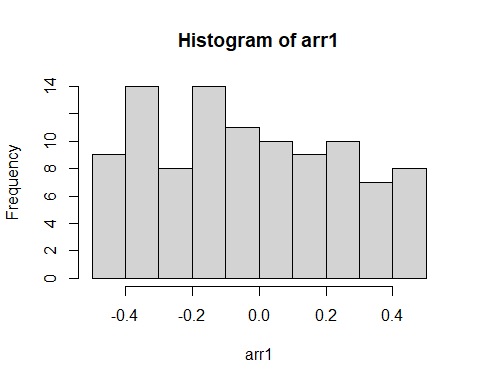
mean of x mean of y

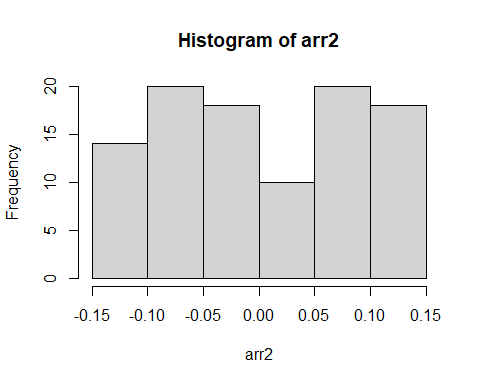
0.3367346 0.3495434

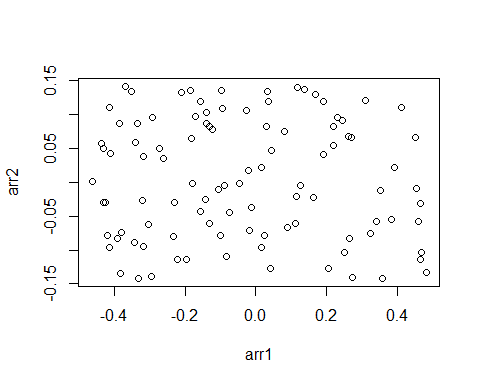
The test show that p-value is .9641 which is greater the 95% confident level so we reject the null hypothesis that state “There is no significant different in the mean of both dataset”

Task 3: Approximate Bayesian Computation (ABC)

ABC is a computational approach used in Bayesian statistics to estimate the posterior distribution of a likelihood function. The ABC algorithm compares the observed data with simulated data sets generated from a proposed model. The distance between the observed data and the simulated data is measured using a distance metric, and simulated data sets close enough to the observed data are saved for further analysis. The saved data sets are then used to estimate the posterior distribution of the parameter of interest.   
I conducted the rejection ABC in the selected model 2, for which I chose the two maximum posterior parameters, fixed other parameters and determined ranges of posterior parameters. I Looped it 100 times.  
Here arr1 and arr2 are the first and second maximum parameters. The result of the ABC is shown in the below diagram. The histogram of arr1 and arr2 and scatter plot of both the parameter.







# Conclusion

With R-lab, assignment given was completed to a satisfactory result. Model 2 represent the dataset, best. For improvement certain thing can be done.

1. Data given was insufficient of accurately model and to define the relationship. More data helps to improve the result
2. Less number of models in the assignment. More models can help to fit the data best.
3. Other method then least square method can be used to define the data better.

# References

<http://rpubs.com/SunilSapkota12/modelling_EEG> (All the code and output of this assignment)

An Introduction to Statistical Learning with Applications in R by Gareth James *•*Daniela Witten *•* Trevor Hastie. Robert Tibshirani

Learn R for Applied Statistics With Data Visualizations, Regressions, and Statistics by **Eric Goh Ming Hui**

Applied Statistics and Probability for Engineers by Douglas C. Montgomery, George C. Runger