

Statistical analysis of EEG signals

Dominik Klepl

Faculty of Engineering Environment and Computing

Coventry University

Coventry, UK

klepld@uni.coventry.ac.uk

1. Introduction

In this paper, the analysis and statistical modelling of two electroencephalogram (EEG) signals is reported and its results interpreted and discussed. We are working with two EEG signals, X and Y, that were measured brain activity of two regions. The goal of this analysis is to analyze the relationship between these signals and identify good model structure that can predict one signal from the other.

For the modelling purposes we refer to the signal X as an input (independent variable) and signal Y as output signal (dependent variable). In other words, signal Y is predicted by signal X.

First an exploratory data analysis was performed in order to acquire better understanding of the data. Next, two approaches were used to identify the best model structure; forward subset selection algorithm and AIC model selection. After the best model structure is identified, we validate the model using cross-validation to inspect its predictive power. Then we also quantify the uncertainty around both the model parameters and its predictions. Finally, rejection approximate Bayesian computation (ABC) is applied to estimate the posterior distribution of the model parameters.

The whole analysis was performed using R 3.6.1 Action of the Toes (ref).

1.1 Data

We are working with two separate EEG signals. The data sample consists of 250 data points. We assume that both signals are independent. Furthermore, we assume that the signals contain additive gaussian noise with zero mean and unknow variance.

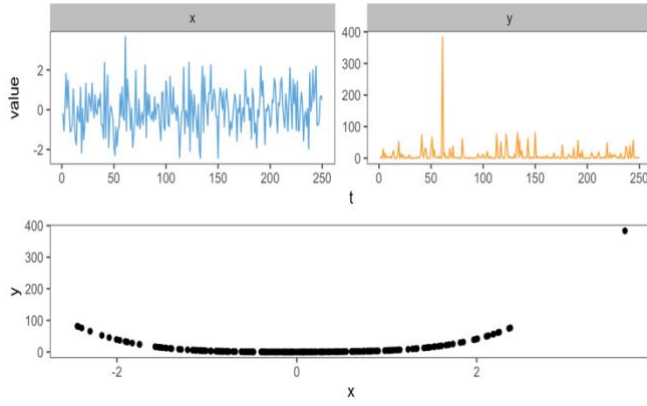


Figure 2 Top: Time series plots of signal X and Y.

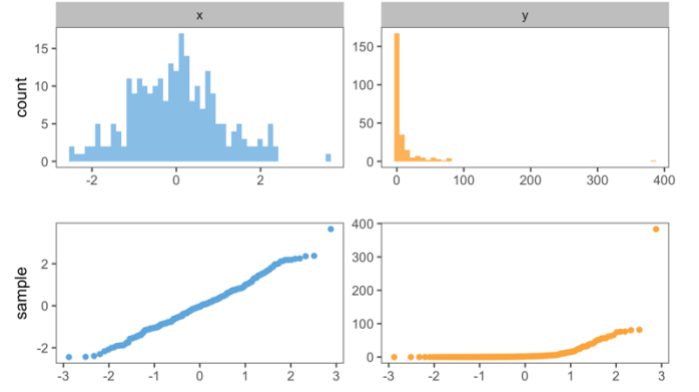


Figure 2 Top: Histograms of signal X and Y. Bottom: QQ-plots of signal X and Y

2. Exploratory data analysis

The aim of this initial exploratory analysis is to gain better understanding of the data and possibly observe some relationships between the signals.

We begin with time series plots of the signals, i.e. plot each signal on y-axis and time (order in which data were recorded) on x-axis (figure 1, top). We can see that the scale of the signals is very different. While signal X oscillates around 0 and approximate range -2 to 2, the signal Y seems to assume only in positive values with some spikes going to 50-80. One value seems to be extreme in both signals, especially in signal Y around 60th data point we observe a large spike that goes nearly to 400. It might be an outlier which would have large influence on model estimation.

Next, we plot the signals against each other in order to observe the relationship between them (figure 1, bottom). We can assume that the relationship between X and Y is nonlinear because the data form a wide parabola. The best model structure is likely to contain a quadratic term. We can again observe the extreme value (top right corner of the plot) that we saw in the time series plot.

Next, we inspect the distribution of the signals. The histograms of both signals are showed in figure 2 (top). We also create QQ-plots to see whether the data is normally distributed (figure 2, bottom). The QQ-plot is plotting quantiles of the data against quantiles of a theoretical normal distribution. Therefore, if the data follows normal distribution then the QQ plot will show a straight diagonal line.

Based on the histogram we might say that signal X is normally distributed, however the line in QQ plot deviates from the diagonal. To test whether signal is normally distributed we used Shapiro-Wilk test which tests the alternative hypothesis that data is not normally distributed with significance threshold 0.05. Results of this test suggest that signal X is normally distributed as we failed to reject the null hypothesis ($W = 0.993$, $p\text{-value} > 0.05$).

From the histogram of signal Y, we can clearly see that it is not normally distributed. Based on the shape it might be following exponential distribution but investigating this further is beyond the scope of this report. The QQ-plot confirms that signal Y is not normally distributed.

2.1 Linear model

From the scatterplot we drew conclusion that the relationship between X and Y is probably nonlinear. However, to explore all options we fitted a simple linear model (1).

$$y \sim \beta_1 x$$

(1)

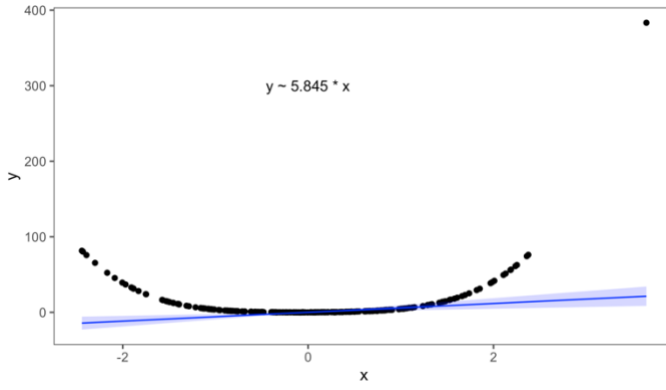


Figure 4 Linear model $y \sim x$, estimated parameters and predictions (blue) with 95% confidence interval (blue ribbon) plotted together with true y values (black points)

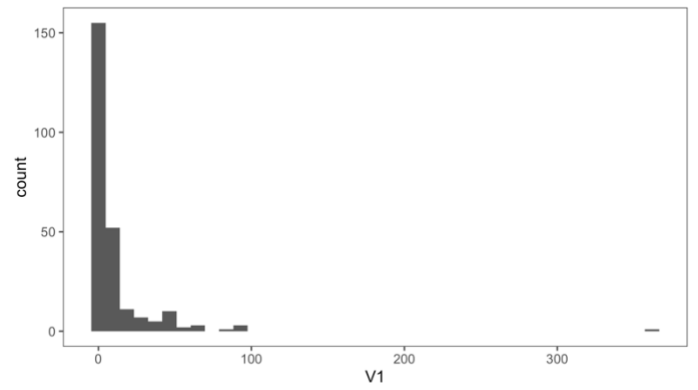


Figure 4 Distribution of residuals of the linear model $y \sim x$

To assess the performance of the fitted model we use mean squared error (MSE) which is calculated using formula given by (2). In the formula n is the number of data points, Y is the true Y value and the \hat{Y} is the model prediction. We also use adjusted R^2 defined in formula (3). SS_{residual} is the residual sum of squares, SS_{total} is the total sum of squares i.e. sample variance, df_i is the sample size - 1 and df_e is sample size - 1 - number of model predictors. Adjusted R^2 is often interpreted as percentage of observed variance explained by the model.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2)$$

$$R^2_{\text{adjusted}} = 1 - \frac{SS_{\text{residual}}/df_e}{SS_{\text{total}}/df_t} \quad (3)$$

Two measurements of the model performance are used in order to be able to assess both predictive (MSE) and explanatory (adjusted R^2) power of the model.

We use least squares to estimate the parameter of the model.

The results of this model are showed in figure 3. The MSE of the fitted model is 886.242. We also generated predictions and 95% confidence intervals. Next residual analysis was performed. In ideal case we would expect the residuals to be normally distributed with mean 0. However, residuals of this model are definitely not normal (figure 4). The histogram of the residuals is rather misleading as it suggests that most residuals are located near 0. This is caused by large range of the residuals (~301). In reality most of the residuals have quite high value with median 2.348.

We can conclude that linear model structure is rather bad fit as MSE is very high as well as residuals which are not normally distributed. Furthermore, from the fitted line (figure 3) we can see that the model captures only very small portion of the variance.

Model selection