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.

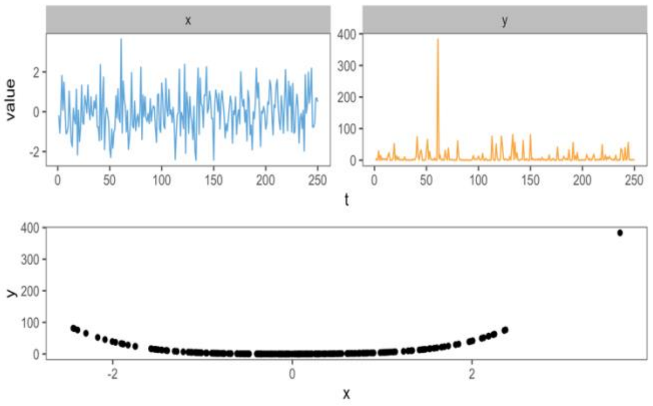

Figure 2 Top: Time series plots of signal X and Y.

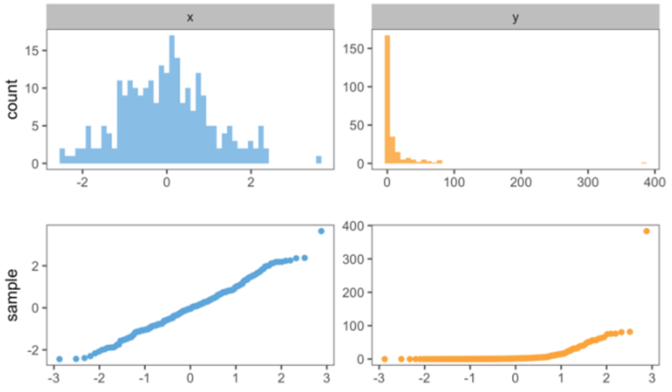

Figure 2 Top: Histograms of signal X and Y. Bottom: QQ-plots of signal X and Y

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-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).

$$y \sim \beta_1 x \tag{1}$$

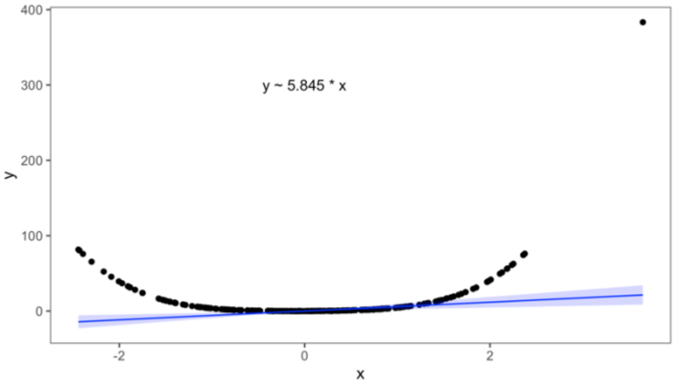

Figure 4 Linear model y ~ x, estimated parameters and predictions (blue) with 95% confidence interval (blue ribbon) plotted together with true y values (black points)

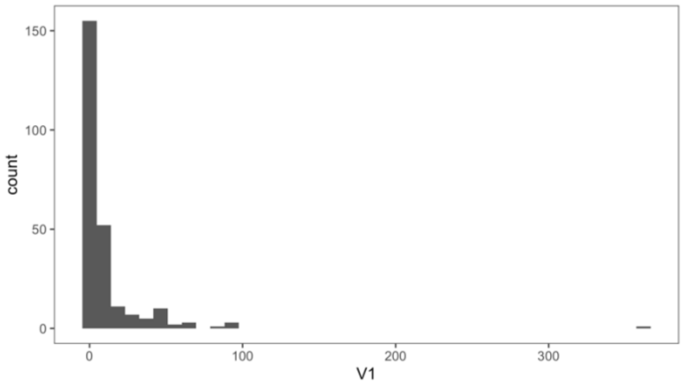

Figure 4 Distribution of residuals of the linear model y ~ x

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_t is the sample size - 1 and df_e is sample size – 1 – number of model predictors. Adjusted R_2 is can be interpreted as percentage of observed variance explained by the model.

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

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

Two measurements of the model performance are used in order to be able to asses 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 online very small portion of the variance.

3. Model selection

To identify the best model structure, two approaches were used. This was done to in order to have more evidence for the selected model. First approach is the forward subset selection and second model selection using Akaike information criterion (AIC).

Table 1: Results of forward subset selection

Model structure	Testing MSE
β_1x4	5.232
β_1x2 + β_2x4	0.308
β_1x + β_2x2 + β_3x4	0.008

Table 2: Parameter covariance matrix

	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$
$\boldsymbol{\beta_1}$	5.1e-05	1.0e-05	-2e-06
$\boldsymbol{\beta_2}$	1.0e-05	4.5e-05	-5e-06
$\boldsymbol{\beta_3}$	-2.0e-06	-5.0e-06	1e-06

3.1 Forward subset selection

Forward subset selection is iterative approach for finding the best model structure. First, the data is split to training and testing sets. We used 80% of data as training.

The training set is used to estimate the model parameters using least squares. Predictions are then generated only for the testing set and performance metric is calculated using these predictions. Performance metric is MSE (formula 2).

Then list of all possible terms are created. In our case this is X raised to power of up to 5 and intercept term. Then all single term models are estimated and their MSE computed. The model that yields the lowest MSE is selected, the parameter stored in the final model and removed from the list of possible terms. Next all two term models are fitted using the parameter of the final model and the remaining terms in the term list. Again, model with lowest MSE is selected and used as the final model. This process is repeated until satisfactory MSE is achieved or all terms are in the final model.

In our analysis, we limited the number of model terms to 3. The results of each iteration of the subset selection is showed in table 1.

3.2 AIC selection

AIC is intended to give an estimate of how close a fitted model is to the true model. The formula for calculating AIC is given in (4). It is based on the model likelihood and penalty for number of parameters.

$$AIC = 2k - 2\ln\left(L_{model}\right) \tag{4}$$

This approach does not require splitting the data, instead the models are fitted using all available data. Then we simply fit all possible combinations of terms and compute AIC for each of these models. The model that provides the lowest AIC is selected.

The best model structure according to AIC is:

$$y = \beta_1 x + \beta_2 x^2 + \beta_3 x^4 + \varepsilon \text{ (AIC = -432.326)}$$

3.3 Evaluation of model selection

Both approaches converged on the same model structure. Therefore, we can be quite confident that this is the best model structure given our data.

4. Model evaluation

The selected model which yields the lowest MSE is then evaluated. The estimated model using the training set is:

$$y = 0.499x + 2.004x^2 + 2x^4 + \varepsilon$$

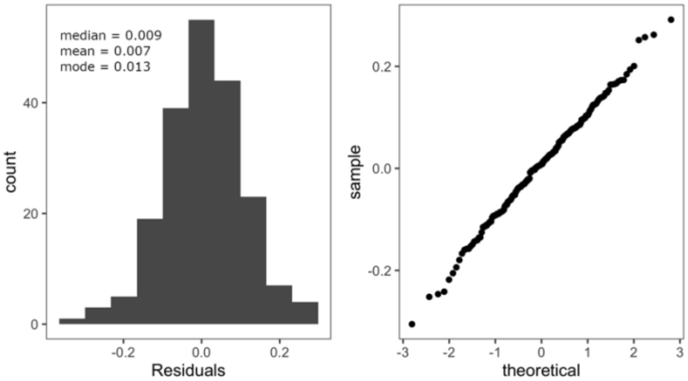

Figure 5 Residuals of the selected model with lowest testing MSE. Right: Histogram of residuals and central descriptive statistics. Left: QQ-plot of residuals

First, we inspect the residuals, especially their distribution. Histogram and QQ-plot are used to investigate the residual distribution (figure 4).

Next, the covariance matrix of the model parameters was calculated (table 2).