STAT 7670, Midterm

D. Pluta

The data provided is from a study interested in quantifying the genetic influence on cognitive activity related to emotionally stimulating images. The genetic data consists of a set of 1000 single nucleotide polymorphisms (SNPs). EEG activity for each subject was recorded over 1 second at 40 time evenly spaced time points while the subject viewed an emotionally stimulating image. Data on $n = 48$ subjects was collected.

The EEG data is contained in the file *EEG.csv*. The SNP data is contained in the file *SNP.csv*.

1. Plot the EEG data, with each time series color-coded by subject. Comment on any patterns or interesting features that you observe.

2. Compute the genetic similarity matrix and produce a heatmap representation. Describe any structure or features you see in the genetic similarity matrix.

3. Use the Mantel (or RV coefficient) test to determine if the data provide evidence of a significant association between a subject's genotype and EEG phenotype.

4. Summarize the EEG data for each subject as the mean of the EEG time series. Plot a histogram of the subject EEG means.

5. Fit a variance components model with the EEG means as response regressed on the full SNP data. Provide a point-estimate and 95% confidence interval for the proportion of EEG means explained by the SNP data.

6. Compute the principal components of the genetic data and produce a screeplot of the proportion of variance explained. Based on this plot, how many components should be retained according to the "elbow method".

7. Using the PCs selected in problem 6, fit a linear regression of the EEG means regressed on the genetic PCs. Is there evidence that the EEG means are significantly associated with the genetic PCs? What is the proportion of variance in the EEG means that is explained by the genetic PCs?

8. Write a brief summary of the results of your analysis in the context of the study setting, including a comparison of the PCA and linear regression approach versus the variance components analysis.