Analysis of EEG Data to study effects of alcoholism

Report by: ARKADEEP ADHIKARI

This report documents the findings from the UCI EEG data to study genetic predisposition to alcoholism. The data is based on recordings from 122 individuals across 64 channels from various parts of the head when a certain stimulus was presented to the subject.

Contact/Email:
+91-9163046728
arka.a92@gmail.com
7/19/2017

Contents

Introduction:

The data is recorded from a study to examine EEG correlates of genetic predisposition to alcoholism. There are 122 total subjects in the dataset and for each subject there is multivariate time series data for a certain number of trials. The data represents measurements from 64 electrodes placed on different part of the scalp and sampled at 256 Hz for 1 second.

In the experiment, there were 2 groups: alcoholic and control. Each subject was either shown a single stimulus S1 or two stimuli S1 and S2. These stimuli were pictures of objects from 1980 Snodgrass and Vanderwart picture set. In subjects where two stimuli were shown, it was either a matched or a non-matched scenario.

The untarred file of the eeg_full data has the following structure:

```
eeg_full/
   co2a0000364/
      co2a0000364.rd.000.gz
      co2a0000364.rd.002.gz
      ...
   ...
   co2c0000337/
      co2c0000337.rd.000.gz
      co2c0000337.rd.002.gz
      ...
   ...
```

Each subdirectory is a subject which has the trials data inside it. The 4th character in the subdirectory name (“a” or “c”) and also in the trial name indicates whether the subject is in alcoholic or in control group.

Each trial file has the following structure:

```
# [subject identifier]
# [experimental info]
# [sampling rate]
# [stimulus type], [trial #]
# [sensor position] chan [channel  #]
[trial #] [sensor position] [sample #] [sensor value]
[trial #] [sensor position] [sample #] [sensor value]
…
For example:
# co2a0000364.rd
# 120 trials, 64 chans, 416 samples 368 post_stim samples
# 3.906000 msecs uV
# S1 obj , trial 0
# FP1 chan 0
0 FP1 0 -8.921
0 FP1 1 -8.433
0 FP1 2 -2.574
0 FP1 3 5.239
0 FP1 4 11.587
0 FP1 5 14.028
...
```

We use the pickle library in python to parse the dataset and store it on the disk. The pickle module in python implements an algorithm for serializing and de-serializing a python object structure. Pickle has two main methods. The first one allows us to dump the data to a file object. The second one loads the data from a file object. Using pickle we serialize each of the trial object to the disk. The data is coerced in a pandas dataframe and stored in as csv dump. Then these individual trials can be merged together to get the data for each subject.

We also use the modules of gzip and os. Module gzip is used to unzip each of the trial files to get the trial data. Module os helps us to set our working directory very easily by providing the necessary path.

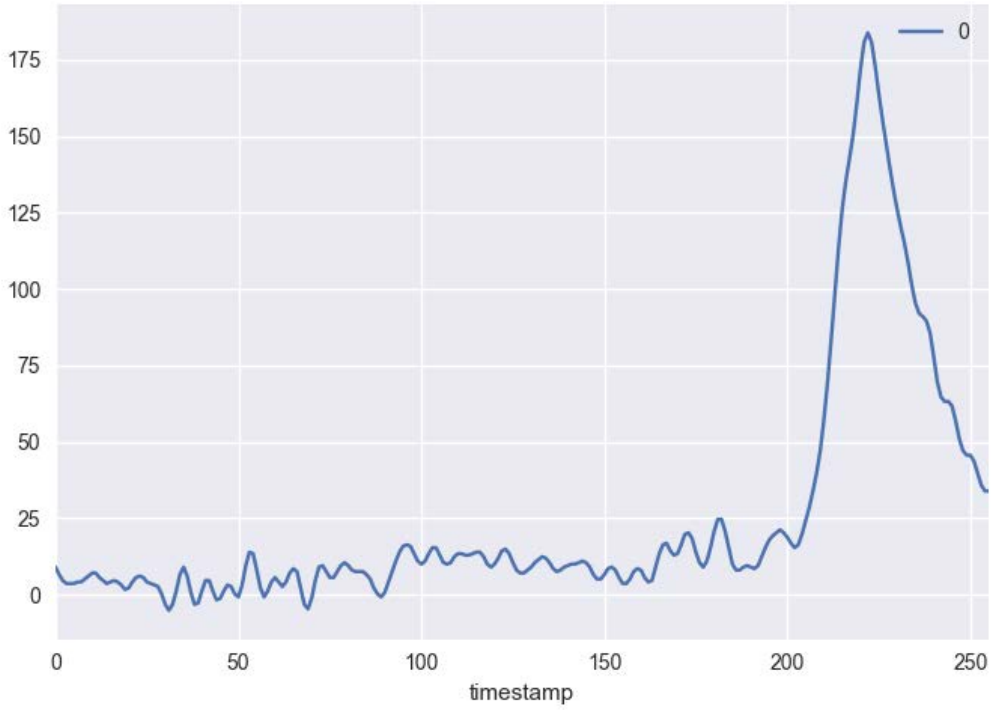
To perform data analysis we mainly use pandas. Pandas is well suited for dealing with tabular data and provides fast and flexible data structures for doing practical real world data analysis using python. For some purposes in our analysis, while dealing with lists or matrices we use numpy. NumPy is a python package which mainly deals with scientific computations in a very simple manner.

For plotting and data visualization we mainly use matplotlib and seaborn. Both of these libraries provide a high level interface for generating various visualizations. Using both these libraries, it is possible to generate plots, histograms, heatmaps, bar charts, scatterplots, etc., with just a few lines of code.

We also have used networkx to express correlations using graphs in this report. NetworkX is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. Using networkx we have tried to generate graphs using the correlations between different channel readings.

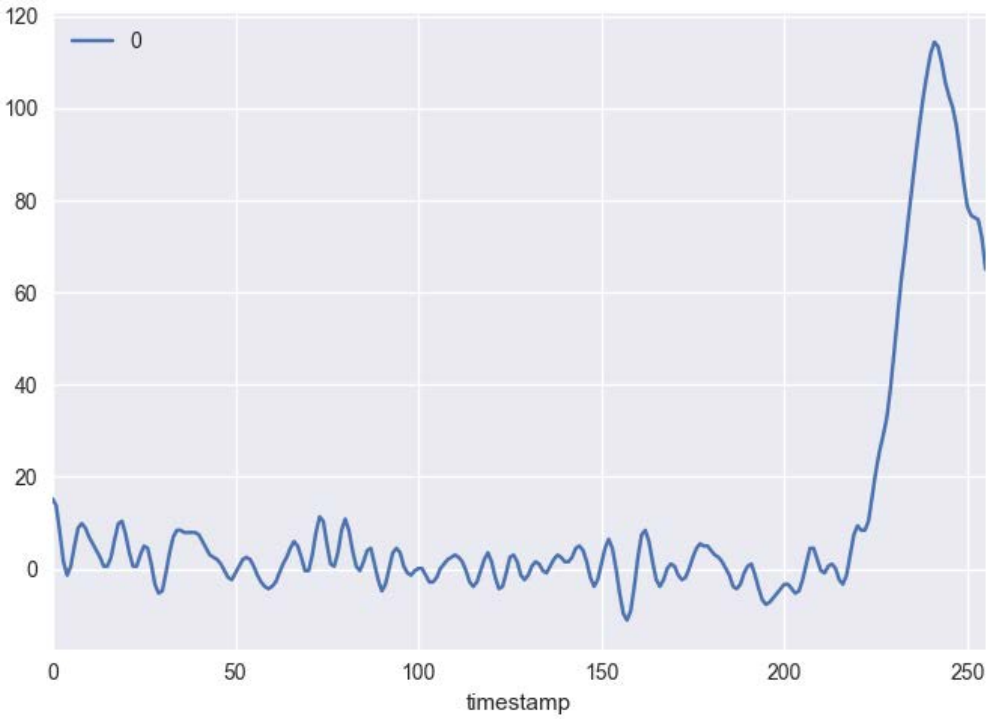
Exploratory Data Analysis:

We start by analysing the variations in the voltage recorded by one channel under a given stimulus over time.
For instance, let us look at channel 0 data for subject co2a0000364 over time. Clearly, this subject is an alcoholic as we can see from its 4th letter of the subject name. This subject under trial number 2 experiences S1obj as the stimulus. Its recording can be viewed as below:

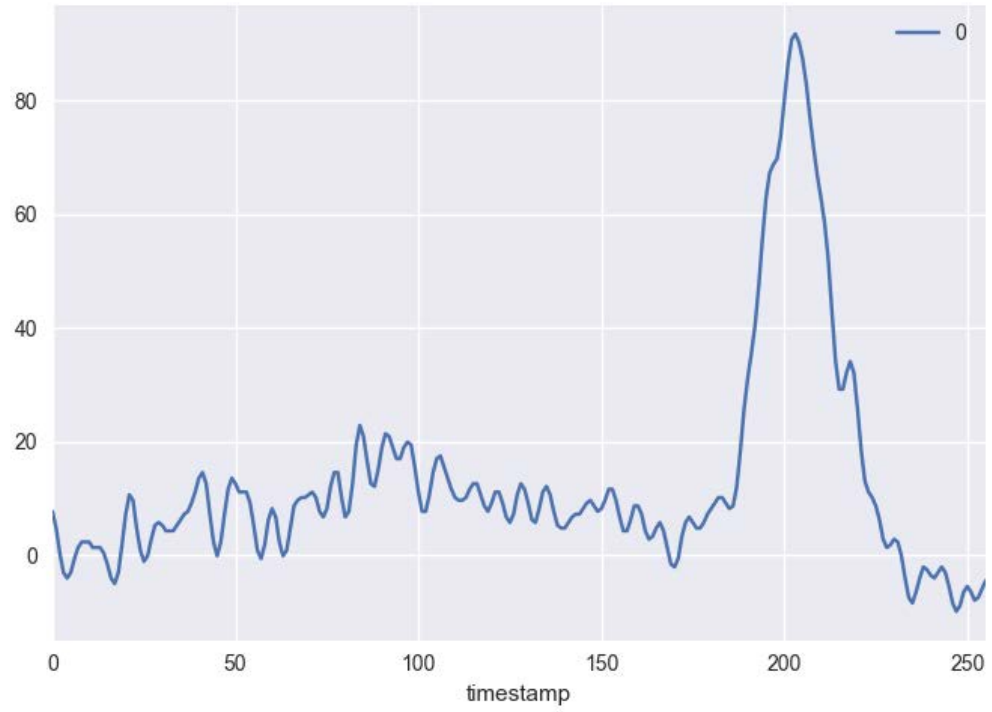


Now, let us look at the pattern of responses recorded by the same channel (channel 0) under the two remaining stimuli to get a better understanding of the data.

The time-series plot below shows the recording of channel 0 under trial 17 under the stimulus of S2match:

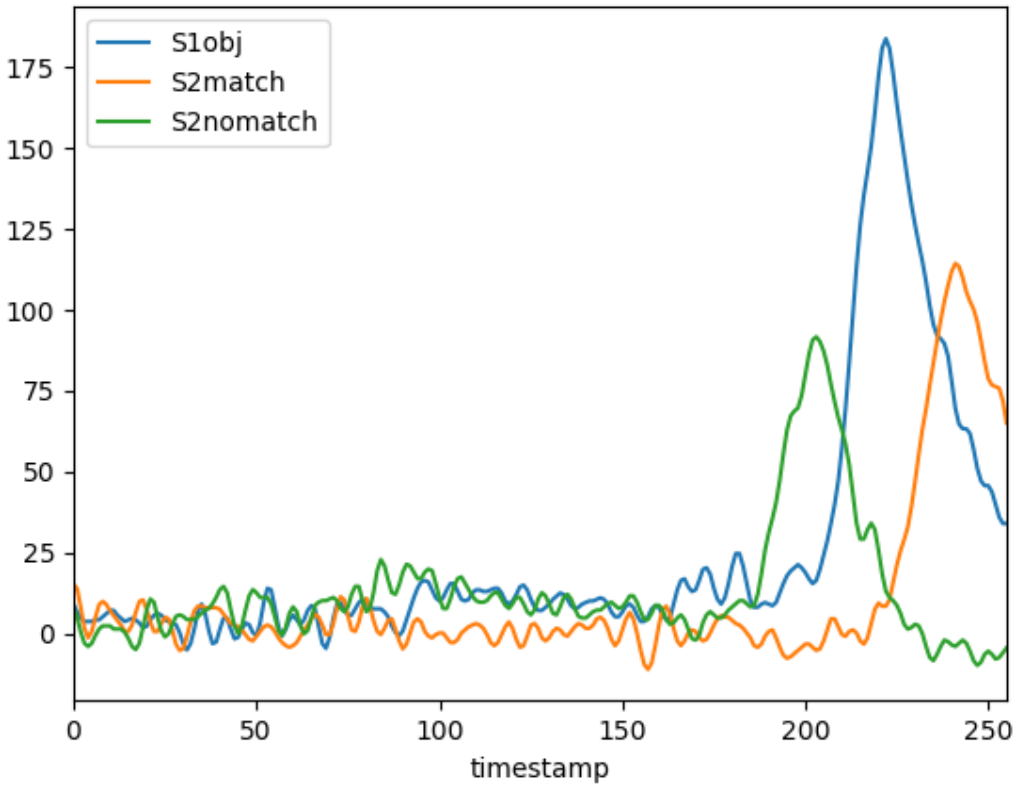


The timeseries plot below shows the recording of channel 0 under trial 7 under the stimulus of S2nomatch:



Impact of the Stimulus on the observations:

We may want to analyse the differences in the effect of the 3 stimuli on a certain subject with respect to a selected channel (channel 0 taken here). Here, we combine the reports made by the individual cases discussed above, i.e we pivot the data with the 3 trials considered above for the 3 different stimuli. The results obtained can be seen below:

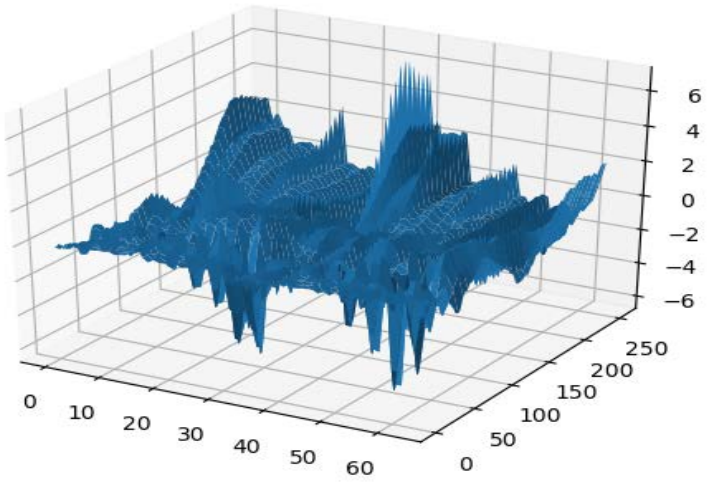


Plotting multiple channels using Surface Plots

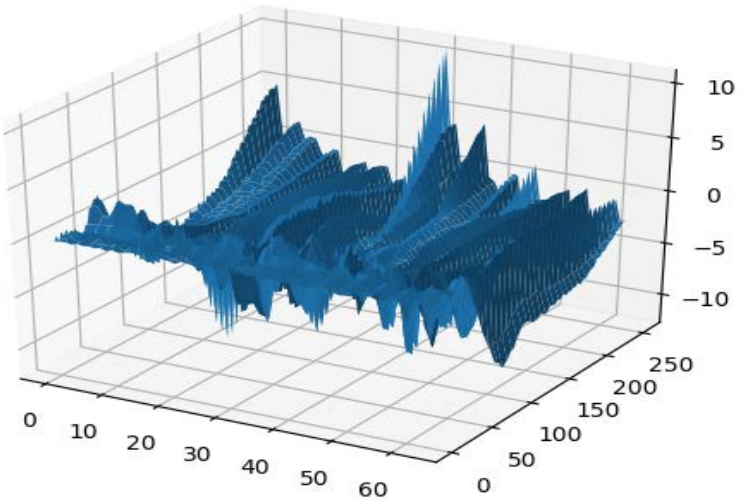
To have a look into multiple channels at the same time, we can try to visualize them as surface plots. These plots can be very easily generated using Axes3D of matplotlib in python. In the plots below, the average voltages across each channel has been considered for the various stimuli given in the two groups of subjects.

In the following plots x represents the channels, y gives us the timestamp and z gives the average voltage recorded.

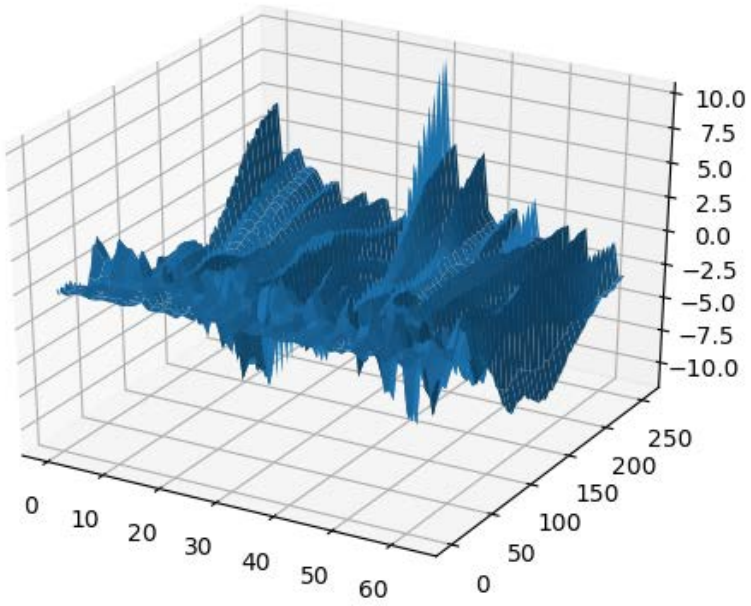
Alcoholic groups with S1obj stimulus:



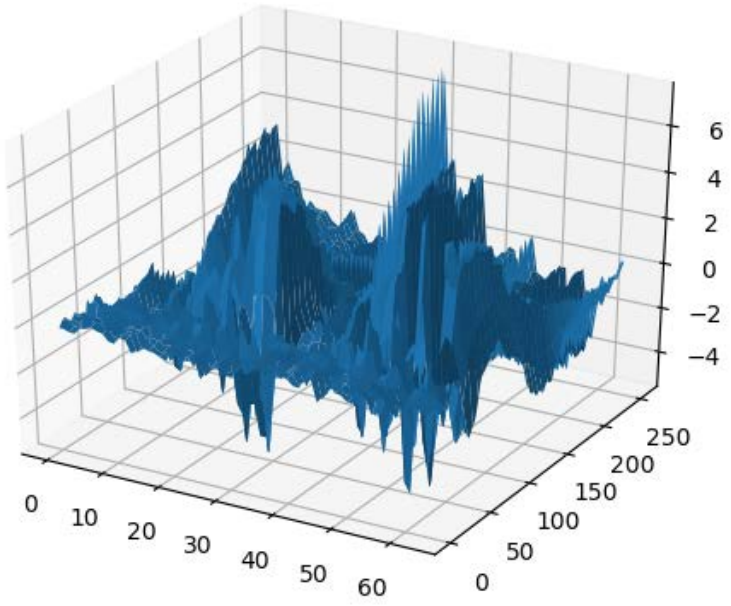
Alcoholic groups with S2match stimulus:



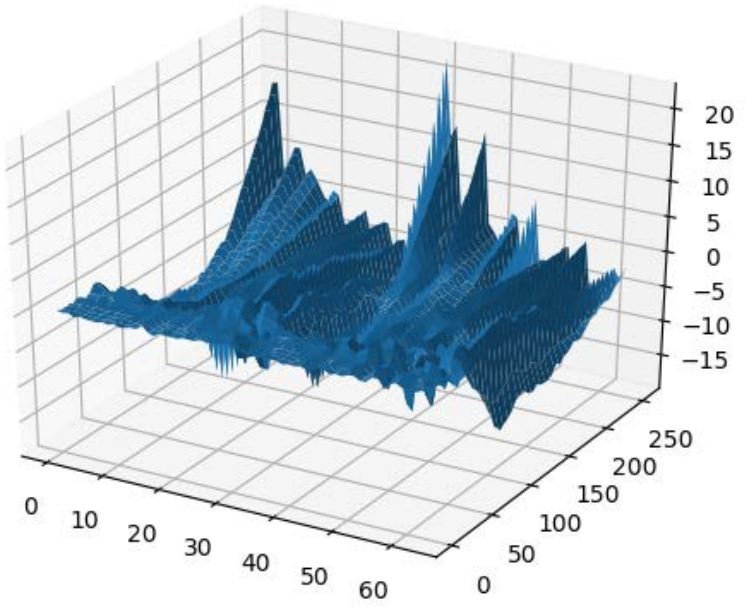
Alcoholic groups with S2nomatch stimulus:



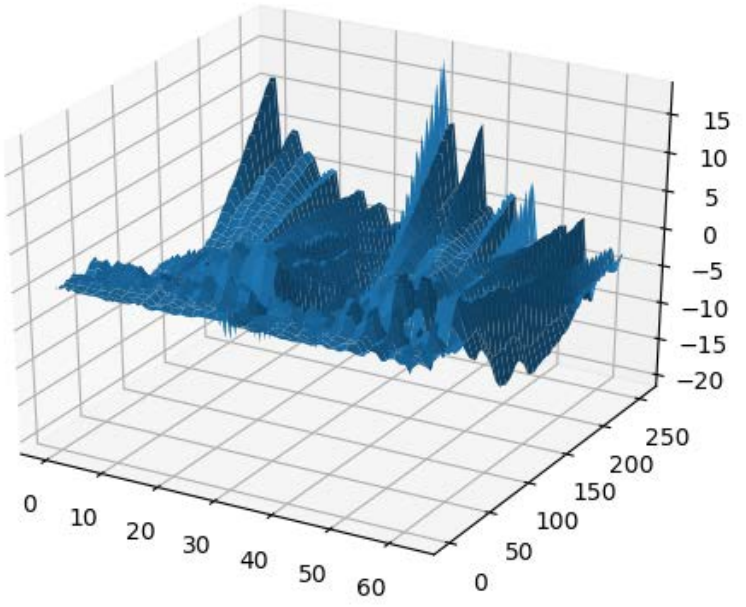
Control groups with S1obj stimulus:



Control groups with S2match stimulus:



Control groups with S2nomatch stimulus:

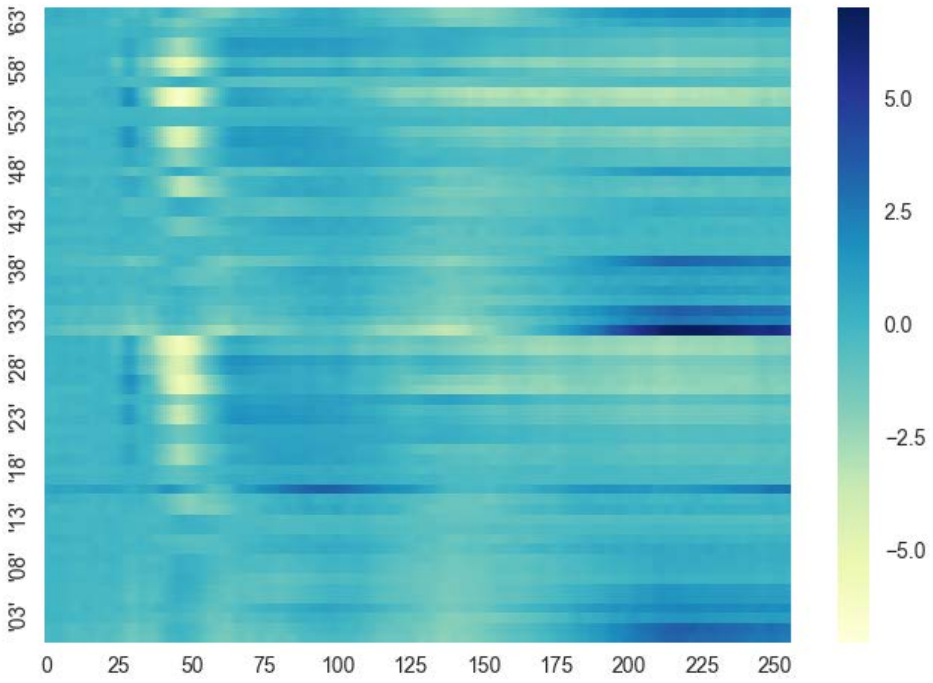


Analysis of Variation of Data across all Channels using Heatmaps:

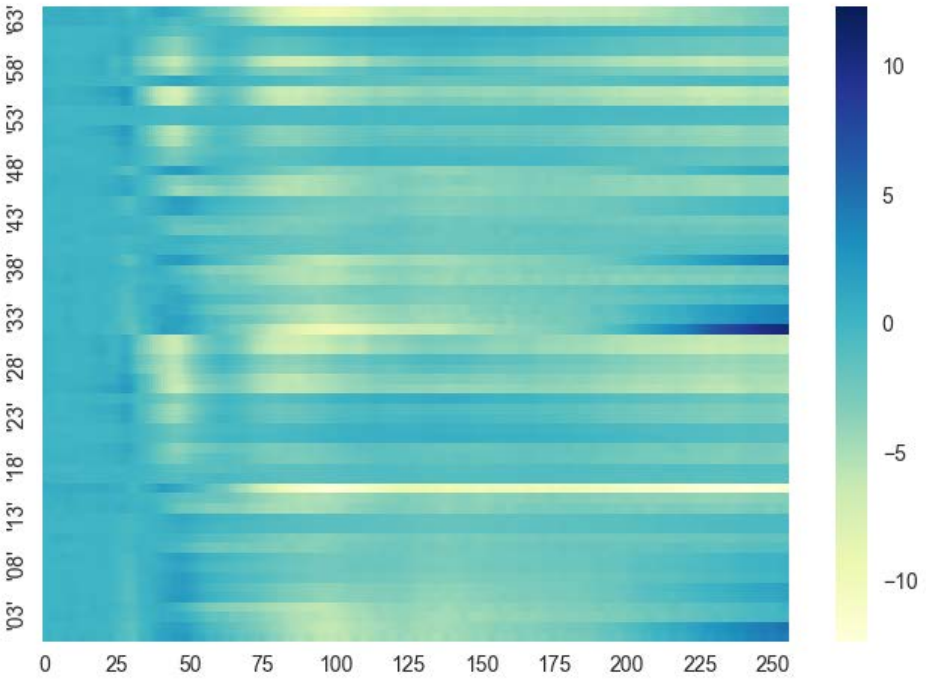
In this analysis, we will look at the average variations of the voltages recorded across all channels and how these changes vary based on the stimulus given to that group. We can see how these changes happen over time across different channels very easily using a heatmap. The module of seaborn in python greatly helps us to generate the heatmaps below.

In the given plots, timestamp in taken along x and the channels are represented along y:

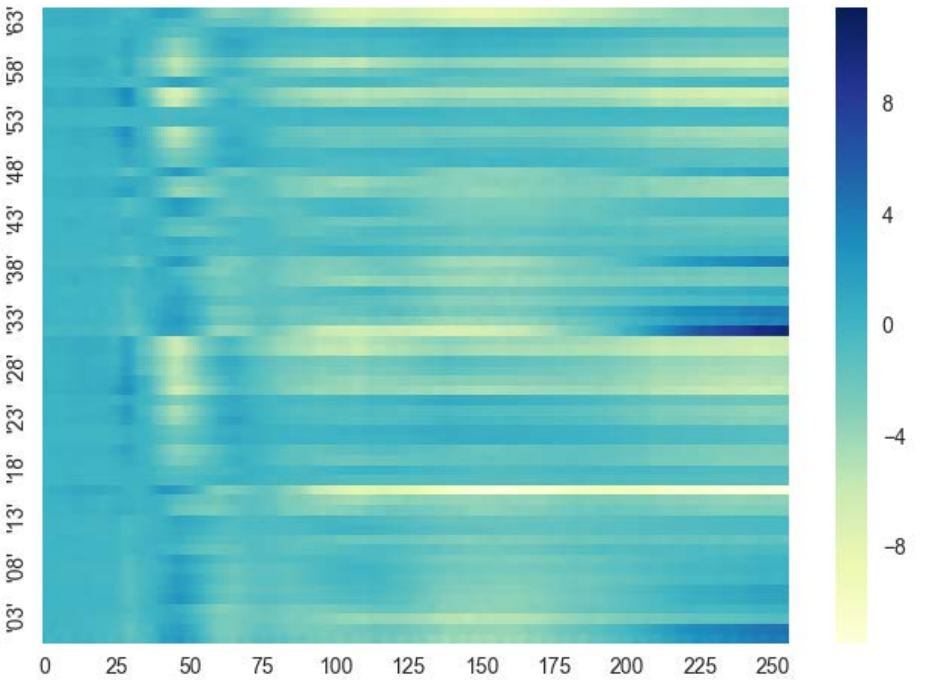
Alcoholic groups with S1obj stimulus:



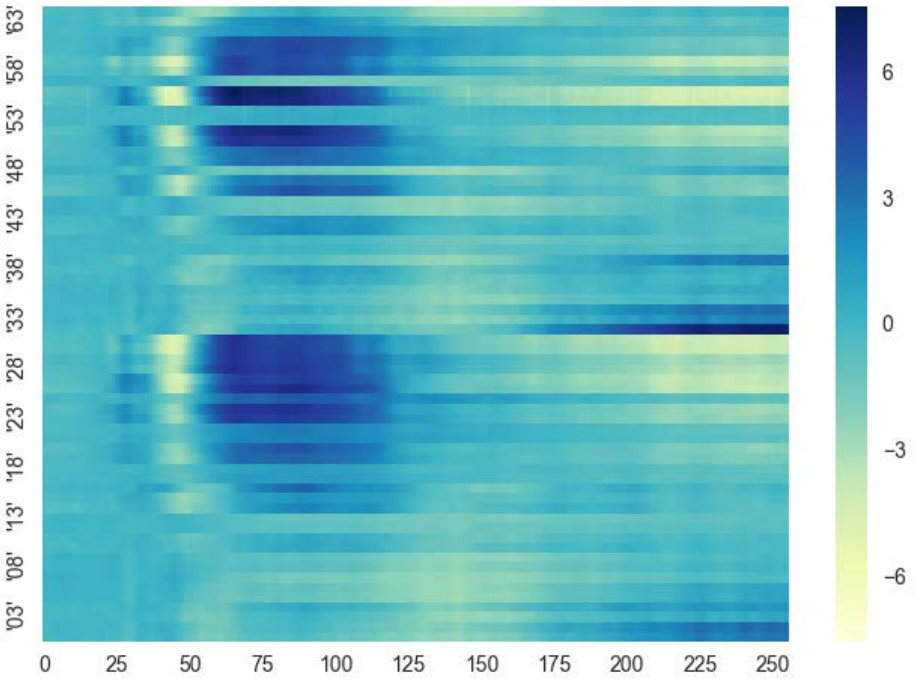
Alcoholic groups with S2match stimulus:



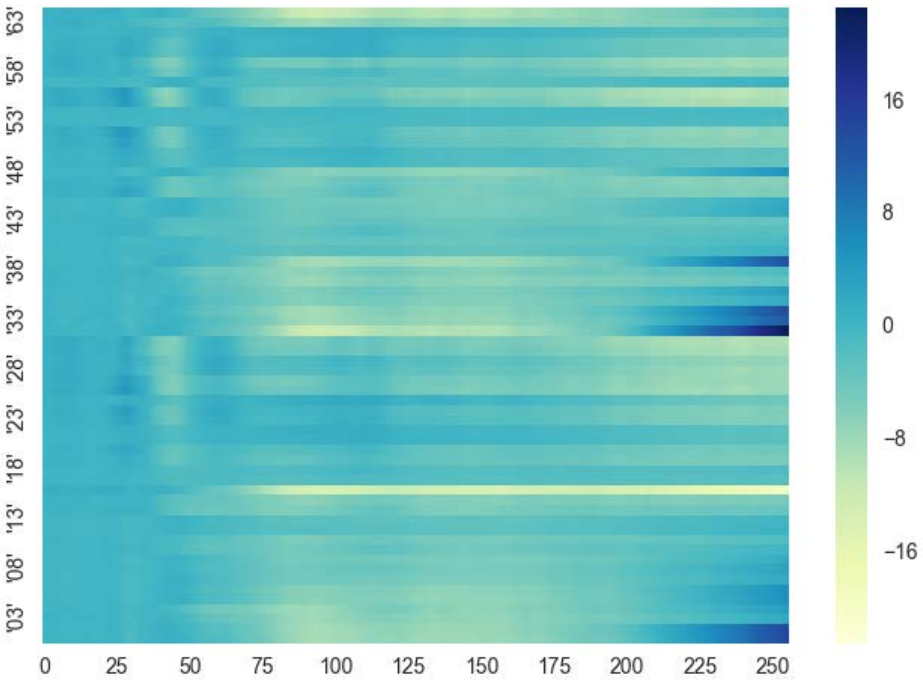
Alcoholic groups with S2nomatch stimulus:



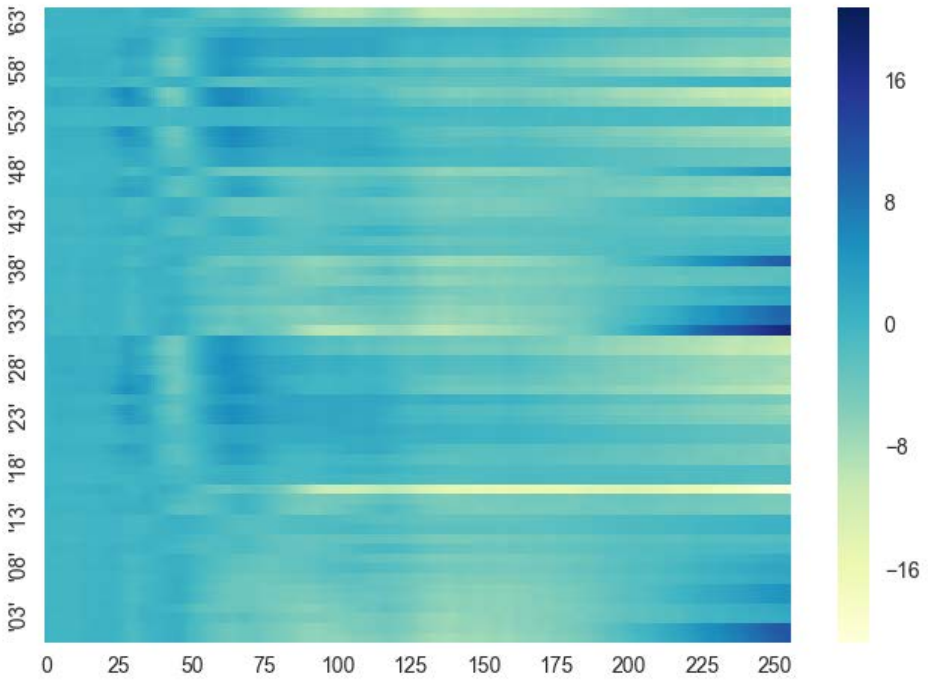
Control groups with S1obj stimulus:



Control groups with S2match stimulus:



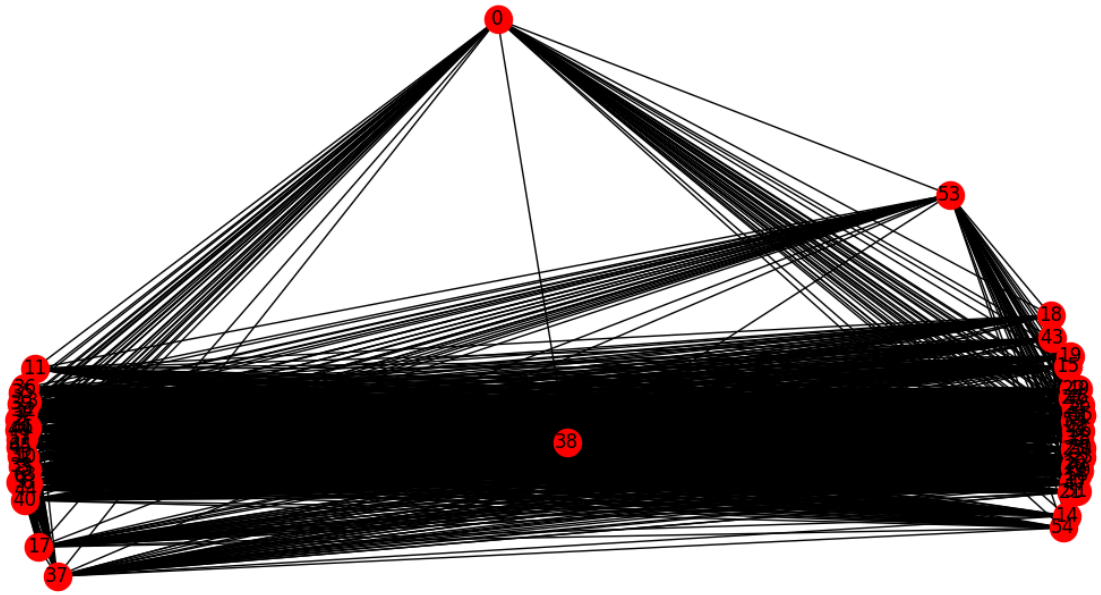
Control groups with S2nomatch stimulus:



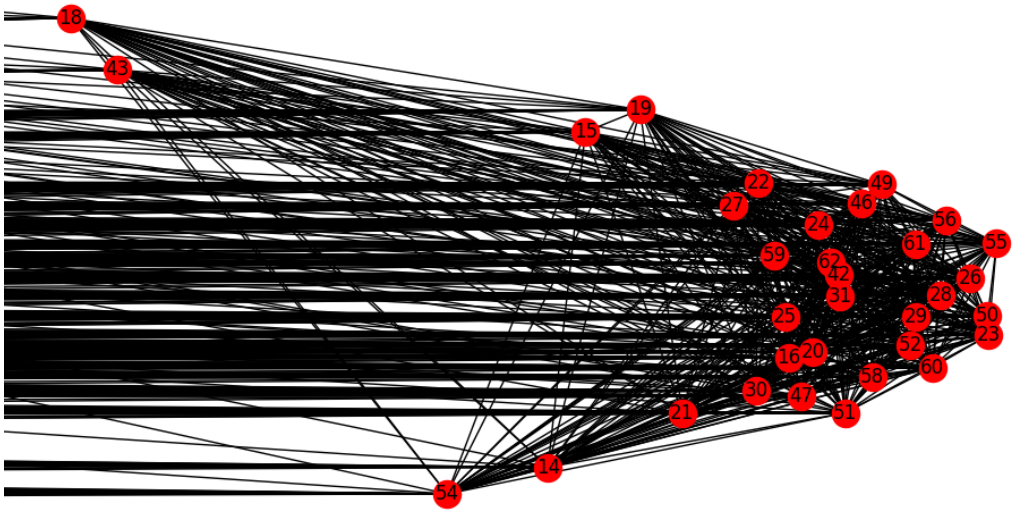
From the above plots, we observe a striking difference between the voltages recorded in control S1obj with the other two stimuli given to control groups for channels in the ranges of 18 to 30 and 45 to 60. These differences, however, are not recorded as such in the case of alcoholic subjects. Also, for the S2match and S2nomatch stimuli, the overall ranges of values of voltages recorded for control subjects were greater than that for alcoholics.

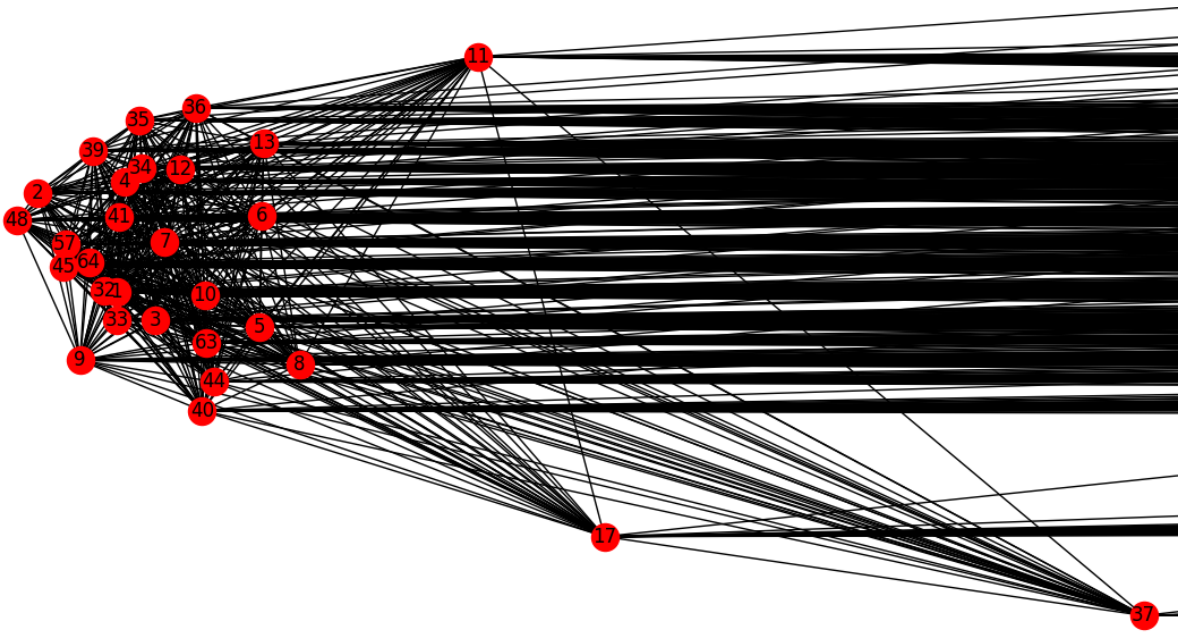
Finding cross-correlations between channels:

For this analysis, we take data from trial 2 (with S1obj stimulus) of subject co2a0000364 to check the correlations. It comes out that the results can be seen in two main clusters of channel numbers. We can easily plot the data as a graph using the networkx library in python. The result is found to be as below:



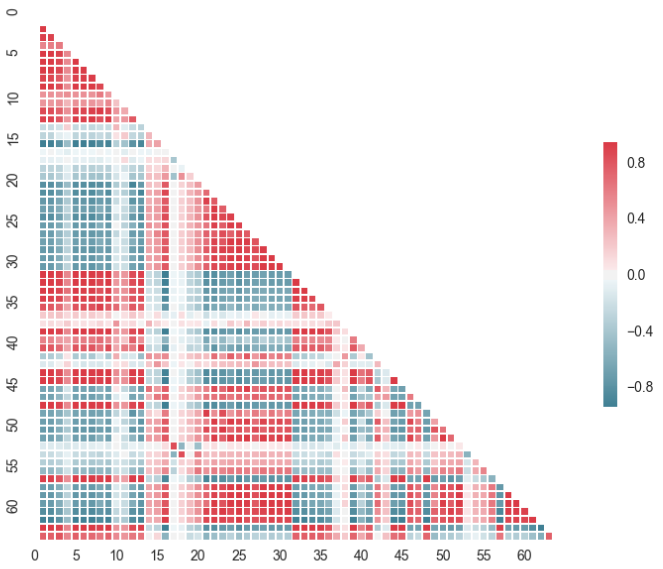
The two crowded clusters can be enlarged as below:



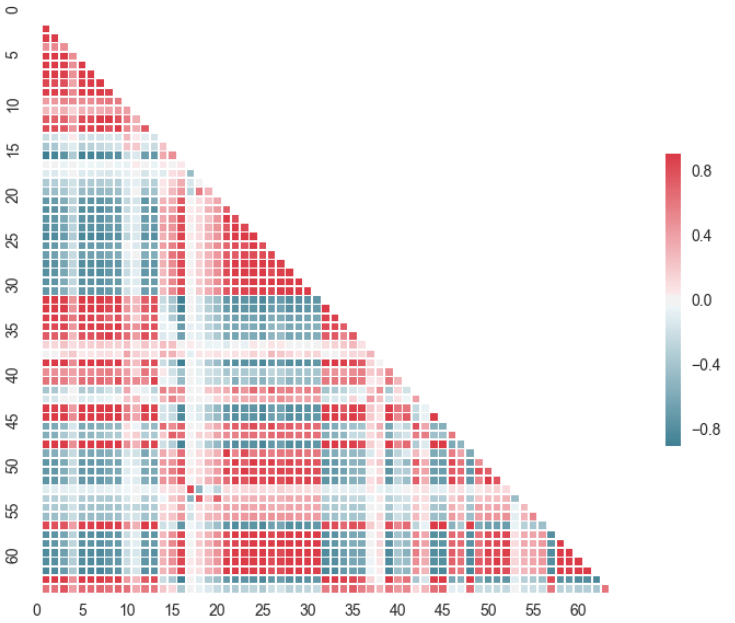


These channels appear to be similarly correlated for different recordings for the same stimulus except for a few rare scenarios.

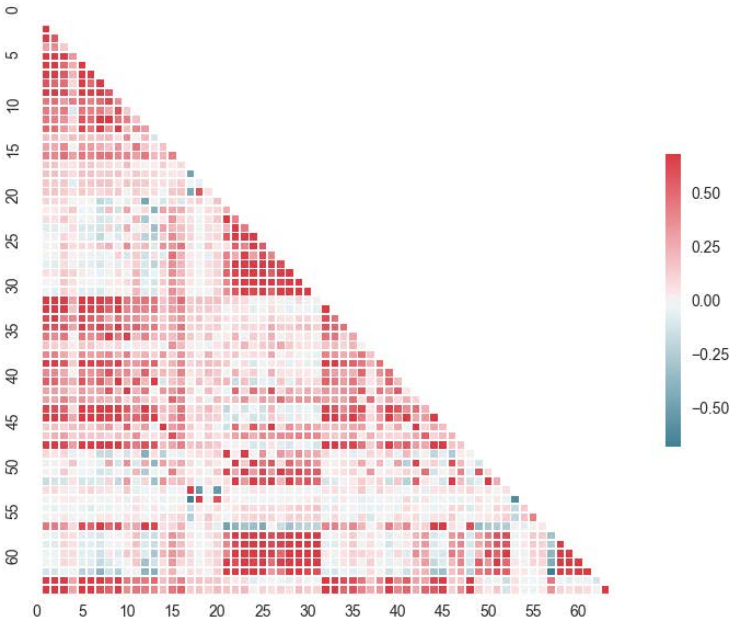
A similar comparison can be made using a correlation triangle.
Analysing a few trials for stimulus S1obj we get the following results:



For Trial no. 2



For Trial no. 10



For Trial no. 12

Significance Testing:

We test for statistical significance in the groups of alcoholics and control. In this we take the mean of the channel voltages of the alcoholics and the control subjects for various stimuli and study them. The following results were obtained:

For Stimulus S1obj:

channel	p	t	Label
'00'	0.060084	1.884393	same
'01'	0.55296	0.593726	same
'02'	3.12E-08	5.621551	diff
'03'	0.000295	3.645425	diff
'04'	5.05E-16	8.390632	diff
'05'	2.29E-14	7.867994	diff
'06'	7.40E-25	11.02611	diff
'07'	5.74E-17	8.714255	diff
'08'	1.04E-31	12.71793	diff
'09'	1.23E-08	5.79108	diff
'10'	2.11E-05	4.295966	diff
'11'	3.43E-36	14.05732	diff
'12'	5.27E-25	11.09446	diff
'13'	0.011038	-2.55282	diff
'14'	0.005316	-2.80188	diff
'15'	1.62E-19	9.438619	diff
'16'	0.292673	-1.05372	same
'17'	0.143026	-1.46781	same
'18'	7.07E-05	-4.0219	diff
'19'	2.86E-06	-4.76207	diff
'20'	7.88E-08	-5.48168	diff
'21'	5.74E-07	-5.09585	diff
'22'	4.77E-06	-4.64798	diff
'23'	5.15E-08	-5.5741	diff
'24'	1.57E-14	-8.01112	diff
'25'	2.01E-09	-6.16263	diff
'26'	1.41E-05	-4.40411	diff
'27'	2.27E-07	-5.28388	diff
'28'	8.60E-06	-4.516	diff
'29'	1.20E-07	-5.41041	diff
'30'	6.25E-05	-4.05376	diff
'31'	0.001262	-3.24297	diff
'32'	0.360363	0.915501	same
'33'	0.697087	0.389477	same
'34'	4.85E-12	7.086946	diff
'35'	1.06E-10	6.595525	diff

channel	p	t	Label
'36'	0.000206	3.739023	diff
'37'	0.000914	3.33548	diff
'38'	0.022952	2.28116	diff
'39'	0.001177	3.264901	diff
'40'	6.89E-21	9.894973	diff
'41'	7.29E-07	-5.0376	diff
'42'	0.0022	-3.08269	diff
'43'	1.85E-20	9.765863	diff
'44'	9.05E-28	11.84131	diff
'45'	7.86E-12	-7.08394	diff
'46'	0.000481	-3.52382	diff
'47'	8.76E-23	10.33492	diff
'48'	7.57E-11	-6.70788	diff
'49'	2.00E-11	-6.93371	diff
'50'	1.05E-07	-5.43082	diff
'51'	6.79E-09	-5.94661	diff
'52'	0.133015	1.505423	same
'53'	2.66E-08	-5.68172	diff
'54'	2.64E-07	-5.25058	diff
'55'	3.12E-08	-5.66396	diff
'56'	9.07E-11	6.669087	diff
'57'	1.15E-05	-4.45426	diff
'58'	1.95E-06	-4.84394	diff
'59'	4.54E-11	-6.80791	diff
'60'	7.76E-10	-6.32443	diff
'61'	4.83E-10	-6.40781	diff
'62'	0.000283	3.660121	diff
'63'	5.52E-19	9.270523	diff

For Stimulus S2match:

channel	p	t	Label
'00'	0.034898	2.11728	diff
'01'	0.075352	1.783203	same
'02'	5.43E-08	5.534788	diff
'03'	9.25E-05	3.948294	diff
'04'	4.94E-09	5.985156	diff
'05'	5.80E-08	5.53354	diff
'06'	5.36E-12	7.101157	diff
'07'	4.30E-10	6.410603	diff
'08'	8.33E-21	9.897295	diff
'09'	4.57E-09	5.983411	diff
'10'	1.77E-20	9.748097	diff
'11'	5.37E-10	6.353975	diff
'12'	8.84E-19	9.255195	diff
'13'	9.07E-10	6.249705	diff
'14'	2.81E-17	8.794144	diff
'15'	6.32E-07	5.049202	diff
'16'	5.58E-33	13.07627	diff
'17'	1.75E-09	6.151331	diff
'18'	6.49E-14	7.769954	diff
'19'	0.000475	3.522837	diff
'20'	0.046963	1.992821	diff
'21'	0.450285	-0.75572	same
'22'	0.000685	3.42397	diff
'23'	0.038511	2.076918	diff
'24'	0.752431	-0.31566	same
'25'	0.157304	1.416649	same
'26'	0.003528	2.933637	diff
'27'	0.018087	2.373979	diff
'28'	0.018062	2.374123	diff
'29'	0.540004	0.613275	same
'30'	0.003654	2.922796	diff
'31'	0.165907	1.387863	same
'32'	0.004597	2.850776	diff
'33'	0.107716	1.61248	same
'34'	3.00E-16	8.535783	diff
'35'	7.30E-09	5.916155	diff

channel	p	t	Label
'36'	4.17E-15	8.133171	diff
'37'	1.34E-08	5.785265	diff
'38'	0.007422	2.691811	diff
'39'	2.14E-07	5.27168	diff
'40'	1.56E-21	10.06853	diff
'41'	1.57E-06	4.863797	diff
'42'	7.25E-16	8.370602	diff
'43'	3.94E-07	5.153135	diff
'44'	1.69E-12	7.277978	diff
'45'	0.011027	2.552027	diff
'46'	9.79E-11	6.626109	diff
'47'	6.20E-10	6.345568	diff
'48'	1.39E-07	5.360335	diff
'49'	0.086512	1.718424	same
'50'	0.000434	3.548655	diff
'51'	0.007896	2.670532	diff
'52'	3.01E-38	14.4581	diff
'53'	0.000304	3.643466	diff
'54'	0.030838	2.166582	diff
'55'	0.009105	2.620398	diff
'56'	5.01E-10	6.354195	diff
'57'	2.77E-06	4.756369	diff
'58'	0.053199	1.938328	same
'59'	0.462126	0.73612	same
'60'	0.019362	2.348544	diff
'61'	0.000329	3.62227	diff
'62'	6.51E-08	5.486391	diff
'63'	2.98E-05	4.215582	diff

For stimulus S2nomatch:

channel	p	t	Label
'00'	0.001171	3.271295	diff
'01'	0.003836	2.909384	diff
'02'	0.147831	1.449769	same
'03'	9.50E-12	7.035976	diff
'04'	6.80E-10	6.335493	diff
'05'	9.67E-10	6.279174	diff
'06'	5.30E-14	7.81406	diff
'07'	1.16E-06	4.943927	diff
'08'	1.96E-13	7.611316	diff
'09'	2.12E-12	7.257571	diff
'10'	2.27E-11	6.862697	diff
'11'	5.08E-06	4.622035	diff
'12'	2.74E-10	6.464348	diff
'13'	8.59E-08	5.449855	diff
'14'	0.00659	2.728784	diff
'15'	3.13E-10	6.434711	diff
'16'	0.097865	1.658859	same
'17'	4.24E-06	4.671926	diff
'18'	0.918525	0.102359	same
'19'	0.012818	-2.50127	diff
'20'	0.789188	-0.26758	same
'21'	0.257065	-1.13527	same
'22'	0.068008	-1.83063	same
'23'	0.062804	-1.86661	same
'24'	0.077744	-1.76924	same
'25'	0.000885	-3.35111	diff
'26'	0.000984	-3.32032	diff
'27'	0.002293	-3.07113	diff
'28'	0.026003	-2.23543	diff
'29'	0.113736	-1.58518	same
'30'	0.114774	-1.58068	same
'31'	0.736101	0.337233	same
'32'	0.004559	2.853573	diff
'33'	3.94E-07	5.165539	diff
'34'	3.00E-07	5.212166	diff
'35'	1.99E-13	7.640386	diff

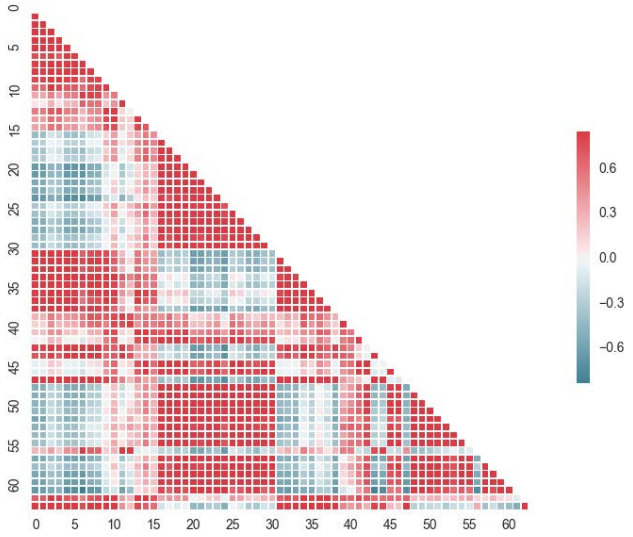
channel	p	t	Label
'36'	1.19E-06	4.923798	diff
'37'	6.66E-10	6.318241	diff
'38'	0.009172	2.61917	diff
'39'	6.44E-15	8.127422	diff
'40'	3.62E-10	6.410088	diff
'41'	0.011265	2.546618	diff
'42'	3.14E-05	4.205785	diff
'43'	8.18E-09	5.895846	diff
'44'	9.25E-13	7.380579	diff
'45'	0.618364	-0.49856	same
'46'	0.718748	-0.36036	same
'47'	2.45E-10	6.509898	diff
'48'	0.062124	1.87168	same
'49'	0.201633	-1.27938	same
'50'	0.034611	-2.12112	diff
'51'	0.044165	-2.01976	diff
'52'	9.55E-08	5.446174	diff
'53'	0.002863	3.004953	diff
'54'	0.024627	-2.25653	diff
'55'	0.025398	-2.24418	diff
'56'	5.06E-08	5.550668	diff
'57'	0.357191	-0.92193	same
'58'	0.179274	-1.34536	same
'59'	0.008559	-2.64484	diff
'60'	0.095588	-1.67135	same
'61'	0.047065	-1.99292	diff
'62'	0.033386	2.133339	diff
'63'	0.004035	2.88924	diff

*The above similarity and dissimilarity of means have been assumed at a confidence level of 95%.

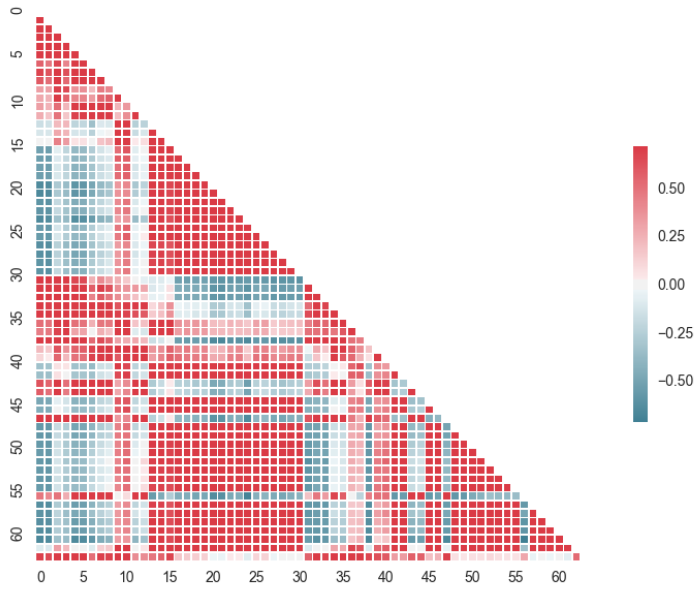
Studying for correlations in Alcoholic and Control Group:

Here we would consider the mean values of recorded voltages across channel over time for the alcoholic and the control groups. Using this data of average voltages, we can easily compute the correlation triangle using seaborn library in python to study the various correlations.

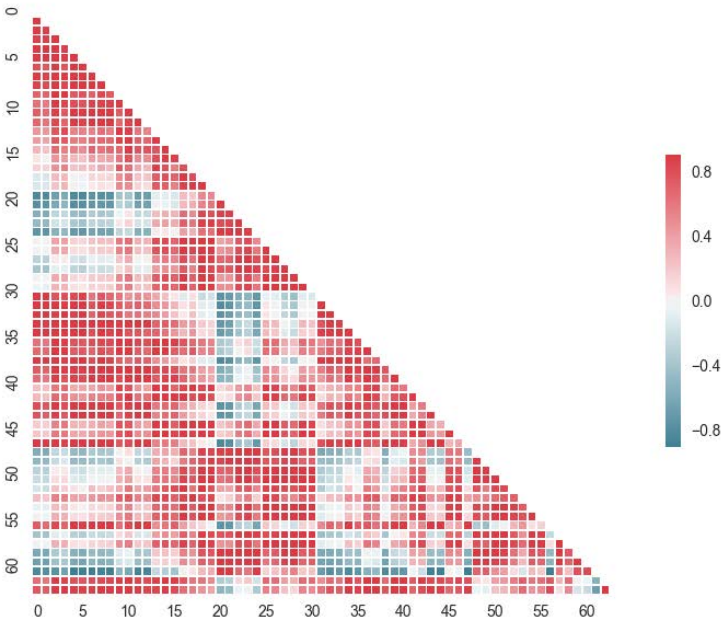
Alcoholics with S1obj stimulus:



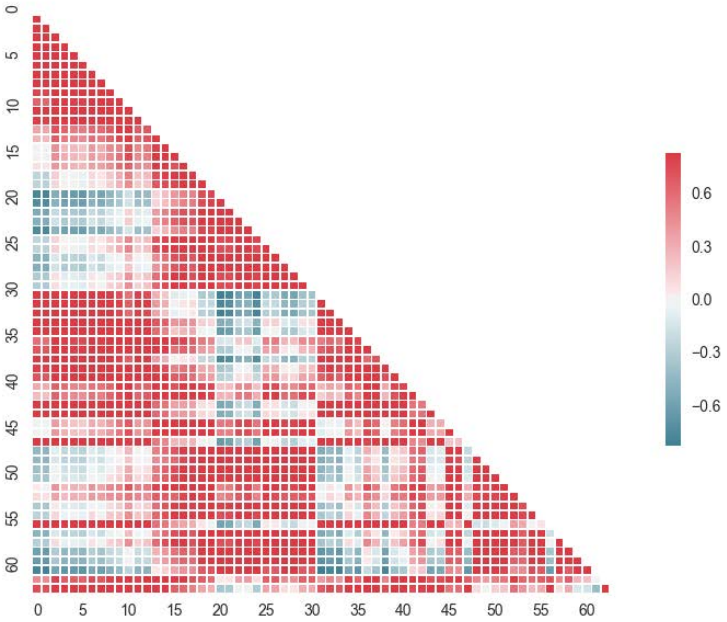
Control with S1obj stimulus:



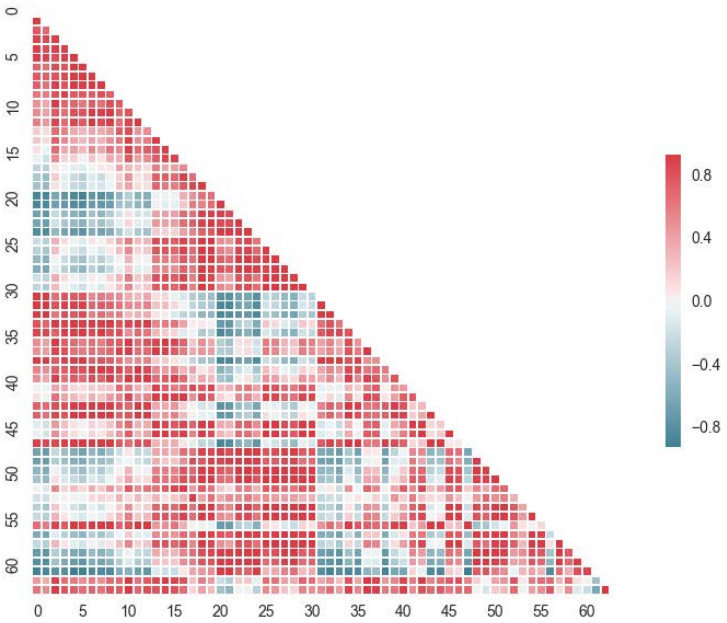
Alcoholics with S2match stimulus:



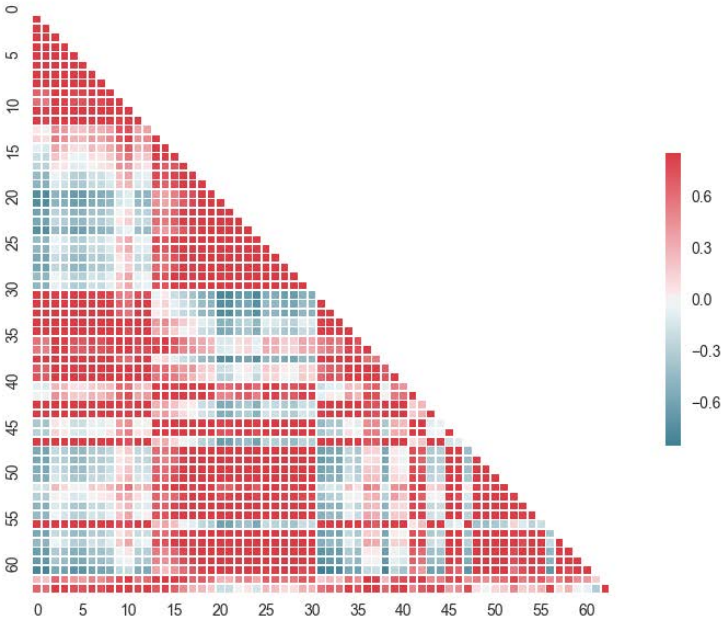
Control with S2match stimulus:



Alcoholics with S2nomatch stimulus:



Control with S2nomatch stimulus:



It can be observed that the positive and negative correlations between the channels in control subjects are higher than for alcoholic subjects. For the alcoholic subjects the correlations as determined by the colours tend to diffuse out more than for the control subjects.

References:

[1] https://docs.python.org/2/library/pickle.html

[2] https://docs.python.org/2/library/gzip.html

[3] http://pandas-docs.github.io/pandas-docs-travis/

[4] http://networkx.readthedocs.io/en/networkx-1.11/

[5] https://seaborn.pydata.org/

[6] https://docs.python.org/2/library/os.html

[7] https://archive.ics.uci.edu/ml/machine-learning-databases/eeg-mld/eeg.data.html