1 Decoding ECoG Data

Team Members: Jake Kornblau, Linda Zheng
Mentors: Genevera Allen, Fredrick Campbell

2 Questions

To determine with maximal accuracy what audio-visual stimulus the subject saw based on the (electrocorticography) ECoG readings, with an emphasis on discovering regions of the brain that are highly associated with correctly distinguishing between the different stimuli.

3 Data

The data consists of ECoG readings from a single epileptic patient over 224 trials. In each trial, the subject saw a video if a woman saying either "rain" or "rock" in which the video image was either clear or blurry and the audio was either clear or noisy. There are 28 trials (repetitions) for each of the 8 possible stimulus. For each of these trials, there are readings of the electrical activity on the subjects brain at 70 different node sites along 96 different frequencies over 301 distinct time-points. It should be noted that these 301 time points are themselves smoothed values from even more-frequent readings. It should be noted that this data set is interesting as there are relatively few trials for the number of variables. In order to help reduce the number of variables, electrical readings along the 96 different frequency were averaged together along the six different frequency bands that compose neural activity.

4 Statistical Methods

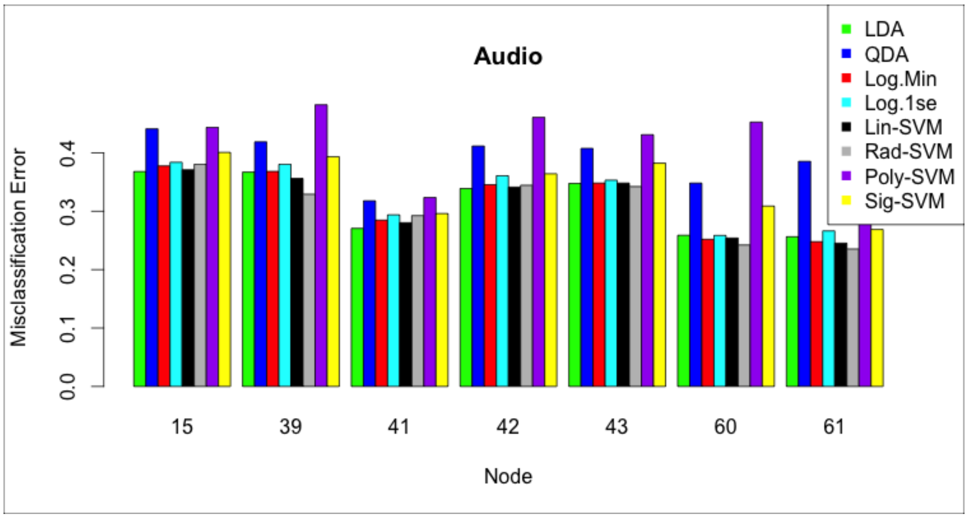
The first problem that we had to overcome was the low number of trials as this prevented us from merely splitting the data into training, validation, and test sets as is typically standard. Thus, we randomly split our data into training, validation, and test sets 100 times. For the rest of the project, we will always for all models and do all dimensionality reduction on these 100 random splits separately. All results, such as validation accuracy, will be averaged across the 100 random splits, helping ensure that the models we build are not over-fit while still using all of the data.

We initially decomposed the overall multi-class classification problem into three binary classification subproblems: (1) identifying stimuli with clear audio, (2) identifying stimuli with clear video, and (3) identifying stimuli in which "rock" was said. As stated above, a major dimensionality reduction was achieved by averaging across the frequency bands as the readings within these bands were effectively identical. The dimensionality was further reduced by projecting the data for each node onto components produced by per-node partial least squares. Specifically, a higher-order partial least squares method created by Fredrick was utilized since our data was in a data-cube. We then attempted to determine the best classifier for our problem by building a variety of classification methods using cross-validation on the training set including LDA, QDA, SVM (linear, radial basis, polynomial, and sigmoid), as well as logistic regression. These methods were then tested on a validation set in order to determine the predicted accuracy.

5 Findings

Overall, our preliminary findings found that gaussian-based decision methods were sufficient for classifying the audio and visual binary classifiers. We were able to get misclassification rates in the 10-20% range using a single node for the problems of determining the clearness of the video and clarity of the audio. However, the lowest misclassification rates that we could get for distinguishing between rock and rain was 40%, which was much higher than we had hoped.

As can be seen in the table below, LDA and Radial-Basis SVMs appeared to consistently have some of the lowest misclassification rate. Overall, the methods were effectively equal as they were all within the standard error limits with the exception of QDA and Polynomial SVMs. We consider these results to be preliminary. While these results show that there is clear signal in the data, it does not sufficiently answer our original question. The results are for the binary classifiers and not the multi-class problem that we are trying to solve. Additionally, these classifiers all only use data from one node and we believe that we can achieve significantly better results through the use of multiple nodes, especially on the currently elusive problem of distinguishing between rock and rain.



6 Next Steps

Overall, we are quite satisfied with our current findings as we have been able to get good accuracies on the binary classifiers. Our next steps are as follows:

- Use ensembling methods to aggregate the per-node classification methods potentially more accurate classifiers for the binary methods. Since only some of per-node classifiers were better than random chance, we will use sparse ensembling methods.
- While we have had much success with Fredrick's higher-order sparse partial least squares method, we plan on trying different dimensionality reduction methods, such as trying some matrix factorization methods on the spectrogram.
- Fitting classifiers to multi-node components calculated using Fredrick's higher-order sparse partial least squares method.