*Please note this paper outlines preliminary research conducted during Fall 2019. Paper finalized on December 14, 2019.*

Connor Capitolo and Catherine Kim

# Analyzing fMRI Data to Predict Cognitive Scores

## Introduction

The general goal for our research is analyzing the effects of the cortex and thalamus in understanding an individual's cognitive and behavioral functions. While there has been extensive research examining different areas of the cortex, the thalamus has typically been treated as one big lump. However, the thalamus is an incredibly important region of the brain that has a unique topological architecture, which may help to strike an optimal balance between globally integrated and locally segregated cortical states (Shine et. al., 2019). It is known as a relay station for sensory signals, and is important in the regulation of sleep, consciousness, and awareness. Specifically for this project, we wanted to understand if the thalamus, cortex, or a combination of the two best predict an individual's cognitive scores.

## Background

We used data from the Human Connectome Project, which describes itself as a "project to construct a map of the complete structural and functional neural connections in vivo within and across individuals." HCP is a compilation of 1206 total subjects to date from young healthy adult twins and non-twin siblings between the ages of 22 and 35 that collects fMRI, EEG, MEG, and behavioral data. For our project, we focused on the resting state fMRI data from 100 Unrelated Subjects. Specifically, we used HCP's package "Resting State fMRI FIX-Denoised (Compact)," which includes only the grayordinate time series for each scan. This package consisted of fMRI preprocessed data, in which spatial distortions have been minimized and data have been aligned across modalities and across subjects using appropriate volume-based and surface-based registration methods. According to HCP's Reference Manual, each individual's data was acquired over a run of approximately 16 minutes where the subject had eyes open with relaxed fixation on a projected bright crosshairs with a dark background. One of our main challenges at the start was actually taking this package, which is a total of 2,200 files and 395.32 GB, and extracting the 100 files (each around 950 MB) corresponding to each HCP subject's resting state fMRI data. With help from Professor Catie Chang and the use of ACCRE to deal with the medium-sized data set, we were successfully able to extract the files.

In order to effectively analyze the data, we used an fMRI technique called resting state functional connectivity. This approach allows for noninvasive parcellation of the brain and relies on the observation that in the absence of any task, spatially distant regions of the cortex exhibit highly correlated patterns of blood oxygenation level-dependent (BOLD) activity (Gordon et. al., 2016). In sum, resting state functional connectivity allows us to determine the correlation between the spontaneous activity of brain areas that are anatomically separate but may have functional communication between the disparate regions (Rodriguez et. al., 2019). Functional connectivity has been used to accurately predict children with ADHD and examine symptoms of Parkinson's disease (Rosenberg et. al., 2016; Engels et. al., 2018).

**Methods**

With help from Luke Chang's course "Introduction to fMRI data analysis," we were able to extract the features for use in our model. We used a Python package for analyzing neuroimaging data called NLTools. First, we used the Brain_Data() class, which stores imaging data as a vectorized matrix. Each image is an observation, and each voxel is a feature. We then had to obtain the cortex and thalamus atlases to identify the regions of interest for the two brain regions. We used BT Yeo's 17-network liberal mask cortical parcellation, which was based on 1000 young, healthy adults using a clustering approach to identify and replicate networks of functionally coupled regions across the cerebral cortex (Yeo et. al., 2011). For the thalamus, we actually used an atlas created at Vanderbilt with 13 total regions of interests, but only 10 ROIs were used due to a small number of voxels in the other three. Using our fMRI data, the average time course within each region of interest was extracted for both the cortex and the thalamus. Now that the nodes were specified, the edges of the graph were calculated using pearson correlations with the pairwise_distance function from scikit-learn. The distance metric was then converted into similarities by subtracting all of the values from 1. Since correlation matrices are symmetric, we took the lower triangular matrix as the features for our model. Therefore, there was a total of 136 features for the cortical matrix (16*(16 + 1) / 2), and 45 total features for the thalamus matrix (9*(9+ 1) / 2).

Since our project aims to examine the prediction accuracy of thalamus, cortex, and the combination of the two on individuals' cognitive scores, we also combined the elements we had taken from the cortical correlation matrix and thalamus correlation matrix. Therefore, the cortical-thalamus matrix had a total of 181 features. Another form of the data we examined was using the Fisher Z-Transformation, which normalizes the distribution of the input features. Therefore, the cortical, thalamus, and cortical-thalamus data all had the same number of features as the original form. Finally, since the Fisher Z-Transform standardizes the data, we applied Principal Component Analysis (PCA) from scikit-learn on the Fisher Z-Transformed matrices for dimensionality reduction so that 95% of variance was explained. The resulting cortical data had 25 features, the thalamus data had 2 features, and the cortical-thalamus data had 25 features. We then had thalamus, cortical, and cortical-thalamus features in three forms: original, Fisher Z-transformed, and PCA transformed. All matrices had 100 total observations.

We decided to analyze the cognitive score Pattern Completion Processing Speed because it provided an age-adjusted scaled score with values ranging from 67.35 to 144.16, meaning that it's a regression problem. The test measures processing speed by asking participants to discern if two side-by-side pictures are the same. A participants' raw score is the number of items correct in a 90-second period. Fortunately, there was a .csv file provided on HCP's website that had all the cognitive and behavioral scores for the 100 Unrelated Subjects. All we needed to do was load the .csv file into a Jupyter notebook and use python to extract the Processing Speed scores.

After processing the 100 participants' brain imaging data (.nii.gz), due to time constraints we decided to focus solely on the regression models with fewer tuning hyperparameters: Lasso, Ridge, Elastic Net, and K-nearest neighbors (KNN). Considering the small sample size, we decided to apply Nested Cross-Validation to train the models to avoid both overfitting and data leakage. We used 4-fold cross validation for both inner and outer loops. For hyperparameter-tuning in the inner loop, we used RandomizedSearchCV from scikit-learn. Random search uses random combinations of hyperparameters to find the best solution for model building and has proven to yield better results than Grid Search (Bergstra et. al., 2012). Each time the

RandomizedSearchCV function was called, we had it sample 100 hyperparameter combinations. The detailed tuning parameters we used for random search are listed in Figure 1.

As explained above, we prepared the correlation matrix features for cortex, thalamus, and the cortical-thalamus in three forms: original, Fisher Z-Transformed, and PCA-Transformed. We trained those 9 feature groups individually on each of the four regression models. For each of the 36 different matrix-model combinations, we ran Nested Cross Validation for 30 trials and evaluated using "Negative Mean Squared Error (MSE)" and "Explained Variance" metrics from scikit-learn separately.* From the resulting scores of the 30 trials for each matrix-model combination, we recorded the median value as the score for the corresponding input matrix and model type. This is because Random Search often yields to high variance during computation, and the median is less affected by outliers than the mean (Bergstra et. al., 2012).

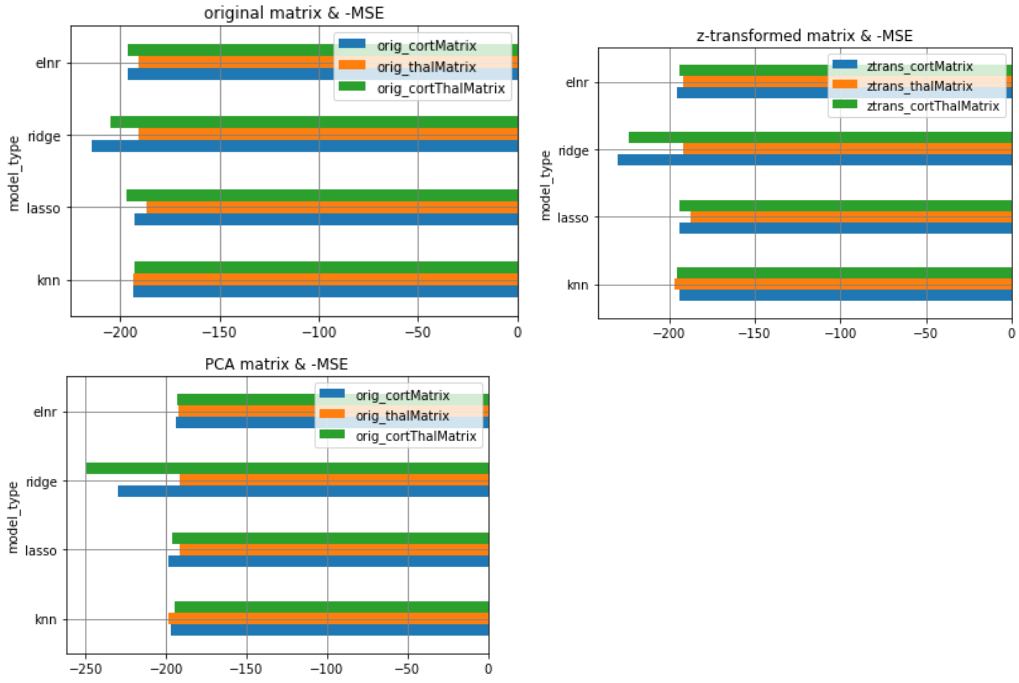
*Note: "Negative MSE" is simply the negative version of the actual MSE. The unified scoring API always maximizing the score in scikit-learn library. Thus, the most optimal negative MSE result would be the one with the lowest absolute value.

| **4 models** | Lasso | Ridge | Elastic Net | KNN |
|---|---|---|---|---|
| **Tuning Parameters** | {'alpha': sp_rand()}* | | | {'n_neighbors':range(1,57),'weights':['uniform','distance']} |

(Figure 1: For the alpha hyperparameter for Lasso, Ridge, and Elastic Net, we used stats.uniform function from Python library Scipy to prepare a uniform distribution between 0 and 1 for the Randomized Search function to sample from. For KNN, we let the Randomized Search function to select from a list of integers ranging between 1 and 56 and either uniform or distance weight.)

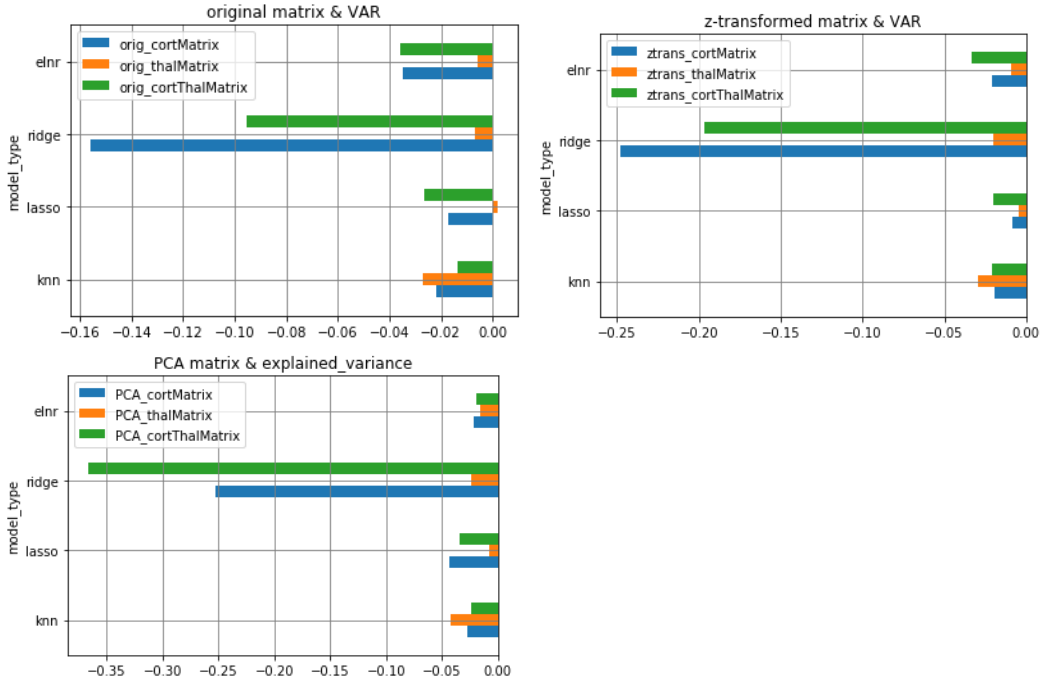
**Results**

Figure 2 shows the negative MSE results for all matrix and model type combinations. Based on our results, the original thalamus matrix with lasso regression most accurately predicted the processing speed for HCP's 100 Unrelated Subjects with an MSE of 186.49. However, the other matrix-model MSE scores don't differ too significantly.




(Figure 2)

Figure 3 shows the Explained Variance results. Surprisingly, all results turned out to be negative, except for the original thalamus matrix using lasso regression. *Scikit-learn's* built-in Explained Variance metrics is calculated using 1- Var (ypred - y)Var (y). A negative explained variance result implies the model's resulting Var(y_pred - y) > Var(y), which means the model is very ill-fitted for the data; the variance of the difference from the actual outputs and predicted outputs should not be greater than the variance of the actual outputs in the case of a working model. Although the result for original-Thalamus-Lasso model is positive at 0.0018, it is still not high enough to indicate any correlation between the predicted outputs and actual outputs.



## Discussion and Conclusions

As discussed in the Results section, none of the matrix and model type combinations we have tested lead to promising results. The results we obtained could be a consequence of any of the following reasons: 1) Cortical and thalamus connectivity at resting state have no correlation with individual's processing speed, 2) None of the regression models we have applied are a good fit for the data. The models may have been too simple to characterize the potential relationship between our input and output variables. 3) The data we used has already been pre-processed. There is a chance that the data has been too over-processed for accurate predictions. Further investigation and research are required for us to determine whether there exists any correlation between individual's resting brain connectivity and cognitive processing speed.

Moving forward, we are hoping to 1) look at different cognitive and behavioral scores. We will start from Fluid Intelligence, since some recent work has successfully identified the correlation between resting-state brain data and fluid intelligence using the same HCP dataset (Dubois et. al. 2018). We will read the paper closely to learn their methods. We might apply their models on our cortical and thalamus correlation matrix input features. Also, we want to make sure that the other scores we choose are uncorrelated, so we'll set a threshold for the correlation coefficients (possibly between 0.3 and -0.3). This approach then also entails performing classification and regression analyses and applying feature selection, such as PCA, on these other scores to see whether prediction accuracy would increase. 2) We should have access to more storage space soon, which means that we would be able to process more subjects' data, hopefully all 1206 subjects. Since our predictions from our current machine learning models are not accurate, we're looking to use more complex models with more hyperparameters, such as support vector machines, random forest, and neural networks. We also want to try out larger alpha hyperparameters and see the effects. 3) We hope to examine the data using different atlases, specifically with the cortex. In addition to the 17-network liberal mask we have used for this project, B.T. Yeo also provides three other cortical parcellation masks, including one containing 7 regions of interest. Also, there is some cortical connectivity that is not captured by Yeo's 17-network parcellation that might be correlated with some individual behavioral scores; we plan to find cortical parcellation masks that contain higher number of networks as well. 4) Finally, we are hoping to take HCP's raw data instead of their preprocessed data in order to preprocess it ourselves. There is a possibility that HCP's preprocessed data is leading to overfitting, so we want to see what results may arise from our own preprocessing.

## Resources

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

## References

Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13(1), pp.281-305, 2012.

Dubois J, Galdi P, Han Y, Paul L, Adolphs R, Resting-State Functional Brain Connectivity Best Predicts the Personality Dimension of Openness to Experience. Personality Neuroscience, 05, 2019. DOI: https://doi.org/10.1017/pen.2018.8

Engels G, Vlaar A, McCoy B, Scherder E, and Douw L (2018) Dynamic Functional Connectivity and Symptoms of Parkinson's Disease: A Resting-State fMRI Study. Front. Aging Neurosci. 10:388. doi: 10.3389/fnagi.2018.00388

Gordon E.M., Laumann T.O., Adeyemo B., Huckins J.F., Kelley W.M., Petersen S.E. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb. Cortex. 2016;26:288–303

Rodríguez-Vidal, L., Alcauter, S., & Barrios, F. A. (2019). The functional connectivity of the human claustrum according to the Human Connectome Project data. *BioRxiv*. doi: 10.1101/705350

Rosenberg, M., Finn, E., Scheinost, D. et al. A neuromarker of sustained attention from whole-brain functional connectivity. Nat Neurosci 19, 165–171 (2016) doi:10.1038/nn.4179

Shine et. al., The Low-Dimensional Neural Architecture of Cognitive Complexity Is Related to Activity in Medial Thalamic Nuclei, Neuron (2019),https://doi.org/10.1016/j.neuron.2019.09.002

Yeo BT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zollei L., Polimeni JR, Fischl B, Liu H, Buckner RL. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 106(3):1125-65, 2011.