Structural Learning of rs-fMRI data in Heavy Smokers

Yiru Gong
yg2832
yg2832@cumc.columbia.edu

Abstract

Recent studies showed alterations of brain structure and function in heavy smokers. However, the specific casual relationship of alteration among different brain functional regions and topological connection changes are yet unclear. In this project, we estimate Gaussian Undirected Graphs based on the graphical lasso algorithm for smokers and non-smokers rs-fMRI data and identify the stable graph over bootstrapping. The graph similarity and node similarity are then compared to identify the brain regions with most significant alterations in brain connections. Our results showed that the estimated graph showed high stability throughout random sampling and several identified brain regions showed relationship with cigarette smoking in clinical studies. Therefore, our model provide informative hints to the topological changes of brain region connections in heavy smokers, which benefit the future clinical research.

1 Introduction

Smoking, especially heavy and long-term smoking, are shown harmful to cardiac, pulmonary, and vascular systems. Meanwhile, recent studies on neuroimaging also suggested an adverse effect on the brain functions [6] and degradation in neural connections [1] for long-time smokers. Multiple brain regions throughout the whole brain has been clinically identified with modifications in heavy smokers, and showed a positive correlation between cigarette addiction in epidemiologic studies [4].

However, the specific topological connection changes and the specific pathological pathways among different brain functional regions are yet unclear. Therefore, in this project, we aimed on applying graphical models, in specific, Graphical Lasso (glasso) algorithm to the brain fMRI data of smokers and non-smokers to identify the Markov Random Field model (MRF) of brain connectivity of each respectively, and compare the node-wide difference between the graphs of smokers and non-smokers.

2 Data

The resting-state functional magnetic resonance imaging (rs-fMRI) scan was performed on 37 heavy cigarette smokers and 36 healthy non-smoking control subjects. The detailed data acquisition and preprocessing procedures were described by [3]. In brief, the normalized fMRI signals are parcellate into 116 anatomical regions of interest according to the automatic anatomical labeling (AAL) atlas [7], and 200 volumns (each last for 2 seconds) of fMRI records are corrected and kept for analysis. The preprocessed data are provided by Lin and was ready for graphical network analysis.

To simplify the time series analysis of fMRI data, we sparsely select the first volume in every three volumns (i.e. every 6 seconds) to eliminate auto-correlation. This results in 66 samples per subject and 2442 samples for smokers, 2376 samples for non-smokers.

Preprint. Under review.

3 Method

3.1 Gaussian Undirected Model

The gaussian undirected graph is a graphical model indicating the adjacency relationship between every pair of variable nodes based on the independence status when condition on all other variables. The edges in the graph thus indicate a direct link between two variables, while eliminating the effect of indirect or multi-level associations. It is thus helpful for our fMRI data to identify direct linkages between two brain regions and imply the connectivity of the functional regions.

Here we assumed the fMRI data follows a multi-variable Gaussian distribution $X \sim N(0, \Sigma)$ with the precision matrix $K = \Sigma^{-1}$. The conditional independence is thus achieved if and only if $K_{ij} = 0$. Then we can learn undirected graphical structure by estimating "structural" zeros in the precision matrix, or estimating the entire precision matrix and treating small entries as zero.

3.2 Graphical Lasso Estimation

To estimate the precision matrix of the undirected graph, we applied the Graphical Lasso algorithm which applies sparsity penalty to the precision matrix K:

$$\hat{K}^{gl} = \arg min_K \{-\log\det(K) + tr(SK) + \lambda||K||_1\}$$

Where S is the sample covariance matrix, λ is a tuning parameter to control the sparsity penalty. The $||.||_1$ is the l_1-form.

As a result, the graph adjacency matrix could be estimated based on the estimated precision matrix. In particular, we used the embedded glasso function in R package 'huge'.

3.3 Parameter tuning and RIC criteria

To pick a proper regularization parameter value of λ in the glasso algorithm, we used the rotation information criterion (ric) for every lambda value and pick the one with the best ric score. In detail, RIC randomly rotates the variables for each sample multiple times and selects the minimum regularization which generates all zero estimated using rotated data. One drawback of the method is the potential of underselection. However, since we focus more on the consistency of neural connection results, it is still acceptable to have a relatively higher false negative rate.

3.4 Correlation matrix comparison

We could then obtain the denoised correlation matrix based on the estimated precision matrix. The estimated correlation matrix would then be compared with the sample Pearson correlation matrix to check if the algorithm identifies the major relationship between variables.

3.5 Similarity comparison

After generating the independent undirected graphs for smokers and non-smokers separately, we would like to compare the similarity of connections for each node (region of interest). In specific, we performed bootstrap to the dataset to generate ten new dataset with 2500 observations each. A graph with adjacency matrix was estimated for each dataset and the frequency of an appeared edge was recorded for the ten graphs. If an edge showed more than 9 times in the ten graphs, we identified it as a stable edge and kept in the final graph. The same process is performed to smoker and non-smoker data separately to get a final graph for each dataset. The overall graph similarity is calculated by Sorensen-Dice coefficient, defined as:

$$SD(A, B) = \frac{\sum(adjA \neq 0 \cap adjB \neq 0)}{\sum(adjA \neq 0 \cup adjB \neq 0)}$$

While the node-specific similarity score is calculated by the Jaccard Similarity Score, defined by:

$$J(a,b) = \frac{|a \cap b|}{|a \cup b|}$$

Nodes with significant differences are filtered out if they have a smaller similarity score than the overall graph similarity score.

3.6 GLMNET

To compare the results of our graphical models with other non-structural-based models, we apply the similar penalized multivariate logistic model (GLMNET) with elastic net penalty on the logistic regression model on smoker status. The λ penalty parameter is tuned on ten-fold cross validation with AUC (area under the ROC curve) as evaluating criteria. The correlation between the brain regions and the response smoking status is then extracted to find any significantly correlated brain regions.

4 Results

4.1 Validation of Gaussian Undirected graphs

We performed the glasso estimation and lambda parameter tuning of the undirected graph for both smokers and non-smokers. A larger lambda would always correspond to a more sparse graph Figure 1. Meanwhile, the optimized graphs of smokers and non-smokers showed similar results, as indicated in Table 1. The similarity score of the optimized graphs for smokers and non-smokers are then calculated based on the Sorensen-Dice coefficient.

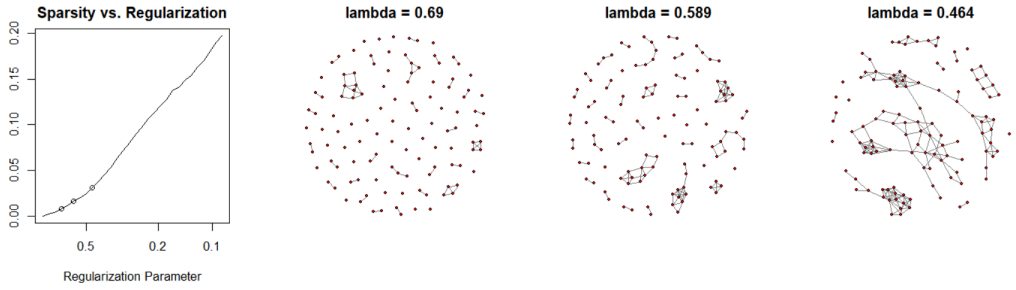

Figure 1: Estimated graph of different lambda values in non-smokers

Table 1: Comparison between the optimized graphs in smokers and non-smokers

	lambda	Number of nodes	Number of nodes (stable)*	Similarity Score*	Similarity Score (stable)
Smokers	0.2814	555	454	0.6025	0.631
Non-smokers	0.2736	586	430		

*The graph-wise similarity score is calculated through Sorensen-Dice coefficient.
*The stable here means the final stable graphs with common edges in >=9 bootstrap graphs throughout 10 in total

Meanwhile, we compare the reconstructed correlation matrix based on the graphic model. As indicated in Figure 2, the graphical model is able to recap most of the variance and significant correlations from the sample.

4.2 Similarity Score Comparison based on stable edges

After filtering out the common edges exists in more than 90 % of the repeat graphs, we obtained an overall graph composed of common edges. The overall graph similarity score is 0.631 Table 1, which is a little bit higher than the single graph model. Meanwhile, the number of stable edges remains

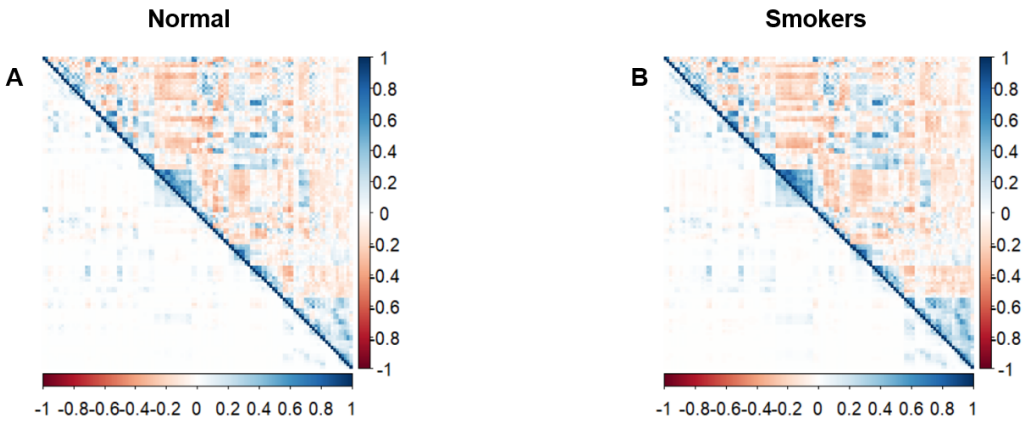

Figure 2: Connectivity between brain regions - Pearson correlation matrices (top triangle) and undirected graphs (bottom triangle; direct connections as discovered by glasso, correlation calculated from estimated precision matrix). Left: non-smokers; right: smokers.

454/555 in smokers, and 430/586 in non-smokers, indicating the graphical estimation method is relatively stable.

The node-wise Jaccard similarity score is also calculated in the stable graph and filtered with a threshold of 0.631 (Sorensen-Dice coefficient). As a result, 50 brain regions out of 116 regions are identified significant Table 2, with a smallest Jaccard Score of 0.29 in the "Temporal_Inf_L". We then take a further look at the specific connection change in the top three changed brain regions, namely “Temporal_Inf_L”, “Thalamus_R”, “Cerebelum_Crus2”. For example, in Temporal_Inf_L, 6 connections are identified in healthy non-smokers, and 3 connections are identified in heavy smokers, while only 2 edges are shared (namely "Temporal_Mid_L", "Temporal_Inf_R"). This corresponds to the neural region functions that they all belongs to the Temporal region. The results also indicated lose in connection with "Frontal_Mid_Orb_L", "Frontal_Inf_Orb_L", "Parietal_Inf_L", "Angular_L" in heavy smokers, which worth further clinical verification Figure 3.

Table 2: Top 20 regions most altered between smokers and non-smokers

name	index	jaccard_score
Temporal_Inf_L	8301	0.285714
Thalamus_R	7102	0.333333
Cerebelum_Crus2_R	9012	0.333333
Cerebelum_10_L	9081	0.333333
Angular_R	6222	0.35
Angular_L	6221	0.352941
Precuneus_L	6301	0.357143
Caudate_R	7002	0.375
Precentral_R	2002	0.4
Frontal_Mid_L	2201	0.4
Supp_Motor_Area_R	2402	0.4
Cingulum_Mid_L	4011	0.4
Frontal_Sup_L	2101	0.4375
Occipital_Mid_L	5201	0.4375
Calcarine_L	5001	0.444444
Parietal_Sup_R	6102	0.444444
Precuneus_R	6302	0.444444
Frontal_Mid_R	2202	0.466667
Temporal_Sup_R	8112	0.466667
Frontal_Inf_Orb_R	2322	0.5

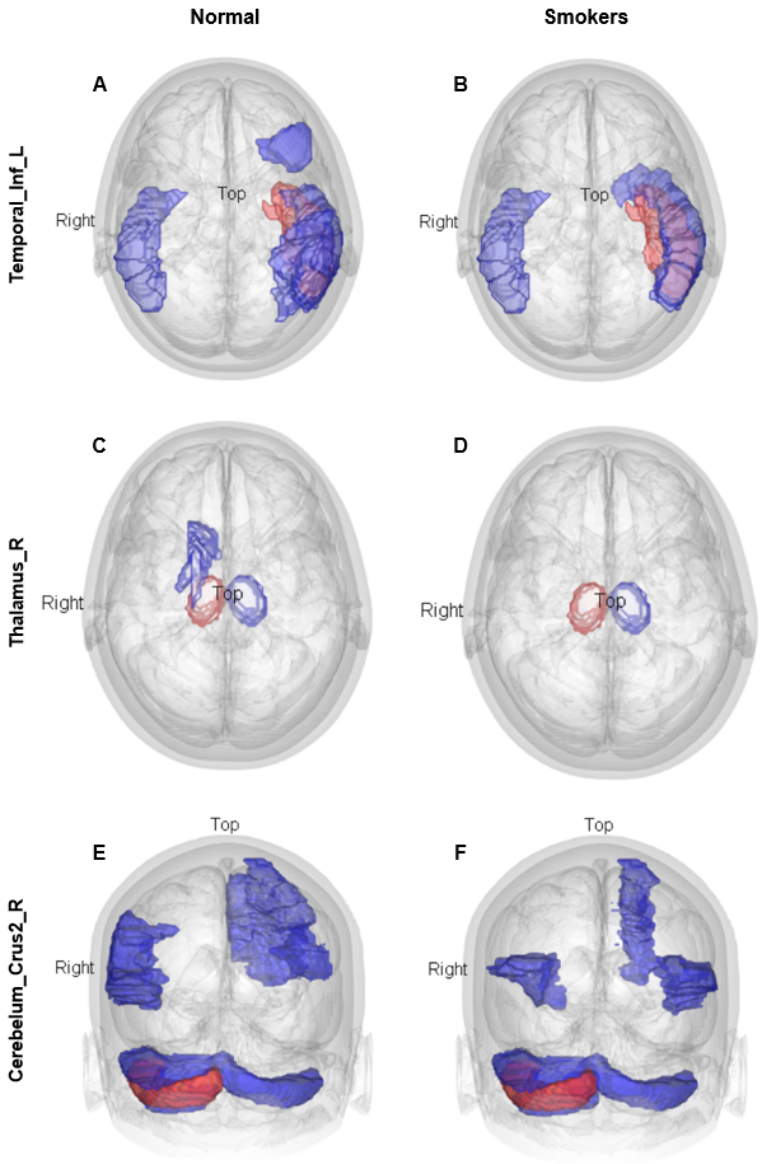

Figure 3: Brain connection visualization in Top 3 largely changed brain regions - Red: node region, blue: directed connected regions; left: non-smokers, right: smokers.

4.3 Comparison between Graphical Model and Non-graphical Model

Other study of fMRI uses regressions to predict the correlation between brain regions and smoking status. Usually the demographic data will also be included in the model. To compare the methods, we also conducted a glmnet classification model as discussed in method and tuned for the best hyperparameters. However, the model does not perform well on the complex fMRI data, resulting in a maximum ROC of 0.5 (with tuned parameter: alpha=0, lambda=148.41), which is totally the same as random effect. The extracted coefficient of the best model also showed no significant results of variable coefficients since all of them are smaller than 10^{-40}. Therefore, we further proved that the regular non-graphical model has difficulty in identifying the difference of brain regions in smokers and non-smokers. Thus a graphical model show more importance in complex fMRI data structure learning.

5 Discussion

5.1 Alteration in Brain Region Connections

According to our graphical model comparison, we identified missing connections in several brain regions in heavy smokers. The Temporal_Inf_L, known as left Inferior temporal gyrus, has been shown related to cigarette smoking in the dALFF experiment verification [8]. Meanwhile, the thalamic glutamate level is shown reduced in cigarette smoking [5]. Similar results are found in the top altered nodes, which indicates a good reflection of our model on the effect of smoking on brain connectivity. Besides, the Cerebelum_Crus2_R, known as the right posterior crus II cerebellum, is responsible for social mentalizing and emotional self-experiences [2]. However, few study exists on this specific brain region and the relationship with smoking is yet unclear. It is thus worth further clinical research to identify potential casual relationship between the loss of connection of cerebellum crus II and smoking.

5.2 Model Limitations and future directions

Although our undirected graph showed great performance on the fMRI data analysis, limitations still exist. Firstly, our model ignore the feature of autocorrelation in time-series data and analyze each time stamp independently, which might results in collinearity problem and missing of some time-dependent features. Also, the previous research applied a 25-second interval to reduce the collinearity, however, our dataset has limited sample size and fMRI test length and if a 25s interval are applied, only 600 samples could be generated for each group, which is relatively limited compared to 116 variables. If possible, a time-series based graphical models could be developed or applied to fMRI data in the future study, or a longer fMRI test time might be applied to get a larger dataset with more efficient data points.

Secondly, the current model did not identify a specific hypothesis testing method to determine the threshold of "nodes with significant alterations". We might assume the jaccard score follow specific normal or t distribution and perform statistical test to compare the node similarity score with the overall graph similarity score. An adjusted multi-test correction should also be applied in this process to ensure the proper interpretation of individual test. Regardless of the limitation on threshold, it is still informative since we sort the nodes with Jaccard score which will provide great hint for future clinical and experimental studies.

References

[1] Nixon SJ Durazzo TC, Meyerhoff DJ. A comprehensive assessment of neurocognition in middle-aged chronic cigarette smokers. *Drug Alcohol Depend*, 122:105–111, 2012.

[2] Elien Heleven Frank Van Overwalle, Qianying Ma. The posterior crus ii cerebellum is specialized for social mentalizing and emotional self-experiences: a meta-analysis. *Social Cognitive and Affective Neuroscience*, 15:905–928, 2020.

[3] Wu G. Zhu L. Lin, F. and H. Lei. Brain networks in smokers. *Addiction Biology*, 20:809–819, 2015.

[4] Zhu L Lei H Lin FC, Wu GY. Heavy smokers show abnormal microstructural integrity in the anterior corpus callosum: a diffusion tensor imaging study with tract-based spatial statistics. *Drug Alcohol Depend*, 129:82–87, 2013.

[5] Hudkins M Oh EY Hellemann GS Nurmi EL London ED O'Neill J, Tobias MC. Thalamic glutamate decreases with cigarette smoking. *Psychopharmacology*, 231(13):2717–24, 2014.

[6] Lessov-Schlaggar CN Swan GE. The effects of tobacco smoke and nicotine on cognition and the brain. *Neuropsychol Rev*, page 259–273, 2007.

[7] Papathanassiou D. Crivello F. Etard O. Delcroix N. et al. Tzourio-Mazoyer N., Landeau B. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15:273–289, 2002.

[8] Yang Z Wei Y et al. Wen, M. More than just statics: Temporal dynamic changes of intrinsic brain activity in cigarette smoking. *Addiction Biology 26(6):e13050*, 2021.