Association Between Mice Connectome and Behavioral Indices

Youngsoo Baek, Yunran Chen, and Brian Kundinger

Apr. 16th, 2020

Abstract

We study the possible associations between brain connectomes and behavioral trait measurements across 38 different mice. Cross correlation analysis results suggest a possible covariance between poorer learning and more connections within the isocortex, and better learning and more connections within the hindbrain. We find no indication of significant associations based on different Beta regression models with connectomes summary statistics as covariates.

1 Introduction

Connectomes are collections of white matter fiber tracts connecting different regions of the brain, representing how the neural connections in the brain react to daily challenges. A recent study collected connectomes of 55 distinct mice, and their behavioral traits related to learning and memory (NormSWTime, NormSWDistance), and novel object recognition (RI_T2, RI_T3). Our goal is to analyze the possible relationship between the connectomes and behavioral traits. Connectomes data are available in the form of 332×332 weighted adjacency matrices of undirected graphs for each of the 55 mice. Each entry of these graphs represent counts of fibers connecting different regions of interest (ROIs) that belong to left or right hemisphere, and one of the 8 known superstructures. A major challenge of this study is the need for novel dimension reduction methods or summary statistics to analyze the connectomes arrays, which are both high-dimensional and sparse, given very few samples.

2 Materials & Methods

Given the time constraints and the broad nature of the problem, we focus on exploratory analyses using canonical correlation analysis (CCA) and simple Beta regression models with behavioral traits as response variables, normalized to a scale between 0 and 1.

2.1 Canonical Correlation Analysis

Canonical correlation analysis is an exploratory method for estimating correlation between two sets of variables by finding linear combinations of variables that maximally correlate. Let $X = (X_1, ..., X_p)'$ and $Y = (Y_1, ..., Y_q)'$ denote two random vectors. CCA aims to find linear coefficients $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^q$ such that $\text{Cor}(U, V)$ is maximized, under unit variance constraints for $U = a'X$ and $V = b'Y$. Similarly as in principal components analysis, a total of $\min(p, q)$ pairs of linear coefficients are determined via singular value decomposition, subject to pairwise orthogonality constraints. The resulting set of uncorrelated pairs (U_i, V_i) are called canonical variates.

2.2 Beta Regression

Beta regression is a generalized linear model for response variables that are normalized to a scale between 0 and 1. Suppose a random variable Y has a Beta(a,b) distribution. Parameterizing the Beta likelihood in terms of the mean of Y, $\mu = \frac{a}{a+b}$, yields the likelihood:

$$f(y|\mu,\phi) = \frac{\Gamma(\mu)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where $\phi = a + b$ is often referred to as the precision parameter, since large ϕ implies a likelihood narrowly centered around a neighborhood of mean μ. The linear model is fully specified by a linear predictor $x_i'\beta$ for each response y_i and a link function $g : (0, 1) \to \mathbb{R}$ (frequently chosen to be logit). An analogous expression that models the precision allows a regression model that can account for heteroskedasticity in the data. We use the method to build an exploratory model for associations within behavioral variables, and between behavioral variables and summaries of connectomes.

3 Results

3.1 Data Preprocessing

Data processing involved matching different observations across connectomes and behavioral traits data sheets based on "runno" identifiers and animal IDs. A mouse identified as "N54891/N54900LR specific," and another with "NA" runno identifier, were removed from the analysis. The final analysis included 38 distinctly identifiable mice.

3.2 Data Visualization

Figure 1 shows a heatmap plot of counts of zeroes for each of the connectomes adjacency matrix entries across 38 mice. The plot illustrates the extreme sparsity of the connectomes counts and the small between-sample variability. Figure 2 shows averaged counts among non-zero entries (regions not greyed out) on log scale. The plot illustrates how some clusters readily visible from the data do not necessarily align with the known superstructure boundaries. These plots evince the challenges addressed in Section 1, and in particular suggest significant interest for community detection methods to summarise high-dimensional connectomes in a data-driven manner. Our current analysis do not include such results due to implementation difficulties, of which some explanation is given in Section 4.

3.3 Results: CCA

For CCA, we first extracted only the connectomes entries for counts within each superstructure, and vectorize the lower triangular entries, which yields a 38×2107 matrix for our samples. Principal Component Analysis was applied to further reduce the dimension of features to 35. The 38×9 matrix of behavioral traits included demographic backgrounds, learning and memory measurements averaged over 5 days, and novel object recognition indices. Based on the first canonical variates, we can estimate how a pair of a principal component and behavioral trait co-varies, and identify the connections between ROIs that contribute the most to this covariance based on PCA loadings.

Figure 3 shows cross-correlation estimates between principle components and behavior traits, several of which values are high. Figure 4 (left) shows the correlation between traits (blue) / principle components (red) and canonical covariates. A noticeable pattern of covariance exists between the 13th principal component and average pool time, which both locate outside the circle

of 0.5 radius. Figure 4 (right) orders in a barplot the subdivisions of brain, which correspond to the top 100 connections contributing the most to this principal component. The connections within the isocortex and the hindbrain contribute the most, in a positive and negative direction, respectively. In the context of the experiment, greater and lesser pool time each implies poorer and better learning ability.

Figures 5 and 6 are heatmap plots of counts within the isocortex and the hindbrain, sorted by average pool time in increasing order. For brevity, only the plots for female, old, C57 genotype mice are shown here. We observe more connections in the isocortex, and less in the hindbrain, as the average pool time increases, suggesting evidence from the data supporting our inference.

3.4 Beta Regression

We fit three different Beta regression models. In the first model, NormSWTime is regressed on recognition index variable RI_T3 and time. In the second model, various summary statistics for connectomes networks have been included as further covariates in the first model. In the third model, recognition index is regressed on a vectorized, raw connectomes matrix. Since this leads to obvious estimability issues, we first aggregate all connectomes counts within the same superstructures, and then vectorize only the diagonal and lower triangular entries in this matrix, as it is symmetric.

While we were unable to determine any evidence of association between the connectomes and behavioral traits, our analysis highlighted several aspects of the data that warrant further study. For instance, Figure 7 shows that the Beta regression model identified positive correlation between recognition index (RI_T3) and learning (NormSWtime). However, we also infer from the same model that the variance of the observations decreases over the course of the experiment, so that NormSWTime for all mice become more similar over time. We therefore conjecture that if a significant association between the connectomes and the learning ability exists, then it will be strongest in the first trial and decrease over time, as the inherent between-sample difference in the brain connectivity is overcome through repeated trials.

The second model, in which we added as covariates standard summary statistics for connectomes networks, such as centrality and betweenness, yielded no meaningful results from which we can infer the effect of connectomes on behavioral traits. Similarly, no meaningful inference was possible from fitting the third model. Modeling the association between connectomes and response variables, therefore, is an ongoing work for us.

4 Discussion

The major difficulties in fitting a regression model stemmed from summarising high-dimensional, sparse networks in a meaningful way. We aggregated the counts over known superstructures, but our EDA results illustrate these clusters are most likely suboptimal. Community detection within graphs, which can account for between-graph variation, is therefore of significant interest for the problem. In particular, a Dirichlet-like regression framework based on a mixture membership model can provide flexible ways to model the association between response and clusters within connectomes. We attempted to implement a relatively simple EM algorithm to estimate the membership probability parameters as introduced in Newman and Leicht (2007), but unfortunately faced numerical issues in implementing it. Given more time, we also hope to further extend the tensor regression framework, introduced in Guhaniyogi, Qamar, and Dunson (2017), such that we can systematically induce strong shrinkage given few samples and explicitly address symmetry constraints.

Reference

(1) "Bayesian Tensor Regression," Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). *Journal of Machine Learning Research*, 18.

(2) "Mixture models and exploratory analysis in networks," Newman, M. E. J. and Leight, E. A. (2007). *PNAS*, 104.

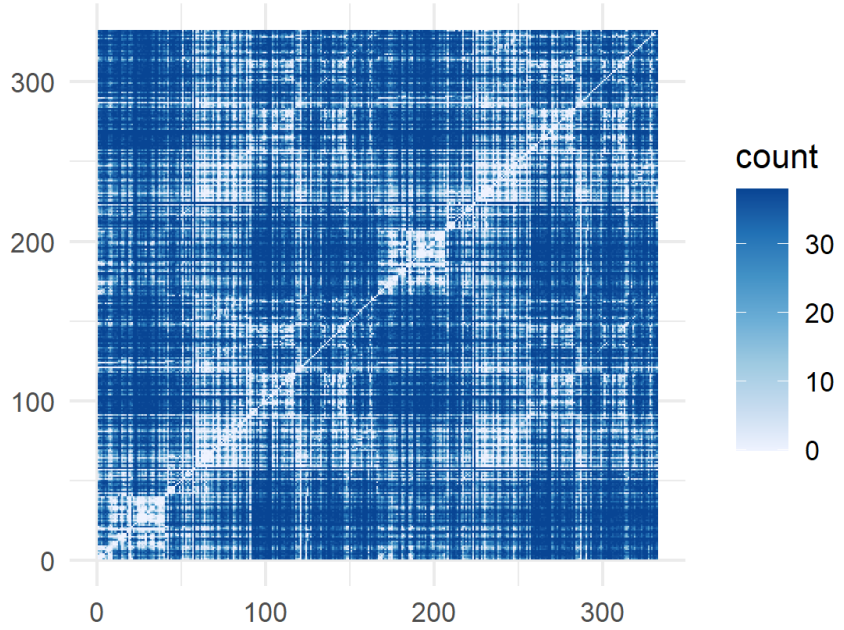

Figure 1: Count of zero entries for 38 mice connectomes. Thicker blue indicates more zeroes.

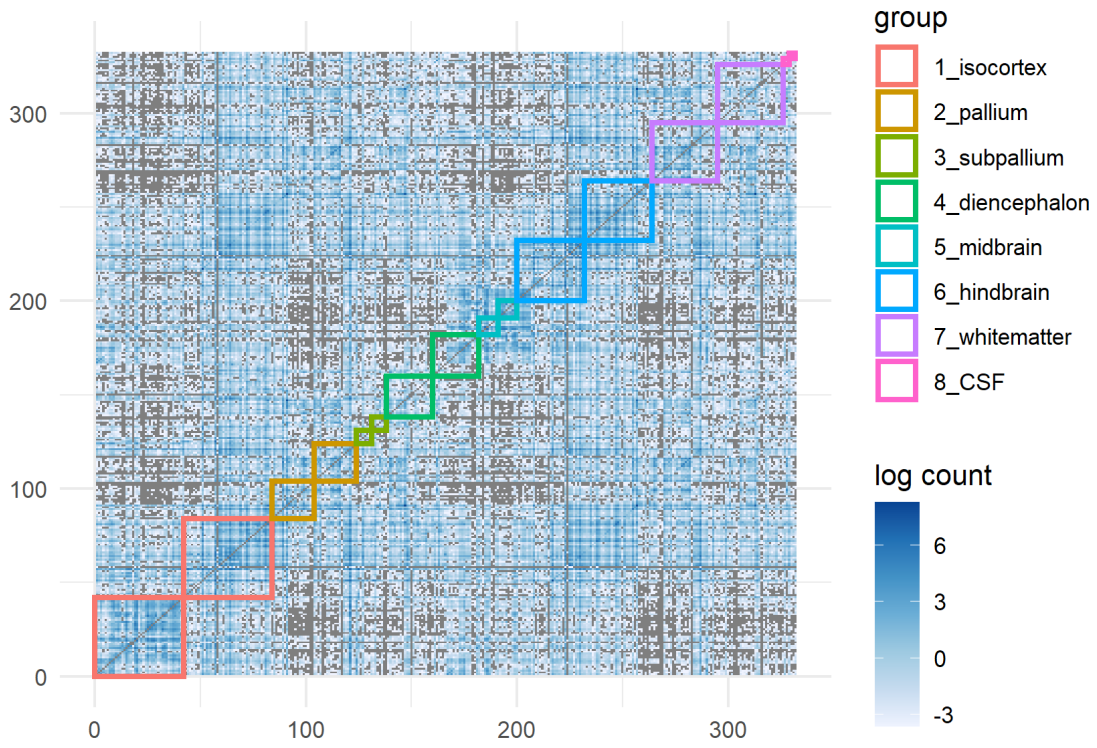

Figure 2: Average log count for non-zero connectomes entries. Diagonal rectangles indicate different superstructures, divided into left and right hemispheres.

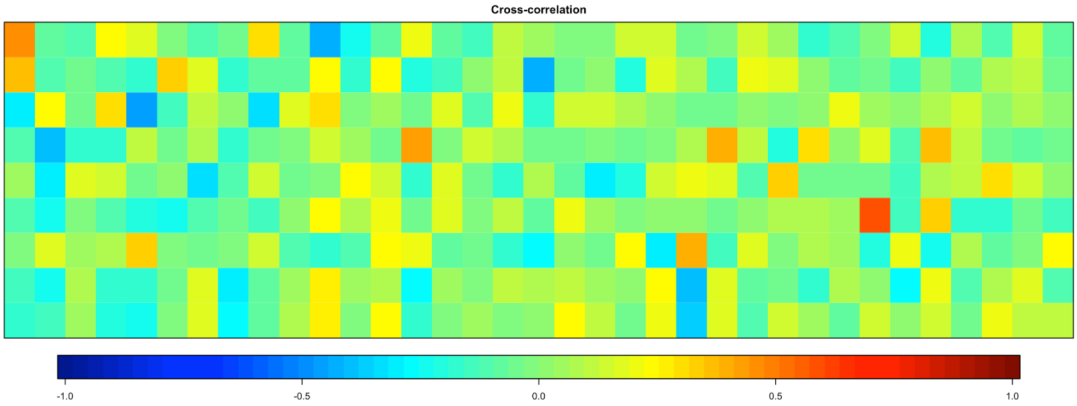

Figure 3: Cross Correlation between 35 Principle Components extracted from Connectomes and Behavioral traits

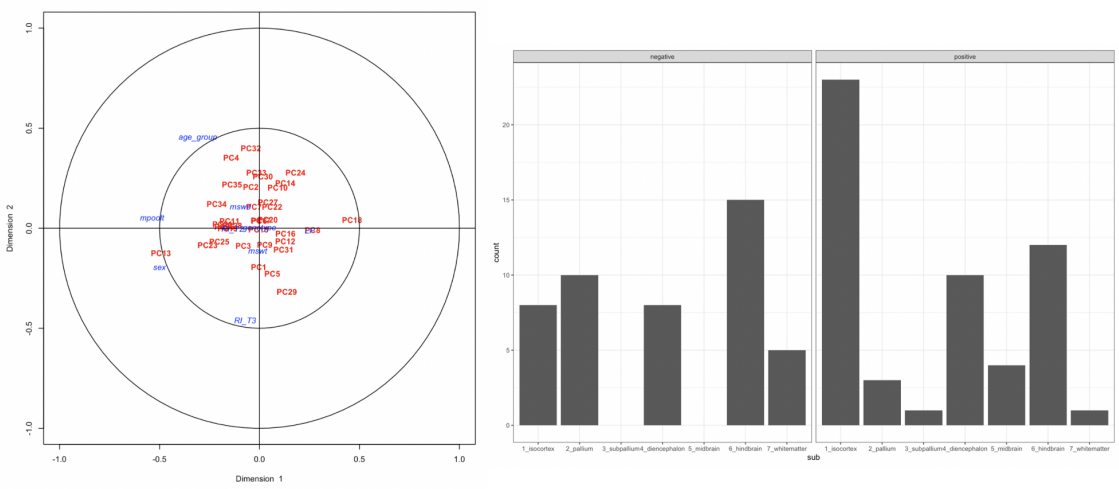

Figure 4: Results from Canonical Correlation Analysis (Left) and the 13th Principle Components(Right)

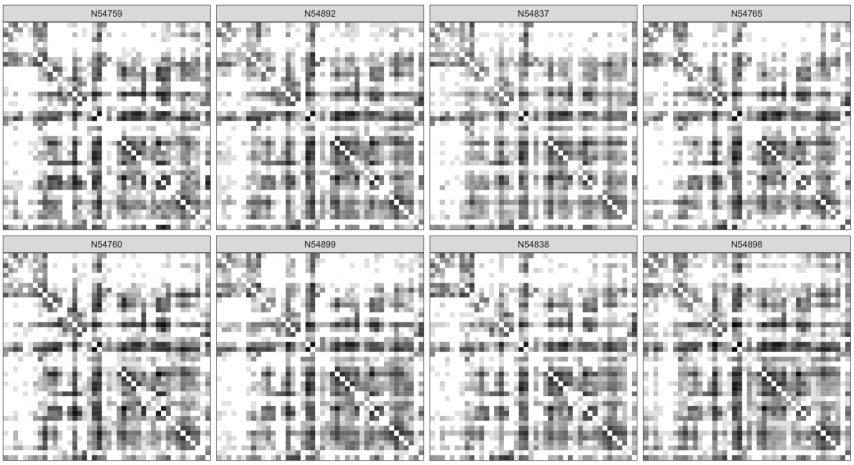

Figure 5: Connectomes within Isocortex Region Sorted by Average Pool Time

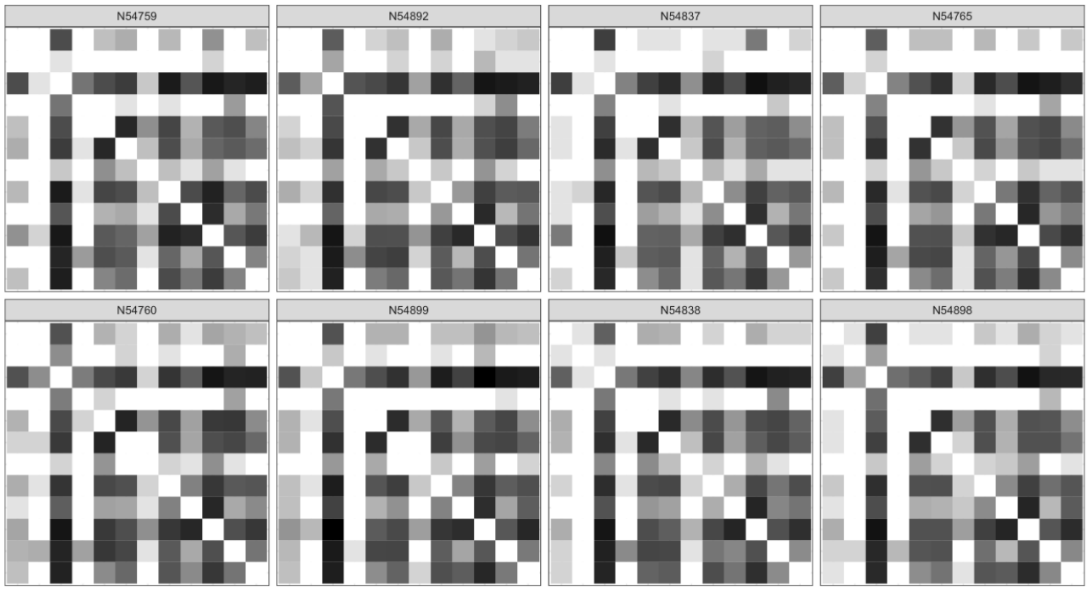

Figure 6: Connectomes within Isocortex Region Sorted by Average Pool Time

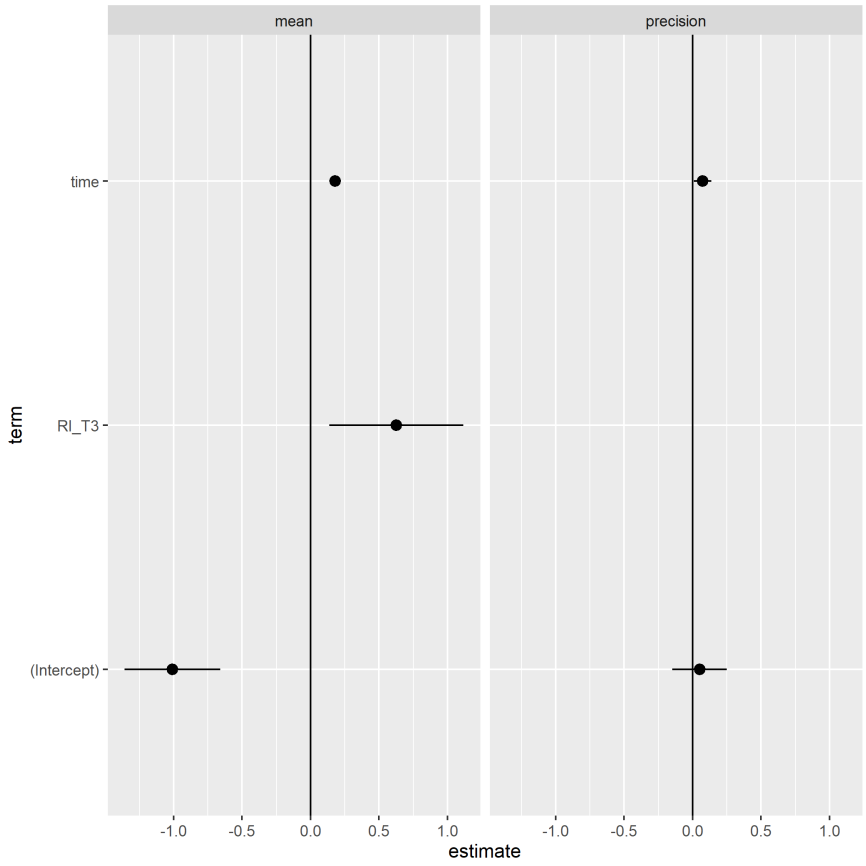

Figure 7: Plot of Beta regression coefficients.