Investigating linked latent constructs in brain connectivity and PTSD symptoms using PLS

Xuan Kan, Ying Guo, Amita Manatunga, Jennifer Stevens, Limin Peng

Emory University, Atlanta, GA 30322, USA

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH

CBIS Center for Biomedical Imaging Statistics

Introduction

In this work, we investigate the phenotypes of neuroimaging (e.g., brain connectivity) in relation to the self-reported clinical assessments of psychiatric scales of PTSD in an inner-city high-risk population. We employ partial least square (PLS) regression method to identify latent factors that explain connections between neuroimaging phenotype and the PTSD psychiatric scales, while accommodating the high dimensionality of the neuroimaging phenotype. We conduct a comprehensive analysis of one dataset in this work. We identify 5 latent components that are associated with sub-clusters of PTSD symptoms. Neural connections that drive each of the latent components are visualized in the brain.

The sample size is 95, including 68 healthy subjects, 27 PTSD patients. All of them are women. Each sample includes a brain imaging and a survey.

Partial Least Squares Regression(PLSR)

Given the original dataset, which size is n, each sample is a $v \times v$ matrix. Since this matrix is a symmetric matrix, we can extract the upper triangular part of the matrix and convert to a m dimension vector, where $m = 0.5 \times (v-1) \times v$. Overall, the input feature $\boldsymbol{X}$ is $\boldsymbol{X} = \{x_1, \cdots, x_n\}$ where $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ and the label is $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$ where $\boldsymbol{Y} \in \mathbb{R}^{n \times c}$, n is the sample size.

Prescreening. Since the number of dimension m of $\boldsymbol{X}$ is really high comparing with the sample size n, $m >> n$. Prescreening is a necessary step to evaluating the quality of each feature, then remove redundant or irrelevant ones. Here we use the correlation between the feature and label to select these highly correlated features. For the i-th column of $\boldsymbol{X}$, its correlated score r_i with $\boldsymbol{Y}$ can be calculated as below,

$$r_i = \max_{j \in \{1, \cdots, c\}} \operatorname{cov}(\boldsymbol{X}_{*,i}, \boldsymbol{Y}_{*,j}) \tag{1}$$

where cov is a function to calculate the pearson correlation coefficient. Next, given a threshold r^*, Columns where $\text{col} = \{i \mid r_i > r^*\}$ in $\boldsymbol{X}$ can be selected. Finally, we can get the new feature $\widetilde{\boldsymbol{X}} = \boldsymbol{X}_{*,\text{col}}$, where $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n \times k}$ and $k << m$.

Partial Least Squares Regression. After feature selection, the PLSR is applied to $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{Y}$. Here the component number we use is l, then we can get,

$$\widetilde{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{P}^\top + \boldsymbol{E}, \boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^\top + \boldsymbol{F} \tag{2}$$

where $\boldsymbol{T} \in \mathbb{R}^{n \times l}$, $\boldsymbol{U} \in \mathbb{R}^{n \times l}$, $\boldsymbol{P} \in \mathbb{R}^{k \times l}$ and $\boldsymbol{Q} \in \mathbb{R}^{c \times l}$. $\boldsymbol{P}$ and $\boldsymbol{Q}$ are loading matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. $\boldsymbol{E}$ and $\boldsymbol{F}$ are error terms.

Edge Selection

From $\boldsymbol{P}_{i,j}$, we can know how a edge i contribute to the component j. We assumes that these more relevant edges will have greater weight in $\boldsymbol{P}$. Hence, we can select these top relevant edges in each component based on $\boldsymbol{P}$. The gained a more robust result, we train the PLSR T times with the o folders cross-validation. Then we obtain $P = \{\boldsymbol{P}^1, \cdots, \boldsymbol{P}^{T*o}\}$,

$$\overline{\boldsymbol{P}} = \frac{\sum_{i=1}^{T*o} \boldsymbol{P}^i}{T*o}, \overline{\boldsymbol{P}} \in \mathbb{R}^{k \times l} \tag{3}$$

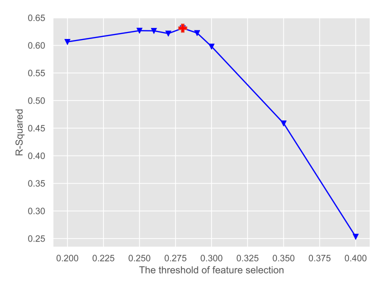
With $\overline{\boldsymbol{P}}$, we can select these top relevant edges $\boldsymbol{E}_i$ for each component i with a threshold τ,

$$Z_i = \text{z-score}(\overline{\boldsymbol{P}}_{*,i}), Z_i \in \mathbb{R}^k; \boldsymbol{E}_i = \{j \mid (Z_i)_j > \tau\} \tag{4}$$

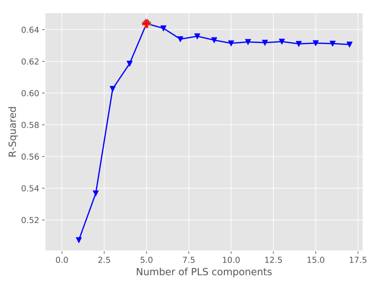
where z-score is a function to compute the z score of each value in a given vector, τ is the chosen threshold.

Hyperparameter Setting

There are two hyperparameters that need to be tuned, the number of latent component l of PLSR and the threshold of feature selection. We employ a grid search to find the optimal number of these two hyperparameters separately.



(a)



(b)

After the grid search shown in the above figure. the optimal threshold r^* of feature selection is 0.28 and the optimal number of component l is 5.

Result

In Table 1, we show the loading matrix $\boldsymbol{Q}$ in a PLSR trained with the whole dataset. From this table, we can see that the Comp1 is almost equally correlated with different PSS subdimension scores, but from Comp2 to Comp5, each of them is correlated with one of these PSS subdimension scores respectively.

Component	Comp1	Comp2	Comp3	Comp4	Comp5
INTRUSIVE	0.089	0.064	-0.033	**0.109**	-0.048
AVOIDANCE	0.082	**0.100**	0.087	-0.004	-0.065
NEGATIVE AFFECT	0.100	0.056	0.023	0.039	**0.090**
HYERAROUSAL	0.086	0.076	**-0.106**	-0.039	0.008

Table: The loading matrix $\boldsymbol{Q}$ of $\boldsymbol{Y}$ in a tuned PLSR.

Visualization

Here we visual these selected edges where τ is 2.32.

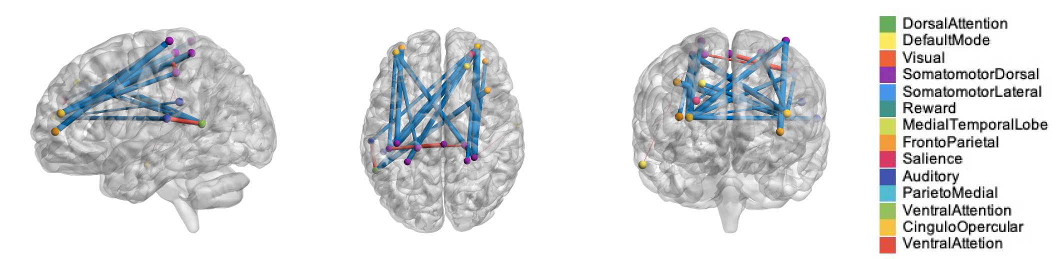


Figure: Visualizing these top ($\tau = 2.32$) brain connectivity edges of the Component1 in the PLSR. 19 edges are selected.

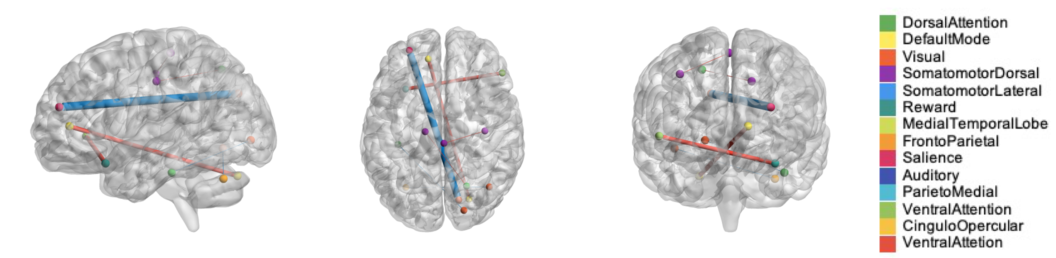


Figure: Visualizing these top ($\tau = 2.32$) brain connectivity edges of the Component2 in the PLSR. 17 edges are selected.

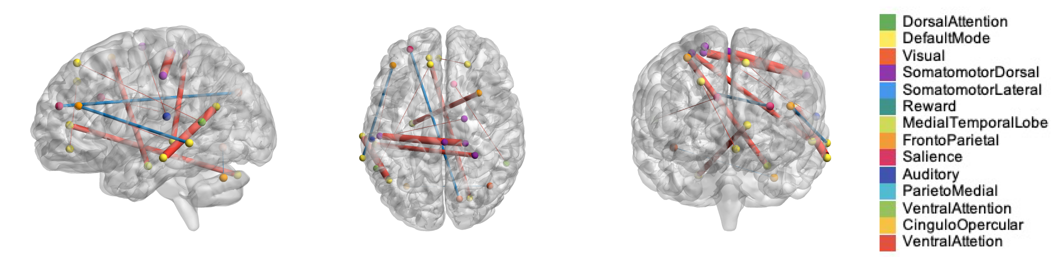


Figure: Visualizing these top ($\tau = 2.32$) brain connectivity edges of the Component3 in the PLSR. 19 edges are selected.

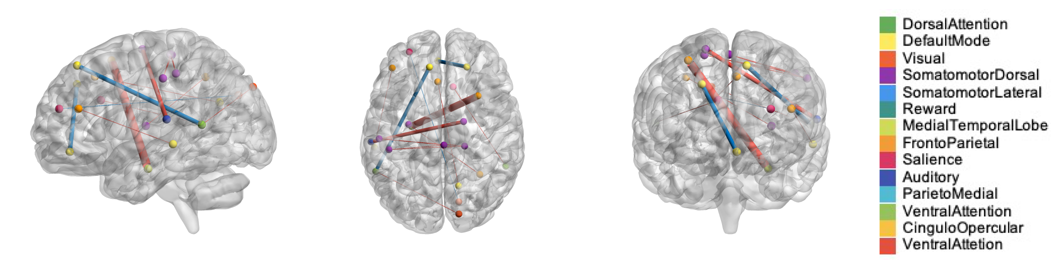


Figure: Visualizing these top ($\tau = 2.32$) brain connectivity edges of the Component4 in the PLSR. 19 edges are selected.

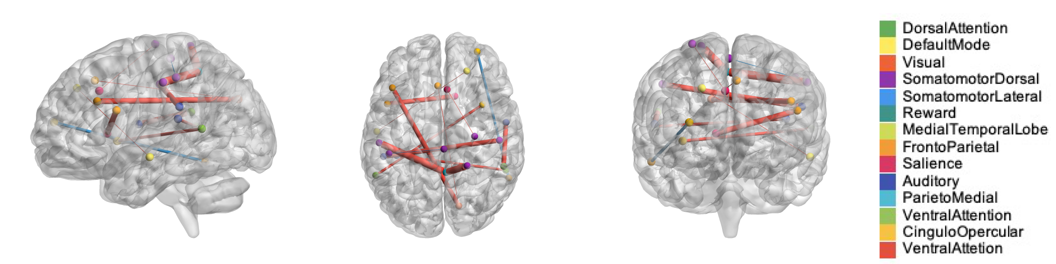


Figure: Visualizing these top ($\tau = 2.32$) brain connectivity edges of the Component5 in the PLSR. 22 edges are selected.