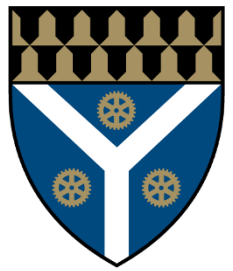
Imputation of Missing Behavioral Measures in Connectome-based Predictive Modelling

Qinghao Liang[1] , Dustin Scheinost[2 1]

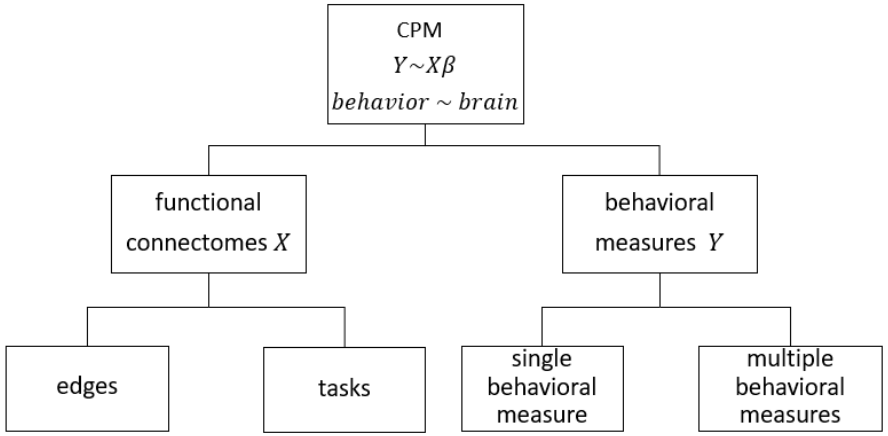
[1]Department of Biomedical Engineering, Yale University, New Haven, CT
[2]Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT

Yale



Introduction

- Currently, most fMRI studies only consider complete cases. In other words, participants with missing behavioral or imaging data are simply removed from analysis, introducing potential selection biases, reducing statistical power, and hurting generalization.
- In this work, we introduce a data imputation step to connectome-based predictive modeling (CPM) to improve brain-based models of behavior by including participants with missing data in model training



Dataset

1. Human Connectome Project (HCP)
500 subjects, 7 tasks, 10 behavioral measures
standard preprocessing pipeline
268-node parcellation
2. Consortium for Neurospsychiatric Phenomics (CNP)
172 subjects, 6 tasks, 7 behavioral measures

method

1. ridge regression Connectome-based Predictive Modeling (rCPM)

- Ten fold cross-validation
- Each connectome is vectorized and the edges are taken as features.
- Then edges of connectivity matrices that are significantly correlated with the phenotypic measure of interest are selected.
- Prediction performance was evaluated by the cross-validated

$$R^2,\ R^2_{CV} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

2. Data imputation methods:

1) mean imputation
2) missForest
   - Implemented in R package **missForest**
3) Regularized Iterative Principal Component Analysis
   - Expectation-maximization algorithm for a PCA fixed-effects model

$$\hat{z}_{ij}^{rPCA} = \sum_{s=1}^{S}\left(\frac{\lambda_s - \frac{np}{min(n-1,p)}\hat{\sigma}^2}{\lambda_s}\right)\sqrt{\lambda_s}u_{is}v_{js}$$

$$= \sum_{s=1}^{S}\left(\sqrt{\lambda_s} - \frac{\frac{np}{min(n-1,p)}\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)u_{is}v_{js}$$

   - Implemented in R package **missMDA**

Simulation

$$Y = \{y_1, y_2, \dots, y_n\}$$

1.Single behavioral Measure y_k

- $\{X, Y\}_{obs}$, $\{X, Y\}_{miss}$
- $\{X, Y\}_{obs} \rightarrow \{X, Y\}_{train_0}$, $\{X, Y\}_{test}$
- concatenate $\{X, Y\}_{train_0}$, $\{X, Y\}_{miss} \rightarrow \{X, Y\}_{train}$
- impute $\{Y\}_{train}$
- $\{X, y_k\}_{train}$, $\{X, y_k\}_{test}$

2.Latent Phenotype

- $\{X, Y\}_{train}$, $\{X, Y\}_{test}$
- impute $\{Y\}_{train}$
- impute $\{Y\}_{test}$
- $\{Y\}_{train} \xrightarrow{pca} y_{pc_{train}}$
- $\{Y\}_{test} \xrightarrow{pca_{train}} y_{pc_{test}}$
- $\{X, y_{pc}\}_{train}$, $\{X, y_{pc}\}_{test}$

Result

Predicting single behavioral measure

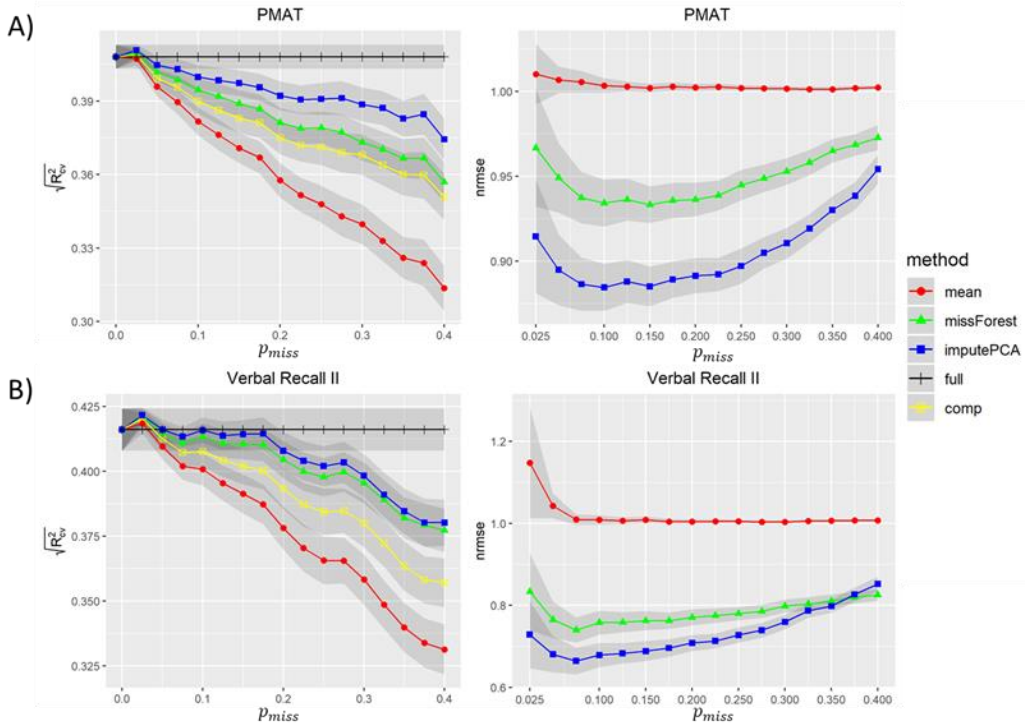


Figure 1. Performance of rCPM with embedded data imputation in predicting A) PMAT (HCP dataset) B) Verbal Recall II (CNP dataset). Prediction performance ($\sqrt{R^2_{cv}}$) and imputation accuracy (nrmse) over a range of missing data percentage from 2.5% to 40% are shown. The shadow areas represent the 95% confidence interval calculated from multiple repeats of missing different data.

Predicting latent variable

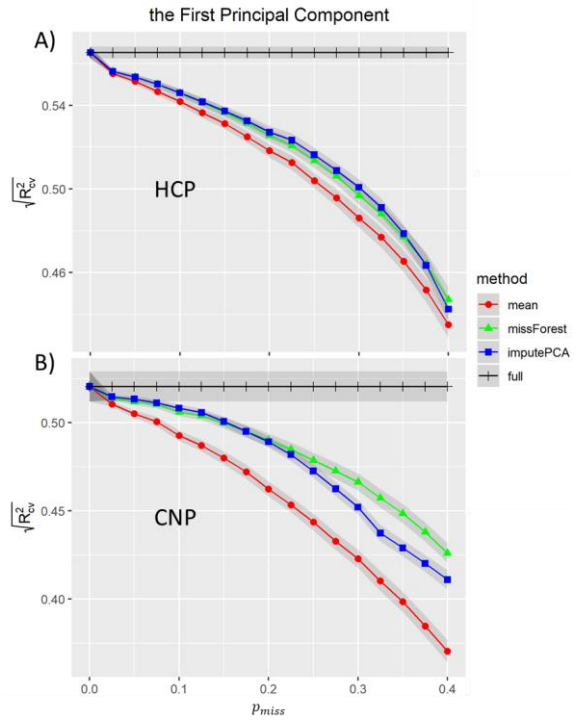


Figure 2. Performance of rCPM when using data imputation in predicting a latent factor (i.e., the 1^{st} principal component) of all behavioral measures in A) HCP dataset and B) CNP dataset over a range of missing data percentages from 2.5% to 40%. The shadow areas represent the 95% confidence interval calculated from multiple repeats of missing different data.

HCP Behavior	full	comp	mean	imputePCA	missForest
PMAT	0.408	0.379	0.361	**0.394**	0.384
ReadEng	0.394	0.378	0.364	**0.394**	0.389
PicVocab	0.457	0.432	0.413	**0.433**	0.430
1^{st}pc	0.565	NA	0.512	**0.519**	**0.519**
CNP Behavior	full	comp	mean	imputePCA	missForest
Verbal Recall II	0.416	0.392	0.378	**0.405**	0.403
CVLT Short	0.377	0.356	0.351	**0.371**	0.368
WMS Digit Span	0.308	0.293	0.284	0.297	**0.298**
1^{st}pc	0.520	NA	0.456	0.478	**0.484**

Table 1. rCPM data performance with different data imputation methods for each tested behavioral variable averaged over all missing data percentage. Bolded values indicate the best performance data imputation method for each tested behavioral variable.

Discussion and Conclusion

- Imputation embedded rCPM using either imputePCA or missForest significantly outperforms simpler methods for handling missing data, such as only using complete cases or mean imputation.

- Future work will include using both the imaging and behavioral data to impute missing behavioral data and testing for cases where the data is not missing completely at random. Overall, our results suggest that data imputation may be valuable for CPM studies with missing behavioral data.

Email: qinghao.liang@yale.edu