PREDICTION METHODS IN LARGE-SCALE NETWORK ANALYSIS FOR NEUROIMAGING DATA

Ziliang Zhu

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:

Hongtu Zhu

Joseph G. Ibrahim

Quefeng Li

Tengfei Li

Weili Lin

ABSTRACT

Ziliang Zhu: Prediction Methods in Large-Scale Network Analysis for
Neuroimaging Data
(Under the direction of Hongtu Zhu)

Brain functional connectivity data are critical for understanding human brain structure and cognitive disease diagnostics. The underlying genetic architecture behind brain functional connectivity is a critical topic in medical studies, which helps unveil the linkages between genetic variants and brain activity and further understand cognitive diseases and brain disorders. The rapid emergence of large scale imaging studies provides researchers with more opportunities to discover the connections between brain system and genes. However, existing methods in imaging genetics are not sufficient in dealing with the high-dimensional data with complex structure, thus limiting the discovery of biological foundation of neuro-development. Therefore, we developed novel statistical approaches for efficient analysis of imaging genetic data. In the first project, we developed a matrix decomposition based method for denoising and recovering the structure of the subject-wise network based on the assumption of factor model. We decompose the subject networks into two parts: a common low-rank basis and subject-specific loadings on the basis. A matrix L0 penalty problem was formulated to accelerate the algorithm. Meanwhile, to avoid iterative computation of high dimensional matrix, we will select a relatively lower dimension basis in the first step, which is a coarse estimator, and then do a fine-tuning in the second step based on the results in step one. In the simulation study, it showed that our approach outperformed other existing approaches in terms of recovering accuracy and computing speed. We also proved that under mild conditions, the algorithm converges fast in an exponential rate. In the second project, we proposed a matrix regression approach for imaging genetic studies. The proposed regression model includes two steps. In the first step, a marginal screening procedure was used to study the univariate associations between genetic variants (SNPs) and imaging phenotype. The theoretical p-value for the marginal screening step was derived using random matrix theories, and important SNPs were selected based on the univariate associations using knock-off. In the second step, a multivariate regression model with all the important SNPs selected as covariates were fitted, and a penalized optimization problem was solved using Nestrov methods. We studied the theoretical properties of the proposed two-stage algorithm thoroughly and simulation studies supported the efficiency and consistency of the proposed method. In the third project, we established a missing data imputation framework to address the issue of missing image modality in real data. The missingness of some imaging modality is common in real imaging data, which may undermine the statistical power in the prediction and inference. However, inaccurate imputation of the missing modality may lead to bias in prediction. Therefore, we thoroughly studied the performance of imputation approaches, including LASSO and ridge models, under different conditions, and concluded the optimal choice of imputation options under the different settings.

ACKNOWLEDGEMENTS

I am grateful to my academic advisors Hongtu Zhu and Joseph Ibrahim for their guidance, encouragement, and support during my study at UNC. I thank Weili Lin for the support during my study at UNC and guidance of statistical collaboration in BRIC. I also thank Quefeng Li and Tengfei Li for their advice and serving in my committee.

I thank my collaborators in UNC BRIC and Nestle for many enjoyable memories. Special thanks to Weiyan Yin for insightful thoughts on research projects. I am grateful to fellows during my internship for giving me guidance on career development and job hunting.

It has been my pleasure to be a member of the Biostatistics and Imaging Genomics Analysis Lab (BIG-S2). I would like to thank all my lab mates in BIG-S2 for priceless friendship. Special thanks to Bingxin Zhao for his enormous help and valuable conversations in discovering interesting scientific questions. Also I'm really grateful to Tengfei Li for the collaboration in all research projects.

I would like to thank my fellow classmates: Jitong Lou, Jiawei Xu, Rujin Wang, Meichen Dong, and my friends: Zengdi Liao, Mengxiao Liu, Haoyang Li, Yujie Su for the great friendship in the past years.

Finally and most importantly, I would like to dedicate this dissertation to my family for their love and support in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In recent years, the emerging of high-dimensional network data has drawn the attraction of researchers in many fields including finance, sociology, geography, and neuro-science Scruggs and Glabadanidis (2003); Patz et al. (2005); Kolaczyk (2009). Powerful statistical tools for analysing these high-dimensional network data can provide a better insight into the research questions associated with these data. Specifically, in the field of neuro-imaging, researchers has found that the functional connectivity between different parts of the brain is associated with cognitive behaviours, brain developments, and neural diseases He et al. (2011); Rogers et al. (2007); Atasoy et al. (2016). A large number of studies have been using brain imaging for detecting potential developmental disorder, or clinical outcomes Vincent et al. (2011); Chupin et al. (2009); Johnson et al. (2012). Teipel found the resting-state fMRI shows a significant difference in the Alzheimer Disease patients, suggesting that neuro-imaging can help diagnose serious diseases Teipel et al. (2017).

However, the real world network data we get are not reliably measured, because the number of nodes in the network is large, while the length of time-series for measuring the correlation is limited. For example, in UK Biobank study, a typical atlas of resting-state fMRI is usually several hundreds (p=90 for AAL Atlas), while the length of each scan is n=468. The high-dimensionality brings in a large amount of noise in the network data we derive Fan et al. (2008), and the noise may lead to unstable results if further statistical analysis is done based on the network. Therefore, it is imperative to develop statistical methods for denoising the covariance matrices.

CHAPTER 2: LITERATURE REVIEW

2.1 High Dimensional Neuro-Imaging Studies

The Human brain has been analyzed from a network perspective with the advent of neuroimaging acquisition techniques and network theory. Functional magnetic resonance imaging (fMRI) is a non-invasive neuroimaging procedure to assess brain neuronal activity that can be measured by changes in blood oxygen level-dependent (BOLD) signal Logothetis et al. (2001). In particular, resting-state fMRI is a stable neural signal when a subject is not performing any tasks. The pairwise correlations of resting-state fMRI between different regions measures the network connection strength and is refered to as connectivity van den Heuvel and Pol (2010), and functional networks could be defined based on these connectivities. It is shown that resting-state networks (RSNs) closely resembles the functional networks of the human brain identified by various sensory, motor and cognitive paradigms Atasoy et al. (2016); Fox et al. (2006); Fox and Raichle (2007).

To understand the human brain connectome and how it relates to other clinical and genetic traits, and it motivated a bunch of large neural-imaging studies, including PNC, PING, UKBiobank, HCP. Also, longitudinal studies are developed in both adults population, including ADNI, and young infants, e.g. BCP. Clinical, genetic, and imaging information are all collected for subjects with an age range from 0-70 years old. A brief summary of these data set are listed in Table 2.1.

PNC study is funded by NIHM and was initially a collaborative research between the Brain Behavior Laboratory at the University of Pennsylvania and the Center for Applied Genomics at the Children's Hospital of Philadelphia. This cohort study focus on characterizing the interaction between brain, behaviours, and genetics. More than $9,500$ subjects aged 8 to 21 years old with diverse medical conditions were enrolled in the study, and 1,445 subjects

Dataset	Data Type	Age Range	Sample Size
PNC	Cross-Sectional	8-21	1445
PING	Cross-Sectional	3-20	1400
UKBiobank	Longitudinal	40-69	500000 (released 40k)
HCP	Cross-Sectional	4-75	1206
ADNI	Longitudinal	50-90	822
BCP	Longitudinal	0-4	500

Table 2.1: Summary of Imaging Cohort Studies

received neuroimaging including fMRI, stuctural MRI, and DTI.

PING study is launched by the UCSD Center for Human Development (CHD) and is funded by NIDA and cofunded by NICHD within NIH. Ten sites throughout the country are involved. The study is aimed at creating a large repository of standardized measurements of behavioral and neuroimaging phenotypes accompanied by whole genome genotyping acquired from 1,493 typically-developing children aged 3 to 20 years.

ADNI study was initially led by Dr. Michael W. Weiner and funded as a private-public partnership between 20 companies and two foundations throught NIH, and NIA. This longitudinal multicenter study was designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimers disease (AD). Neuroimaging including structural MRI, fMRI, T2 weighted imaging, DTI, FLAIR, and ASL were collected.

UK Biobank study was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. The study aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses, including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. Comprehensive information including 1) questionnaire collected data (diet, cognitive function, work history and digestive health), 2) imaging (brain, heart, abdomen, bones carotid artery), 3) electronic health records (cancer, death, hospital episodes, general practice), 4) blood biochemistry (such as hormones cholesterol) from $100,000$ subjects are collected and analysed.

Moreover, genotyping has been undertaken on all $500,000$ participants aged 40 to 69 years old.

In RSNs, researchers are focusing on the following topics:

- Development of RSNs in early infancy
- Identifying regions corresponding to different functional domains
- Constructing the underlying network structures based on resting fMRI signals
- Disease diagnostics and prediction using RSNs
- Genetic foundation study for brain strucures

While resting fMRI data evaluates the connectivities between different gray matter regions using BOLD signals, DTI is another type of brain imaging which measures the diffusion of water molecules in white matter tissue O'Donnell and Westin (2011). Current studies on DTI data mainly focus on

- Generating anatomically plausible tract reconstructions of major projection pathways
- Clustering of white matter tracts generated from DTI imaging
- Disease diagnostics using DTI data
- Genetic analysis on white matter structures

2.2 Multiple Covariance Matrices Estimation

In a neuro-imaging study, an rfMRI scan will be acquired for each subject, thus leading to multiple-covariance matrices data. The challenge here is how we can find a parsimonious way to represent the multiple-covariance data, with the similarities and differences between different subjects kept. There has been bunch of literatures studying a low-dimensional representation for single covariance matrices. For example, PCA finds a low-dimensional representation of the original network and keeps most of the information of the original matrix.

Also, topic model Ke and Wang (2017); Ke (2016); Hofmann (1999) is another approach to find the low-dimensional representation of the network, which focus on finding the typical region representation of each network. Another approach for finding the low-rank structure is stochastic block model (SBM), which aims at finding potential block–wise clusters in the nodes Mao et al. (2018); Holland et al. (1983); Jin et al. (2017).

A direct extension of these approaches could be applied to multiple networks by treating individual networks as independent and apply theese methods separately. This extension is able to reduce the dimension of our networks data significantly, by reducing the dimension for each one separately. However, this kind of extension assumes that all networks are independent, and there is no common information across different brain networks. This assumptionis not reasonable for human brains, since human functional networks could be viewed as replicated graphs Durante et al. (2017) and should share common information. Therefore, we are aiming at find a parsimonious representation of the networks that is able to account for the common basis of the networks, and also extract the individual-specific differences as well.

Factor models are commonly used to model the low-rank network structures for network data Wang et al. (2011); Fan et al. (2011). The factor models Fan et al. (2011) assumes

$$Y_{it} = b_i' f_t + u_{it}$$

where $f_t \in \mathbb{R}^d$ is the random factor and b_i is the factor loading of ith subject. The underlying dimension of the data is equal to the number of effective factors, i.e., the dimension of f_t. Therefore, the dimension of the data could be reduced significantly if $d << p$. The distribution of factors f_t describes the common information across different subjects, while the subject-specific factor loadings contains the subject-specific information.

An similar approach of factor model is called common component analysis Wang et al.

(2011). In CCA, the subject level network data is assumed to have the form

$$X_i = UY_iU' + E_i,$$

where $U \in \mathbb{O}^{p\times r}$ is the orthonormal loading matrix on a low-dimensional data space and $Y_i \in \mathbb{R}^{r\times r}$ is the connectivity matrix for the low-rank space. In this approach, the error E_i is assumed to be Gaussian, and estimates for U and Y_i are attained by minimizing the Frobenius norm of residual connectivity matrices. The optimization problem, which is not convex and involves high dimensional matrix derivitive if using Newton-Raphson approach, was accelarated by using an iterative SVD to approximate the original problem. This approach is fast and can get a reasonably good estimates if the dimension is not extremely high. However, in sup-high dimensional cases, the approach still fails due to computational burden coming from iterative SVD of high dimensional matrices.

To study how well the denoising of connectivity is based on factor model, the most straightforward metric is the Frobenius error of the recovered matrices and true matrices. Besides this, Sin-Θ distance Wedin (1972); Davis and Kahan (1970) was also proposed and used to depict the similarity between the recovered spaces and the original spaces. For two p-dimensional linear spaces $\mathcal{A}$ and $\mathcal{B}$ with orthonormal basis A and B respectively, the principle angle is defined as

$$\Theta(\mathcal{A}, \mathcal{B}) = \text{diag}(\cos^{-1}(\sigma_1), \ldots, \cos^{-1}(\sigma_p)),$$

where $sigma_1 \geq \ldots \geq \sigma_p$ are the singular values of A^TB. Then the Sin-Θ distance is defined as $||\text{Sin}\Theta(\mathcal{A}, \mathcal{B})||$ or $||\text{Sin}\Theta(\mathcal{A}, \mathcal{B})||_F$. It actually measures the similarity between the estimates of common factor space and the true factor space. A rate-optimal upper perturbation bound was derived Cai and Zhang (2018) for SVD for a single matrix.

Another issue with the current denoising algorithms is that the determination of the optimal rank is not yet well-built. Commonly accepted criterion includes EBIC Chen and

Chen (2008) and GCV Josse and Husson (2012). However, both approaches requires an estimation of the degree of freedom of the model, which has been proved to be different from the number of parameters in a low rank approximation Yuan (2016). Adjustment has to be made on the degree of freedom in order to get a consistent estimation of degree of freedom and thus achieve optimal prediction accuracy.

2.3 Network Regression Models

Aside from reducing the dimension of the connectivity matrix, another question of great interest is that how to evaluate the association of the connectivity derived from imaging and clinical or genetic traits. Such association analysis could provide us with information about how to improve brain functional development in childhood and infancy and also predict potential brain malfunction from clinical assessment. However, due to the complexity of network data as mentioned in the previous section, not a lot of efficient approaches has been developed.

One commonly used approach is to extract some summary statistics from the network data, including small-worldness, global efficiency, local efficiency, modularity, etc, and evaluate the association between these statistics and the traits of researchers' interest. These summary statistics provide insight about how the efficient the information is transfered in the brain, and how the brain is functionally segregated. However, since our brain is a complicated world of mixed structures, these extracted features may capture only a small part of the information of our brain, and a lot more information might be neglected. Also, since the original high dimensional connectivity matrices include massive noise, which will further affect the quality of these summary statistics. Therefore, it may further bias the results of association and regression analysis.

Another problem of calculating the summary statistics is that how the network should be defined. A commonly used approach is to define a binary network by including the top connectivities only at a given threshold Achard et al. (2006); van den Heuvel et al. (2017). However, the network could be sensitive to the choice of the threshold Bassett et al. (2012);

Bullmore and Bassett (2011). Therefore, a more stable approach was proposed by considering the graph curves Bassett et al. (2012), which calculates a series of topologic measures across a wide range of thresholds. The mean value of the curve is a typically used summary statistic of the topologic metric Gao et al. (2011).

Another approach to construct the network without using a threshold on the original connectivity matrices is to use sparse representation (SR, Yu et al. (2017); Zhang et al. (2019). In SR, the original time series for each region is assumed to be represented by a small amount of other regions, thus resulting in a optimization problem of minimizing the following loss function.

$$\min_{W} \frac{1}{2}||X - XW||_F^2 + \lambda ||W||_1, \text{ s.t. } diag(W) = 0$$

In neuro-science, previous studies suggests that there are groups of connectivities and regions are closely related and the inference based on these regions should be similar. Therefore, a group level selection approach, namely WGSR Yu et al. (2017); Simon et al. (2013) was proposed by adding a penalty for group selection. The problem is formulated as minimizing the following object problem,

$$\min_{W} \frac{1}{2}||X - XW||_F^2 + \lambda_1 ||cW||_1 + \lambda_2 \sum_{g=1}^{G} ||W_{O_g}||_2, \text{ s.t. } diag(W) = 0$$

where $||W_{O_g}||_2 = \sqrt{\sum_{(i,j)} w_{i,j}^2}$ is the l_2 norm of group g.

The approaches above will give individual specific estimate of networks, thus making the inter-subject variability to be large and resulting in unstable estimates. To overcome this, group sparsity representation Wee et al. (2014) was proposed and reduces the inter-subject variability by forcing the non-zero connectivities across different subjects to be similar. Assuming there are M subjects, then the proposed approach is to minimize the following object function,

$$\min_{W_i} \frac{1}{2}||x_i^m - X_i^m W_i||^2 + \lambda ||W_i||_{2,1}$$

where x_i^m is the time series for ith ROI of subject m, and X_i^m is the time series matrix for subject m excluding region i. Since the subjects may come from different groups, like in an ADHD study, a further improvement allowing for larger between-group variability was proposed (SSGSR, Zhou et al. (2019)).

In SSGSR, the between-group variability could be included, and a comparison between these groups could be done. However, this approach can only work for categorical covariates, and it's extension to continuous covariates is not straightforward.

Aside from using the summary statistics of connectivity matrices, matrix regression analysis using the raw connectivity matrices is another approach and will include more information about the data. Therefore, a few new approaches which use the entire network as response has been developed. Tensor based approach are developed Zhang et al. (2019); Sun and Li (2016) to handle to specific network structure. In Zhang's approach, the regression coefficients are assumed to be tensors as well, such that the network stucture is kept in the space of regression coefficients, i.e. the associations. Within the GLM framework, an low-rank interect, which stands for a low-rank population baseline or mean is assumed and the coefficient tensors were assumed to be sparse. A Newton-Raphson algorithm based MLE was used and a further post hard thresholding was applied. In this model, the assumption is that each variable contributes to only a few of the connectivities. The problem for this model is that if the true underlying structure is not sparse, but low-rank, than this algorithm is not easy to handle, and ignoring the low-rank structure may lead to overfitting of the data.

Another Low-rank based approach, Multi-Scale Network Regression (MSNR,Xia et al. (2019)) was proposed where a low-rank community structure was assumed. The model is formulated as

$$A^i = \Theta + \sum_{f=1}^{q} X_i^f \cdot (W\Gamma^f W') + \epsilon^i, i = 1, \ldots, n$$

where $A^i \in \mathbb{R}^p \times p$ is the network matrix of ith subject, Θ is a low-rank baseline network, X_i^f is the fth covariate for ith subject, and $W \in \mathbb{R}^p \times r$ is a known community structure matrix. $\Gamma^f \mathbb{R}^r \times r$ is the effects of X^f on the connectivities of r communities. The assumption of MSNR is that each of the p nodes belongs to exactly one of the r communities, and the allocation of the nodes are known. However, in a real network, the community structure of the nodes may not be known, so that this assumption is too strong in practice.

The recent L2RM approach Kong et al. (2020) is another low-rank based approach for the network regression problem. The model is formulated similarly, but without any assumptions about the structure of the regression coefficients, which makes the model to be more flexible. A marginal screening procedure is performed before the estimation, to remove the redundant covariates information and make the inference more stable. However, a problem for this approach is that the screening procedure is performed with permutation test, so that it requires large computational resources.

2.4 Missing Data in High-Dimensional Inferences

In large scale neuro-imaging studies, missing data is a major issue that researchers have to face. Due to limited budget, participants' availability, data quality, or data management, some of the information or modalities might be missing for some subjects. With missing data, we are not able to directly use the subject for inference or statistical learning, unless carefully handling it. For example, in ADNI study Petersen et al. (2010), out of the 1628 subjects with imaging data, only 783 participants have both resting-state fMRI and DTI imaging. On the other hand, studies Zhang et al. (2021); Enciso-Olivera et al. (2021) have shown that combining DTI and resting-state fMRI data to predict AD outperforms using a single channel models. Therefore, developing a powerful imputation approach to handle missing imaging modality is critical.

Another situation in imaging genetics where missing data is commonly involved is

Transcriptome-Wide Association Studies (TWAS) Gamazon et al. (2015); Gusev et al. (2016), where gene expression data were used to study the genetic association with phenotypic features. However, due to technical and financial limit in the ability to collect gene expression data, the gene expression data for only a small subset of the participants are collected. The gene expression information for most of the participants needs to be imputed from SNP data, which is collected for all subjects.

Specifically, assume Y is a vector for phenotype, X is an $n \times p$ matrix for population SNP information, and Z is a $n \times q$ matrix for population gene expression information. TWAS is aiming at fitting the model

$$Y = X\beta + Z\gamma + \epsilon$$

However, due to the availability of gene expression data, the matrix Z is not observed, and needs to be estimated as $\hat{Z} = f(X)$, which is typically estimated as a linear function $f(X) = XD$, where D is a $p \times q$ matrix of covariates. The final model to be fitted is

$$Y = X\beta + \hat{Z}\gamma + \epsilon$$

There has been wide research about missing data imputation. The most widely used approach is to impute the missing values using a similar subject with complete data Joenssen and Bankhofer (2012) or imputing with mean value Kalton (1983). However, these methods does not work in high-dimensional missing settings as in imaging genetics. Regression models are also commonly used for imputation. However, in high-dimensional settings, the appropriateness of simple regression models is doubtful.

Beyond the fore-mentioned traditional approaches, some machine learning based approaches have been proposed to handle missingness more effectively. Matrix completion methods based on matrix factorization has been proposed Candès and Tao (2009). The approach is based on a convex relaxation of PCA, and a low-rank structure was assumed for the data. As extensions to regression models, advanced machine learning techniques have been used for imputation, including random forest Stekhoven and Bühlmann (2011) and SVM Yang et al. (2012). Beyond these approaches, generative adversarial networks (GAN) are also used for high-dimensional imputation Yoon et al. (2018); Dong et al. (2021), due to its power of handling complicated missing mechanism.

However, due to the complexity of imaging genetics data, how these methods could help inference is not studied and arbitrary use of these methods without a justification may lead to inaccurate or even biased conclusions. Therefore, we are eager in an investigation of the performance of different methods on imaging genetics imputation.

CHAPTER 3: FACTOR MODEL FOR MULTIPLE NETWORK DATA

3.1 Introduction

In recent years, the emerging of high-dimensional network data has drawn the attraction of researchers in many fields including finance, sociology, geography, and neuro-science Scruggs and Glabadanidis (2003); Patz et al. (2005); Kolaczyk (2009). Powerful statistical tools for analysing these high-dimensional network data can provide a better insight into the research questions associated with these data. Specifically, in the field of neuro-imaging, researchers has found that the functional connectivity between different parts of the brain is associated with cognitive behaviours, brain developments, and neural diseases He et al. (2011); Rogers et al. (2007); Atasoy et al. (2016). A large number of studies have been using brain imaging for detecting potential developmental disorder, or clinical outcomes Vincent et al. (2011); Chupin et al. (2009); Johnson et al. (2012). Teipel found the resting-state fMRI shows a significant difference in the Alzheimer Disease patients, suggesting that neuro-imaging can help diagnose serious diseases Teipel et al. (2017).

However, the real world network data we get are not reliably measured, because the number of nodes in the network is large, while the length of time-series for measuring the correlation is limited. For example, in UK Biobank study, a typical atlas of resting-state fMRI is usually several hundreds (p=90 for AAL Atlas), while the length of each scan is n=468. The high-dimensionality brings in a large amount of noise in the network data we derive Fan et al. (2008), and the noise may lead to unstable results if further statistical analysis is done based on the network. Therefore, it is imperative to develop statistical methods for denoising the covariance matrices.

In a neuro-imaging study, an rfMRI scan will be acquired for each subject, thus leading to multiple-covariance matrices data. The challenge here is how we can find a parsimonious way to represent the multiple-covariance data, with the similarities and differences between different subjects kept. There has been bunch of literatures studying a low-dimensional representation for single covariance matrices. For example, PCA finds a low-dimensional representation of the original network and keeps most of the information of the original matrix. Also, topic model Ke and Wang (2017); Ke (2016); Hofmann (1999) is another approach to find the low-dimensional representation of the network, which focus on finding the typical region representation of each network. Another approach for finding the low-rank structure is stochastic block model (SBM), which aims at finding potential block–wise clusters in the nodes Mao et al. (2018); Holland et al. (1983); Jin et al. (2017).

A direct extension of these approaches could be applied to multiple networks by treating individual networks as independent and apply theese methods separately. This extension is able to reduce the dimension of our networks data significantly, by reducing the dimension for each one separately. However, this kind of extension assumes that all networks are independent, and there is no common information across different brain networks. This assumptionis not reasonable for human brains, since human functional networks could be viewed as replicated graphs Durante et al. (2017) and should share common information. Therefore, we are aiming at find a parsimonious representation of the networks that is able to account for the common basis of the networks, and also extract the individual-specific differences as well.

To account for the difference of networks across different subjects, but maintaining a common low-rank structure, a Multiple Random Eigen Model (MREG, Wang et al. (2019)) was proposed. In MREG, each network was decomposed as

$$S_i = U\Lambda_i U'$$

where $U \in \mathbb{R}^{p\times r}$ is a loading matrix and Λ_i's are diagonal matrices representing the individual specific information for each subject. Although MREG models the common information across different subjects, the model has several drawbacks. First of all, the common information is not assumed to be orthogonal to each other, which is in contrast to the network structures in reality. Typically in a neuro-imaging study, our brain are separated into different disjoint regions, which means all these regions are spatially mutually exclusive, thus orthogonal to each other. On the other hand, the subject-specific community connection matrices is assumed to be diagonal, which means that for each subject, all communities are independent of each other functionally. This assumption is against the brain functional structures, where different networks are usually connected each other. Therefore, although the MREG provides a efficient algorithm for multiple matrix decomposition, the general assumption of this model is not adequate for functional connectivity modeling.

To overcome the issue of disjoint region parcellations, factor models, which assumes orthogonal community memberships, are commonly used to model the low-rank network structures for network data Wang et al. (2011); Fan et al. (2011). The factor models Fan et al. (2011) assumes

$$Y_{it} = b_i' f_t + u_{it}$$

where $f_t \in \mathbb{R}^d$ is the random factor and b_i is the factor loading of ith subject. The underlying dimension of the data is equal to the number of effective factors, i.e., the dimension of f_t. Therefore, the dimension of the data could be reduced significantly if $d << p$. The distribution of factors f_t describes the common information across different subjects, while the subject-specific factor loadings contains the subject-specific information.

An similar approach of factor model is called common component analysis Wang et al. (2011). In CCA, the subject level network data is assumed to have the form

$$X_i = UY_iU' + E_i,$$

where $U \in \mathbb{O}^{p\times r}$ is the orthonormal loading matrix on a low-dimensional data space and $Y_i \in \mathbb{R}^{r\times r}$ is the connectivity matrix for the low-rank space. In this approach, the error E_i is assumed to be Gaussian, and estimates for U and Y_i are attained by minimizing the Frobenius norm of residual connectivity matrices. The optimization problem, which is not convex and involves high dimensional matrix derivitive if using Newton-Raphson approach, was accelarated by using an iterative SVD to approximate the original problem. This approach is fast and can get a reasonably good estimates if the dimension is not extremely high. However, in sup-high dimensional cases, the approach still fails due to computational burden coming from iterative SVD of high dimensional matrices.

In this paper, we made the following contributions. First of all, we build a L0FM framework for high-dimensional connectivity matrices decomposition based on factor model with matrices version of l_0 penalty. Next, we gave theoretical guarantee of the L0FM estimator. Third, we adjusted the degree of freedom estimator in multiple matrices decomposition problem. Finally, we showed that our proposed L0FM approach outperforms existing approaches for large-scale matrices decomposition methods with numerical studies.

3.2 Model

Assume there are I individuals, and for each individual, a time series $X_i \in \mathcal{R}^{T\times p}$ is observed, where p is the number of regions of interest (ROI) and T is the number of timepoints in the time series. We assume the time series comes from a low-rannk factor model

$$X_i = F_i B' + E_i, i = 1, \ldots, I \tag{3.1}$$

where $F_i \in \mathcal{R}^{T\times r}$ is the time series in the low rank underlying basis, which consists of only r components. We assume these r components are independent of each other, i.e., $Cov(F_i) = \Lambda_i = diag(\Lambda_{i1}, \ldots, \lambda_{ir})$. In brain functional analysis, these r components play similar roles as independent networks. $B \in \mathcal{O}^{p\times r}$ is an orthogonal matrix, and characterizing the loading of each ROI on the independent basis. E_i is a noise matrix with i.i.d. elements from $N(0, \sigma^2)$ and independent of F_i.

Under the setting of model (1), we have that the covaraince matrix of X_i is

$$\Sigma_i = B\Lambda_i B' + \sigma^2 I$$

Our goal is to estimate the covariance structure Σ_i from the sample covariance matrix S_i.

Under the low-rank structure, the loss function can be written as

$$\begin{aligned} L(B, \Lambda, \sigma^2, \gamma) =& \sum_{i=1}^{I} ||\Sigma_i - S_i||_F^2 + \gamma \text{rank}(\Lambda_1) \\ =& \sum_{i=1}^{I} ||B\Lambda_i B' + \sigma^2 I - S_i||_F^2 + \gamma \text{rank}(\Lambda_1) \end{aligned}$$

3.3 Estimation

The optimization of this object function is not straightforward, given the non-convex behavior in the parameters. Therefore, we update the parameters block-wisely in each iteration.

3.3.1 Estimating B

Given $\sigma^{2(k)}$ be the estimator of σ^2 after the k-th iteration, let $S_i^{(k)} = S_i - \sigma^{2(k)} I$. We can rewrite the loss function as

$$L_1(B, \Lambda, \gamma) = \sum_{i=1}^{I} ||B\Lambda_i B' - S_i^{(k)}||_F^2 + \gamma \text{rank}(\Lambda_1)$$

To solve this problem, we first introduce the following lemma.

Lemma 1. When a rank-k orthogonal matrix B is given, the rank-k optimizer of Λ_i for the following problem

$$\min_{\Lambda_i} \sum_{i=1}^{I} ||B\Lambda_i B' - S_i||_F^2$$

is given by $\hat{\Lambda}_i = B' S_i B$.

From Lemma 1, by plugging in $\hat{\Lambda}_i$ to L_1, we have the following loss function

$$\tilde{L}_1(B,\gamma) = \sum_{i=1}^{I} ||BB'S_i^{(k)}BB' - S_i^{(k)}||_F^2 + \gamma\text{rank}(B)$$

The optimization of this function is difficult, but we can use approaximation to update it iteratively. Actually we can rewrite

$$\begin{aligned}
\tilde{L}_1(B,\gamma) =& \sum_{i=1}^{I} ||BB'S_i^{(k)}BB' - S_i^{(k)}||_F^2 + \gamma\text{rank}(B) \\
=& \sum_{i=1}^{I} \text{tr}(BB'S_i^{(k)}BB' - S_i^{(k)})^2 + \gamma\text{rank}(B) \\
=& \sum_{i=1}^{I} \text{tr}(BB'S_i^{(k)}BB'S_i^{(k)}BB' - 2BB'S_i^{(k)}BB'S_i^{(k)} + S_i^{(k)2}) + \gamma\text{rank}(B) \\
=& \sum_{i=1}^{I} \text{tr}(-BB'S_i^{(k)}BB'S_i^{(k)}) + \gamma\text{rank}(B) + C \\
=& -\text{tr}(B'(\sum_{i=1}^{I} S_i^{(k)}BB'S_i^{(k)})B) + C \\
=& -\text{tr}(B'A(B)B) + C
\end{aligned}$$

Given a current estimate $B^{(k)}$, we replace $A(B)$ with $A(B^{(k)})$, and minimizing $\tilde{L}_1(B,\gamma)$ is equivalent to maximizing the trace of $B'A(B^{(k)})B$. The rank-r orthogonal matrix maximizing the quantity is the top-r eigenvectors of $A(B^{(k)})$.

However, since $A(B^{(k)})$ is an approximation of $A(B)$ and the top eigen-space could be sensitive to the perturbation. Therefore, when estimating $B^{(k)}$, we will allow for a larger rank R to make sure the actual subspace is included in $B^{(k)}$.

3.3.2 Estimating Λ_i

In the previous section, actually we have already shown that Λ_i could be estimated by

$$\hat{\Lambda}_i = B^{(k+1)\prime}S_i^{(k)}B^{(k+1)}$$

However, in the previous section, we allowed for a larger space for $B^{(k+1)}$, thus the rank of $\hat{\Lambda}_i$ is not optimal and may include redundant information. Therefore, a further tuning of Λ_i is needed. Inspired by the L_0 penalty algorithm in linear model Huang et al. (2018), by solving the KKT condition can significantly improve the performance of this non-convex problem. Similarly, we will find the KKT condition for Λ_i in the matrix case.

Following the proof of Lemma 1, the optimization problem in terms of Λ is

$$L_2(\Lambda) = \sum_{i=1}^{I} ||\Lambda_i - B^{(k+1)\prime} S_i^{(k)} B^{(k+1)}||_F^2 + \gamma ||\text{diag}(\Lambda_1)||_0$$

The penalty is given on the common matrix Λ_1 because we require all subjects share a common subspace.

Lemma 2. The KKT condition for

$$\min_{\Lambda} \sum_{i=1}^{I} ||\Lambda_i - A_i||_F^2 + \gamma ||\text{diag}(\Lambda_1)||_0, i = 1, \ldots, I$$

is given by

$$d = 1[A_2(d \otimes 1_p') + (I - 2D)\text{diag}(A_2) \geq \gamma]$$

where $d = (d_1, \ldots, d_p), D = diag(d)$, and $\hat{\Lambda}_i = DA_iD$. $A_2 = (A_{2,jk})_{p \times p}$ where $A_{2,jk} = \sum_{i=1}^{I} A_{i,jk}^2$.

Following Lemma 2, the optimization of the problem could be done iteratively by satisfying the KKT condition. Given a current candidate index $d^{(k)}$, update

$$d^{(k+1)} = 1[A_2(d^{(k)} \otimes 1_p') + (I - 2D^{(k)})\text{diag}(A_2) \geq \gamma]$$

The rank-r version of the update is

$$d^{(k+1)} = 1[A_2(d^{(k)} \otimes 1_p') + (I - 2D^{(k)})\text{diag}(A_2) >= M_r^{(k)}]$$

where $M_r^{(k)}$ is the r-th largest component of $A_2(d^{(k)} \otimes 1_p') + (I - 2D^{(k)})\text{diag}(A_2)$. The optimal Λ_i will be obtained when the algorithm converges, i.e., $d^{(k+1)}$ remains unchanged after update.

After we get the updated $\Lambda_i^{(k)}$, we should also update $B^{(k+1)}$ by keeping the corresponding columns only, this will reduce the computational complexity during the iteration.

3.3.3 Estimating σ^2

Although in most studies, the term σ^2 is regarded as noise parameter and not modeled in the algorithm. However, the removal of the baseline noise can help increase the signal to noise ratio and enhance the accuracy of subspace estimation.

The update of σ^2 when $B^{(k)}$ and $\Lambda_i^{(k)}$ are given is straightforward, since the loss function is a quadratic function of σ^2. Actually, let $R_i^{(k)} = S_i - B^{(k)}\Lambda_i^{(k)}B^{(k)\prime}$, the loss function of σ^2 is

$$L_3(\sigma^2) = \sum_{i=1}^{I} ||R_i^{(k)} - \sigma^2 I||_F^2$$

Therefore, we can get the closed form solution as

$$\sigma^{2(k)} = \frac{1}{Ip} \sum_{i,j} R_{i,jj}^{(k)}$$

3.3.4 Estimation Algorithm

In this section, we summarize the optimization algorithm in Algorithm 1. [ht] InputInput OutputOutput

```
L0FM($S_i, i = 1 \ldots n$; $R$, $r$,$\epsilon$)
$R$, $r$,$S_i, i = 1 \ldots, I$ $\sigma^2, B, \Lambda_i, i = 1, \ldots, I$
$\sigma^{2(0)} = 0$; $B^{(0)} = \text{Eigvec}_r(\sum_{i=1}^{I} S_i)$;$k = 0$;
$|| \sin(B^{(k)}, B^{(k-1)}) >= \epsilon$ $S_i^{(k)} = S_i - \sigma^{2(k)} I$;
$A^{(k)} = \sum S_i^{(k)} B^{(k)} B^{(k)\prime} S_i^{(k)}$;
$\tilde{B}^{(k+1)} = \text{Eigvec}_R(A^{(k)})$;
$L_i^{(k)} = B^{(k+1)\prime} S_i^{(k)} B^{(k+1)}$;
```

```
$L_2 = (L_{2,jk})_{R\times R}; L_{2,jk} = \sum_{i=1}^{I} L_{i,jk}^{(k)2}$
$d^{(0)} = (1'_r, 0'_{R-r})'$; s=1;
$d^{(s)} \neq d^{(s-1)}$
$D^{(s-1)} = \text{diag}(d^{(s-1)})$
Let $M_r^{(s-1)}$ be the $r$-th largest component of $L_2(d^{(s-1)} \otimes 1'_p) + (I - 2D^{(s-1)})\text{diag}(L_2)$;
$d^{(s)} = 1[L_2(d^{(k)} \otimes 1'_p) + (I - 2D^{(s-1)})\text{diag}(L_2) >= M_r^{(s-1)}]$
$B^{(k+1)} = \tilde{B}^{(k+1)}[:, d^{(s)}]$;
$\Lambda_i^{(k+1)} = B^{(k+1)\prime} S_i^{(k)} B^{(k+1)}$;
$R_i^{(k+1)} = S_i - B^{(k+1)} \Lambda_i^{(k+1)} B^{(k+1)\prime}$
$\sigma^{2(k+1)} = \frac{1}{Ip} \sum_{i,j} R_{i,jj}^{(k+1)}$
k=k+1;
```

L0FM Algorithm

3.3.5 Tuning Parameters

In Algorithm 1, we need to specify two tuning parameters r and R, which is the rank of final model and the rank of the pre-screened model, respectively. The choice of R is less important, since the goal for pre-screening is purely to do a dimension reduction to save computational resources. Therefore, we can choose a moderate R, and we suggest to use $2\sqrt{p}$.

However, the choice of the rank of final model is critical, because it determines what the underlying structure of the data is. A incorrect specification of r may lead to missing information of data or over-fitting to the data. Information type of criterion are typically used to determine the optimal tuning parameters. In these criterion, the degree of freedom is always used to measure the complexity of the model. Typically, the number of free parameters is used as the degree of freedom in most questions Chen and Chen (2008). However, in low-rank matrix type of problems, people have shown that the number of free parameters is a biased estimator of 'true' degree of freedom, even in a single matrix case Yuan (2016). Therefore, we need to adjust for the degree of freedom in L0FM to determine the optimal rank

r. The following lemma gives an asymptotic unbiased estimator for the degree of freedom in L0FM.

Lemma 3. The following estimator is a consistent estimator for degree of freedom in L0FM with rank r.

$$\hat{df} = pr + \frac{r(r+1)(I-1)}{2} + 2\sum_{k=1}^{r}\sum_{l=r+1}^{p}\frac{(\sigma_l^2 - \hat{\sigma}^2)_+}{\sigma_k^2 - \sigma_l^2}$$

where σ_k^2 is the k-th eigenvalue of $\sum_{i=1}^{I} S_i$.

Using Lemma 3, the information type of criterions of L0FM can be calculated using the estimated degree of freedom. Specifically, we will consider the following two criterion. Mallow's C_p is defined as

$$C_p(M) = \sum ||\hat{\Sigma}_i - S_i||_F^2 + 4\hat{\sigma}^2\hat{df}$$

GCV is defined as

$$GCV(M) = \frac{1}{p^2 I - 2\hat{df} - 1}\sum ||\hat{\Sigma}_i - S_i||_F^2$$

The optimla rank could be determined by minimizing C_p or GCV.

3.4 Theoretical Properties

In this paper, we evaluate the performance of our algorithm using Frobinius norm and $\mathrm{Sin}-\Theta$ distance.

For any $\delta \in (0,1)$, let $\tau = \frac{56}{c}[\frac{p}{n}\log\frac{2I}{\delta} + \frac{r}{p}]$.

Conditions.

- All X_i's are sub-Gaussian with variance proxies Σ_i.
- Assume there exist constants $k_1(n,p,I), k_2(n,p,I) \in (0,1)$, s.t.

$$\frac{\tau\sum_{i=1}^{I}\sigma(\Lambda_i)[\sigma(\Lambda_i)+\sigma^2]}{\sigma_{\min}(\sum_{i=1}^{I}\Lambda_i^2)} \leq k_1$$

$$\frac{\tau \sum_{i=1}^{I} [\sigma(\Lambda_i) + \sigma^2]^2}{\sigma_{\min}(\sum_{i=1}^{I} \Lambda_i^2)} \leq k_2$$

and

$$\rho = \frac{2(1-k_1)k_2}{(1-k_1-3k_2)^2 - 18k_2^2} < 1$$

Under the conditions above, we have the following theorem about the convergence of the algorithm.

Theorem 1. (Convergence Rate) The convergence rate of $Sin - \Theta$ distance in Algorithm 1 satisfies

$$\lim_{k\to\infty} \frac{\mathrm{Sin}\Theta(B^{(k)}, B^*)}{\mathrm{Sin}\Theta(B^{(k+1)}, B^*)} \leq \rho$$

Theorem 1 ensures that Algorithm 1 will converge in a exponential rate. The following Theorem 2 and 3 ensures the error bounds of subspace estimation and covariance matrices reconstruction.

Theorem 2. (Sub-Space Estimation Error Bound) The $\mathrm{Sin} - \Theta$ error of subspace estimation of Algorithm 1 satisfies

$$||\mathrm{Sin}\Theta(\hat{B}, B)|| \leq \frac{\rho}{2(1-\rho)}$$

Theorem 3. (Covariance Matrices Reconstruction Error Bound) The error bound of reconstruction the subject-wise low-rank matrix $\hat{\Sigma}_i = \hat{B}\hat{\Lambda}_i\hat{B}'$ of Algorithm 1 satisfies

$$||\hat{\Sigma}_i - \Sigma_i||_F^2 \leq \frac{2\rho}{(1-\rho)}||\Lambda_i||_F + \tau\sqrt{r}[\sigma(\Lambda_i) + \sigma^2]$$

Theorem 1-3 actually tells us if the ratio between the largest eigenvalues of Λ_i and the smallest eigenvalue of Λ_i is a lower order term of the quantity $\max\{\frac{p}{n}\log I, r/p\}$, the algorithm will converge at exponential rate and will result in a consistent estimator.

3.5 Numeric Results

In this section, several numeric studies has been conducted to show the performance of the proposed method.

3.5.1 Simulations

Simulation studies has been done to evaluate the estimation of covariance matrices in terms of both subspace estimation and covariance matrices recovery. The **Sin**Θ distance and the relative error which is defined as

$$RE(\hat{\Sigma}_i, \Sigma_i, i = 1, \ldots, I) = \frac{\sum_{i=1}^{I} ||\hat{\Sigma}_i - \Sigma_i||_F^2}{\sum_{i=1}^{I} ||\Sigma_i||_F^2}$$

were used as metrics for evaluating subspace estimation and covariance matrices recovery, respectively.

Data were simulated from Model (1), were B is a random $p \times r$ orthonormal matrix. Λ_i is generated as

$$\Lambda_i = \Gamma_i \Gamma_i'$$

where $\Gamma_i \in \mathbb{R}^{r \times r}$ with elements from iid $N(0, 1)$. The number of subjects is set to be $I = 1000$, the number of nodes were chosen to be $p = 30, 100, 500, 1000$, the number of true underlying factors is $r = 5$, and the length of signal for each subject is fixed at $n = 1000$.

We compared our L0FM method with several other methods for multiple matrices decomposition, including FGSC Wang et al. (2011), group PCA Smith et al. (2014), and MREG Wang et al. (2019).

Figure 3.1 shows the comparison of **Sin**Θ distances under different methods. When the noise level σ^2 is low, all the methods perform pretty well except for MREG, in which a diagonal structure is assumed on Λ_i. However, as the data becomes noisier, our L0FM approach outperforms FGSC and Group PCA, in both low dimensioal or high dimensional situations. Moreover, we can see that even as the dimension goes higher, the subspace estimation accuracy does not increase significantly.

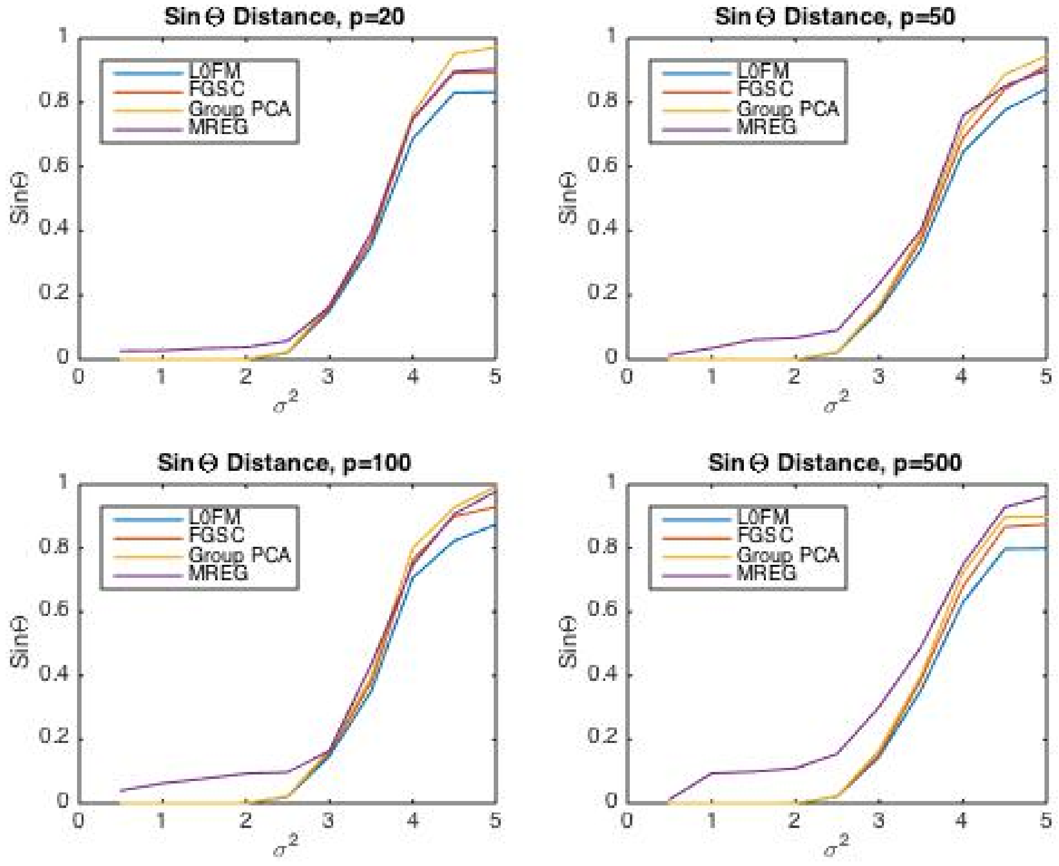

Figure 3.1: Sin Θ Distance of Subspace Estimation

Figure 3.2 shows the relative error of different methods. L0FM outperforms other methods when the noise level is high. In contrast to the **Sin**Θ distance, where MREG performs worst, MREG achieves a comparible performance as L0FM. There might be two potential reasons for this phenomenon. Firstly, MREG aims at recovering the true covariance matrices, thus the relative error is minimized during the optimization algorithm. On the other hand, the simulated Λ_i's are close to diagonal, so that it is close to the model that MREG assumes. Therefore, MREG achieves good relative error although performs not very good in terms of subspace estimation.

Finally, to illustrate the appropriateness of the definition of degree of freedom, we chose the optimal rank based on our adjusted dof and compared it to the optimal rank chosen

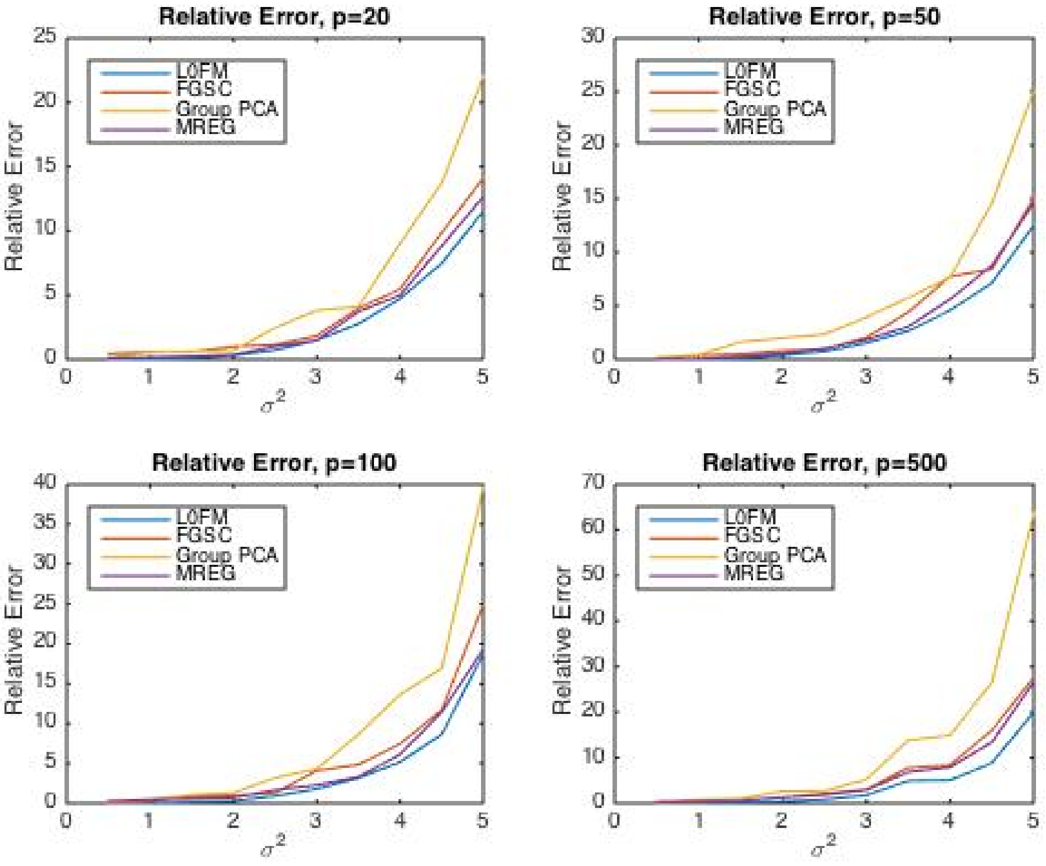

Figure 3.2: Relative Error of Covariance Matices Recovery

by naive method. In naive method, the degree of freedom is equal to the number of free parameters in the model, so it is

$$\hat{df}_{naive} = pr + \frac{r(r+1)(I-1)}{2}$$

We determined the optimal rank using different $\hat{df}$ with two information criterion, the generalized cross validation (GCV) and Bayesian Information Criterion (BIC). The relative error under different rank choice are shown in Figure 3.3. Based on the adjusted degree of freedom, the reconstruction of the covariance matrices are better than using the naive degree of freedom, which suggested that the adjustment is necessary.

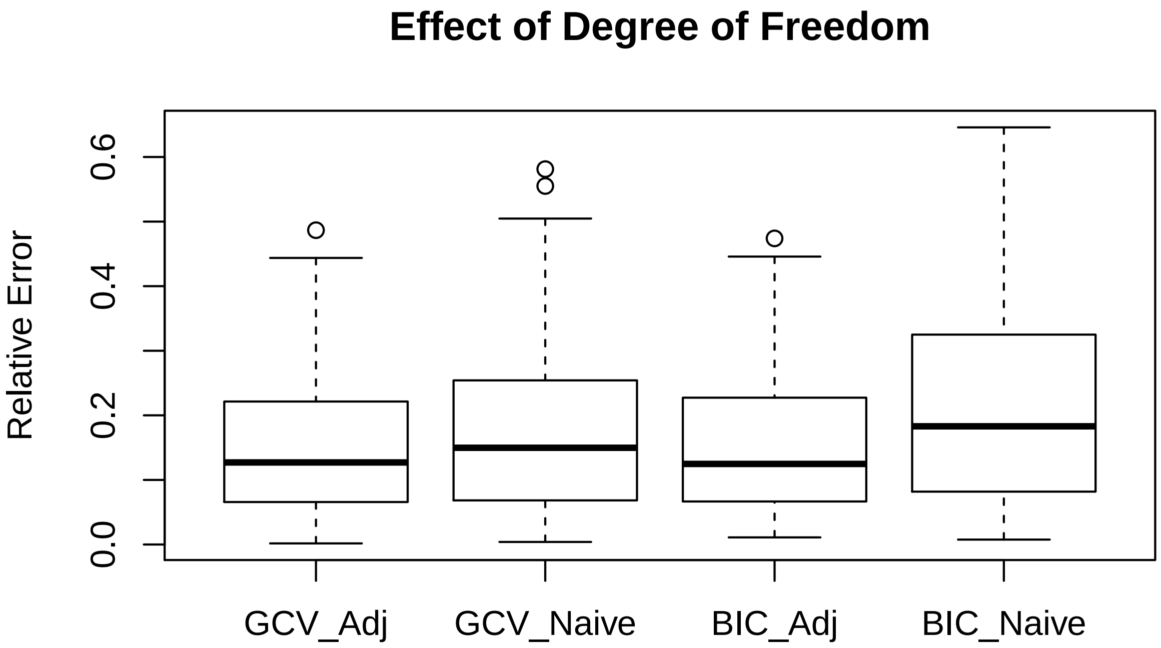

Figure 3.3: Relative Error of Covariance Matrices Recovery Under Different Choice of Optimal Rank

3.5.2 Real Data Analysis: UKBioBank Functional Connectivity

We applied our method on UKBioBank data which involves more than 30k subjects. Functional connectivity was calculated using resting-state fMRI scans for each subject. 37848 subjects with rsfMRI data were included in the study.

The resting-state fMRI data were processed using UKBioBank standard pipeline Alfaro-Almagro et al. (2018). After processing, a 55 × 55 Pearson's correlation matrix for 55 independent network components were acquired for each subjects. Our $L0FM$ approach was applied to the functional connectivity data to further determine the underlying network structures of our brain.

Furthermore, SNP heritability analysis was performed using Λ_i as phenotype to show the genetic contribution on functional connectivities. Non-British subjects were removed from the analysis to minimize the potential ethnicity confounding, and covariates including age, age-squared, sex, age-sex interaction, age-squared-sex interaction, site, and top-40 genetic principle components were adjusted in the analysis. As a result, 31053 subjects were included in the heritability analysis. GCTA tool Yang et al. (2011) was used for calculating genetic heritability.

Results: Based on BIC with the derived degree of freedom, there are 8 underlying networks. The masks of the 8 networks could be found in Figure 3.4. Network 1 and 3 are visual networks, where Network 1 is mainly primary visual cortex, while network 3 involves not only the primary visual cortex, but also the occipital lobe. Network 2 is a mixture of bilateral parietal cortices and precuneus, both of which belongs to the default mode network (DMN). Network 4 is dominated by the sensory-motor network and auditory network. Network 5 is mostly the frontoparietal network and part of the precuneus which belongs to DMN. Network 6 is a mixture of frontal parietal network and primary visual cortex. Network 7 is the mixture of primary visual network, primary motor cortex, and part of auditory area (Wernicke's area). Network 1-7 are more symmetric networks, while network 8 represents a lateralization of the motor area and Wernicke's area. This is in accordance with the previous finding that lateralization of language areas van Ettinger-Veenstra et al. (2010) and motor cortex Tozakidou et al. (2013) are common in brains.

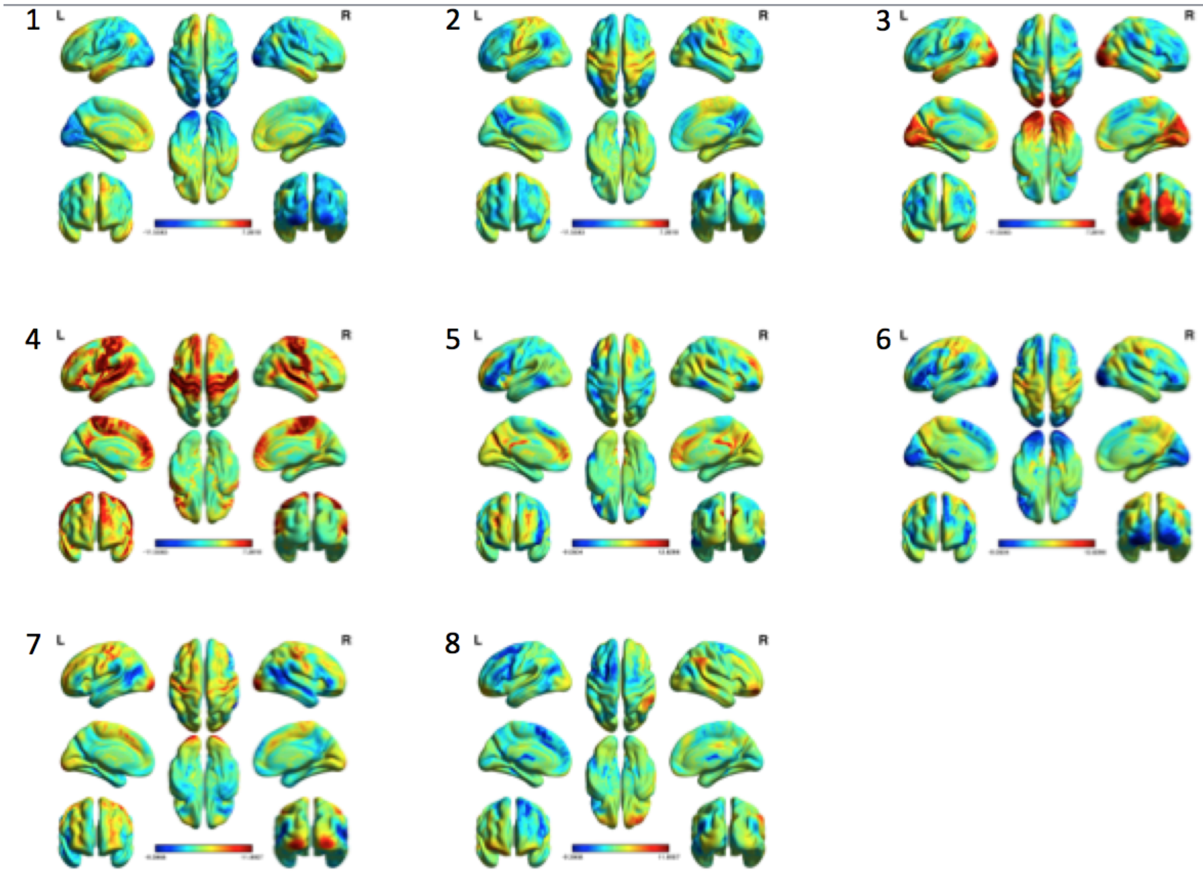

Figure 3.4: Masks of Networks in UKBioBank Resting-State fMRI Data

The heritability analysis suggests that the connectivity between network 2 and 8 ($h^2 = 18.5\%$) and the within network connectivity of network 2 ($h^2 = 16.0\%$) are the most heritable , see Figure 3.5. This finding is in accordance with previous findings that handedness is related to brain cortical lateralization Sainburg (2014) and handedness is genetically heritable Nurhayu et al. (2020); Medland et al. (2009). Meanwhile, the high heritability of within network connectivity of DMN is in accordance with previous findings Teeuw et al. (2019).

3.6 Conclusions

In this paper, we proposed L0FM, a multiple matrices decomposition approach based on factor models. The approach used a matrix form of l_0 penalty, which is based on matrix

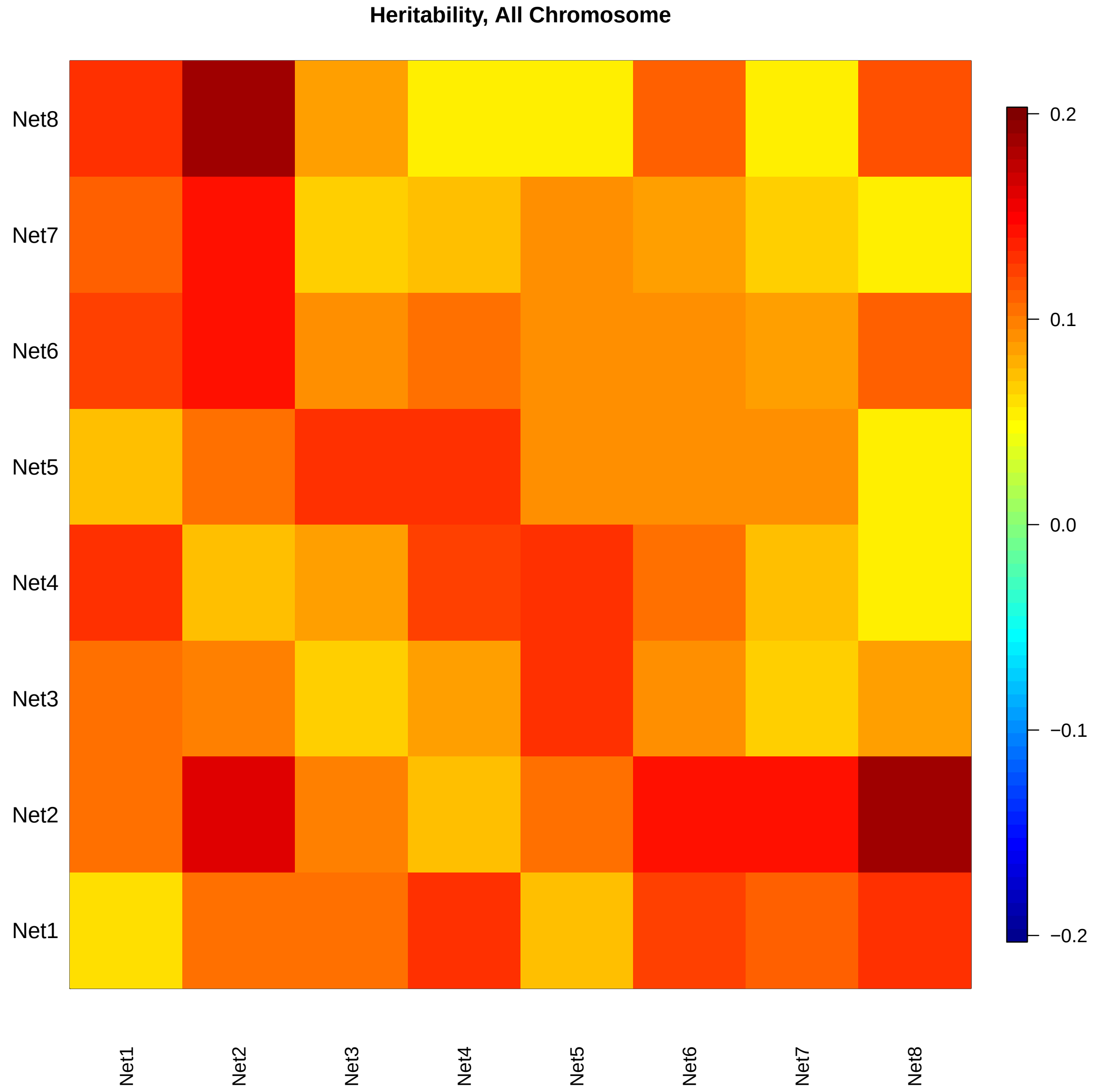

Figure 3.5: SNP Heritability of the Connectivity across 8 Networks in UKBiobank Resting-State fMRI Data

multiplication, so that the computationally complexity is reduced. With the proposed L0FM approach, large scale neuro-imaging datasets like UKBioBank could be handled efficiently. Theoretical properties guarantees that not only the connectivity matrices recovery, but the subspace estimation are consistent with mild condition. This approach gives us possibility to study large scale nerro-image datasets, and learn the underlying brain functional structure with large sample size.

CHAPTER 4: REGRESSION MODELS IN IMAGING GENETICS

4.1 Introduction

In imaging genetics, a fundamental problem of interest is the association between imaging features and genetic biomarkers. Since imaging features provide valuable information about neural development or malfunctions, and thus affecting behavioural outcomes or assisting disease diagnoistics, unveiling the association between imaging features and genetic biomarkers can provide more insights about understanding the biological pathway of cognitive growths and diseases.

However, the high-dimensionality of both the imaging features and genetic variants makes the association analysis become challenging. High-dimensionality brings in cumulatied noise that brings difficulties in separating true signals from spurious noise. Meanwhile, the correlation structures in imaging feature, especially the spatial correlations, makes the problem even harder. Finally, the high-dimensionality also brings challenges in how to do the inference and association analysis efficiently.

There have been a rich body of literature trying to draw inference about the high-dimensional data with complex structures. The simplest way is to perform univariate analysis and use a summary statistics to summarize the results. However, the univariate analysis suffers from the risk of cumulating noises and neglecting complex structure of imaging data. To overcome the issue of complex structure, summary statistics like network efficiency of networks are used for association analysis to reduce dimension Achard et al. (2006); van den Heuvel and Pol (2010); Gao et al. (2011). However, summary statistics may only affecting a certain aspect of the imaging data, and other information might be lost. A series of network regression based approaches Yu et al. (2017); Simon et al. (2013); Zhang et al. (2019); Xia et al. (2019) have been proposed to better model the associations with complex imaging data. However, all of these approaches suffers from computational burden because of high dimensionality. The L2RM approach Kong et al. (2020) used a screening procedure to accelerate the estimation, but still suffers from the computational burden from permutation test.

To fill the gaps as mentioned above, we propose a novel low-rank approach for estimating the association between brain imaging and genetic variants. A screening procedure with theoretical guarantee was proposed to efficiently estimate the association.

4.2 Models

Let $Y_1, \ldots, Y_I \in \mathbb{R}^{m \times n}$ be the imaging feature matrices from I different subjects. For each subject, a p-dimensional vector of covariates $x_i = (x_{i1}, \ldots, x_{ip}$ is observed. In imaiging genetics, the set of covariates include genetic markers (SNPs), age, gender, among others. Without loss of generality, we assume Y_i's have mean 0, and all the covariates are standardized to mean 0 and unit variance.

A low-rank based regression model

$$Y_i = \sum_{j=1}^{p} x_{ij} B_j + E_i, i = 1, \ldots, I \tag{4.1}$$

is considered, where $B_j \in \mathbb{R}^{m \times n}$ is coefficient matrix characterizing the effect of the jth covariate on Y, and $E_i \in \mathbb{R}^{m \times n}$ is the random error matrix with mean 0. Due to the high dimensionality of the covariates and limited sample size in most imaging genetic studies, a few assumptions need to be made to make the model identifiable. Thoughout the paper, we make the following assumptions:

- Only a small portions of the covariates has effect on connectivity matrices, i.e. $\exists S \subset \{1, \ldots, p\}$, where $|S| = s << p$, s.t. for $j \notin S, B_j = 0$.

- For non-zero coefficients $B_j, j \in S$, there is a low-rank structure, i.e., $\texttt{rank}(B_j) << \min(m, n)$

- Noise E_i has distribution

$$E_i = BF_i + G_i$$

where $F_i \in \mathbb{R}^{r \times n}$ and $G_i \in \mathbb{R}^{m \times n}$ has i.i.d. standard normal distribution for each entry.

4.3 Estimation

The estimation of the model given in the previous section suffer from the burden of high dimensionality. Therefore, we separate the estimation procedure into two steps. In the first step, a marginal screening is done by fitting a series of univariate models. The summary statistics will be calculated using random matrix theory and justify the choice of significant variables to enter the next step. In the second step, important features selected in step one will be considered together and a penalized matrix regression model will be fitted using these important features.

4.3.1 Marginal Screening

Since the dimension of covariates p is usually ultra-high in genetics studies (1 million to 10 million SNPs), follow the idea of sure independence screening (SIS) Fan and Lv (2008), a marginal screening step to reduce the dimension of covariates will help enhance the performance by removing redundant information while keeping potential informative predictors.

Similar to ordinary linear models, the correlation between X_j and Y could be used as the testing statistic for marginal screening. The correlation between X_j and Y is calculated element-wise correlation, i.e.,

$$C_j = \sum_{i=1}^{n} x_{ij} Y_i$$

As we can see, C is not a scalar, therefore, we use the largest eigenvalue $T_j = \lambda_{\max}(C_j' C_j)$ as the testing statistic.

This testing statsitics has been proposed by Kong et al. (2020), and has been shown to be robust to signal structure B_j's. However, they failed to derive the theoretical asymptotic null distribution of the testing statistic T_j. As a consequence, a resampling approach was used to construct the null distribution, which requires huge computational resources. In this paper, we derived the theoretical null distribution of the testing statistics, so that no bootstrap or resampling is needed.

Null Distribution of T_j Based on the model assumption, under the condition that $B_j = 0$, we have $x_{ij} Y_i$. Since x_{ij} is standardized to unit length, i.e.,

$$\sum_{i=1}^{j} x_{ij}^2 = 1$$

we have $C_j = \sum_{i=1}^{n} x_{ij} Y_i$ has the same distribution as Y_1.

Note that $Y_1 = BF_1 + G_1$, the columns of Y_1 have i.i.d. distribution from $N(0, BB' + I)$, and thus $C_j C_j'$ has the same distribution as the covariance matrix of $N(0, BB'+I)$ distribution with n observations.

Let $y = \frac{m}{n}$ and α be the largest eigenvalue of $BB' + I$, we have the following lemma about the distribution of largest eigenvalue of $C_j C_j'$ similar to the results shown in **?**.

Lemma 4.1 (Null Distribution of T_j) Under the condition that $\alpha > 1 + \sqrt{y}$ and $x_{ij} \perp Y_i$, as $\min(m, n) \to \infty$,

$$\sqrt{n}(T_j - \lambda) \xrightarrow{d} N(0, \sigma_T^2)$$

where

$$\lambda = \alpha + \frac{y\alpha}{\alpha - 1}$$

and

$$\sigma_T^2 = \frac{2\alpha^2[(\alpha - 1)^2 - y]}{(\alpha - 1)^2}$$

Following Lemma 4.1, we can calculate a normalized summary statistic for each covariate as $Z_j = \frac{\sqrt{n}(T_j - \lambda)}{\sigma_T}$ and the corresponding p-values from the standard normal distribution.

Selecting Important Features Based on the null distribution of T_j, we can order the importance of each feature based on the testing statistics Z_j's and the p-values p_j's. The set of results of p_j's could be considered as the results for GWAS analysis with matrix-valued pheonotypes.

For example, based on the series of p-values p_j, we can calculate a series of the adjusted p-values $\tilde{p}_j$ using FWER or FDR control, and select the top SNPs with a certain threshold. However, a drawback of using traditional FWER of FDR control approach is that the correlation structures between the covariates are ignored, so that there might be power loss during the screening procedure.

To overcome this issue, knockoff filters were proposed **?**. The idea of knockoff filter is that we artificially construct a design matrix that is similar to the original design matrix but independent of the response. Based on the artificial design matrix, we are able to get a null distribution of the testing statistic with the correlation between the variables considered. Therefore we are able to get a data driven estimator of the FDR or FWER.

First of all, we need to construct the knockoff design matrix $\tilde{X}$ as proposed by **?**. Let $\Sigma = X'X$, we select $s > 0$, s.t., $diag(s) \preceq 2\Sigma$, the knockoff matrix could be constructed as

$$\tilde{X} = X(I - \Sigma^{-1} diag(s)) + \tilde{U}C \tag{4.2}$$

where $\tilde{U}$ is a $I \times p$ orthonormal matrix which is orthogonal to the span of X and C is the Cholesky decomposition of $2diag(s) - diag(s)\Sigma^{-1}diag(s)$. With the construction above, we can see that $\tilde{X}$ satisfies $\tilde{X}'\tilde{X} = \Sigma$ and $X'\tilde{X} = \Sigma - diag(s)$.

However, as we can see that the construction above is only valid when $p \leq 2I$ because we need to find a p-dimensional space orthogonal to the span of X. In imaging genetic studies, this assumption does not hold since we usually have $p > n$. To overcome this issue, we used a screening procedure first as proposed by **?**. We devided the samples into 2 blocks X_1 and X_2 with sample sizes I_1 and $I_2 = I - I_1$ respectively. On the first set of samples X_1 and Y_1, we calculate the marginal screening testing statistics Z_j^1, and pick the top d features with $d \leq 2I_2$. We can choose $d = 2I_2$ to minimize the probability of missing any important features in this step. Next, we consider only the d variables selected in step 1 in X_2 (indexed

S_1, denoted as $\hat{X}_2$. Then we can construct the corresponding knockoff features based on $\hat{X}_2$ as $\tilde{X}_2$.

After constructing the knockoff matrix $\tilde{X}_2$, we need to select important features based on the testing statistics. Now the augmented design matrix becomes $[\hat{X}_2\tilde{X}_2]$ and using marginal screening we can get a series of testing statistics (p-values) $W_j^1 = -\log \hat{p}_j^2$ and $W_j^2 = -\log \tilde{p}_j^2$. As we can see, W_j^2 is similar to a sample from the null distribution of W_j^1 if there is no correlation. Therefore, we can further define $\hat{W}_j = \max(W_j, W_j^*)\texttt{sign}(W_j > W_j^*)$. As we can see, if $\hat{W}_j < 0$, it is most likely that X_j is a false discovered feature since its significance is even below the null distribution. For any $t > 0$, if we select the features based on $\hat{W}_j > t$, a sample based estimator of FDR becomes

$$F\hat{D}R(t) = \frac{\#\{j \in S_1, \hat{W}_j \geq t\}}{\#\{j \in S_1, \hat{W}_j \leq -t\}} \tag{4.3}$$

We can select the minimal t such that $F\hat{D}R(t) \leq q$, where q is our targeted FDR rate.

4.3.2 Multivariate Regression

After selecting the important features $\hat{S}$ using a certain threshold, we then include all the important variables in one linear regression model. With our assumption of low rank structure of coefficient matrices, we can formulate the following penalized matrix regression problem.

$$f(B) = \frac{1}{2n}\sum_{i=1}^{I} ||Y_i - \sum_{j \in \hat{S}} x_{ij}B_j||_F^2 + \lambda \sum_{j \in \hat{S}} rank(B_j)$$

To optimize the object function above, we used the coordinate descend approach, where in each iteration, we update B_j for a certain $j \in \hat{S}$.

Given a current estimate of $\hat{B}_k$ for $k \neq j$, let $R_i^j = Y_i - \sum_{k \neq j} x_{ik}B_k$. The object function becomes

$$
\begin{aligned}
f(B_j) =& \frac{1}{2n}\sum_{i=1}^{I}||R_i^j - x_{ij}B_j||_F^2 + \lambda rank(B_j) \\
=& \frac{1}{2n}||B - \sum_{i=1}^{I} x_{ij}R_i^j||_F^2 + \lambda rank(B_j)
\end{aligned}
$$

The solution to the obtimization problem above has a closed form solution as

$$\hat{B} = UD(2n\lambda)V'$$

where UDV' is the SVD decomposition of $\sum_{i=1}^{I} x_{ij}R_i^j$, and $D(2n\lambda)$ is the thresholded diagonal matrix of D with entries smaller than $2n\lambda$ shrinked to 0.

4.4 Numerical Studies

In this section, we did several simulation studies to evaluate the performance of our proposed approaches.

4.4.1 Null Distribution of T_j

First of all, we used simulations to illustrate the accuracy of the proposed null distribution of testing statistic T_j.

We simulated the data with $I = 1,000$ subjects, $p = 2,000$ features, and with different imaging sizes $(m, n) = (30, 30), (100, 100), (30, 100), (100, 30)$. All the regression coefficents B_j's were set to be 0. The rank in the noise E_i is set to be 4, and the largest eigenvalue of B was set to be $2\sqrt{y}$. The QQ plot of the $2,000$ p-values of the features were presented in Figure 4.1. We can see that the derived distribution is a very good approximation to the true distribution of the testing statistics.

4.4.2 FDR Control using Knockoffs

In this section, we evaluated the appropriateness of using knockoff for FDR control, and compared the performance with other FDR control strategies including Benjamini–Hochberg procedure and Benjamini–Yekutieli procedure.

In the simulation, we set $I = 1,000, p = 2,000$, and the true non-zero coefficient matrices

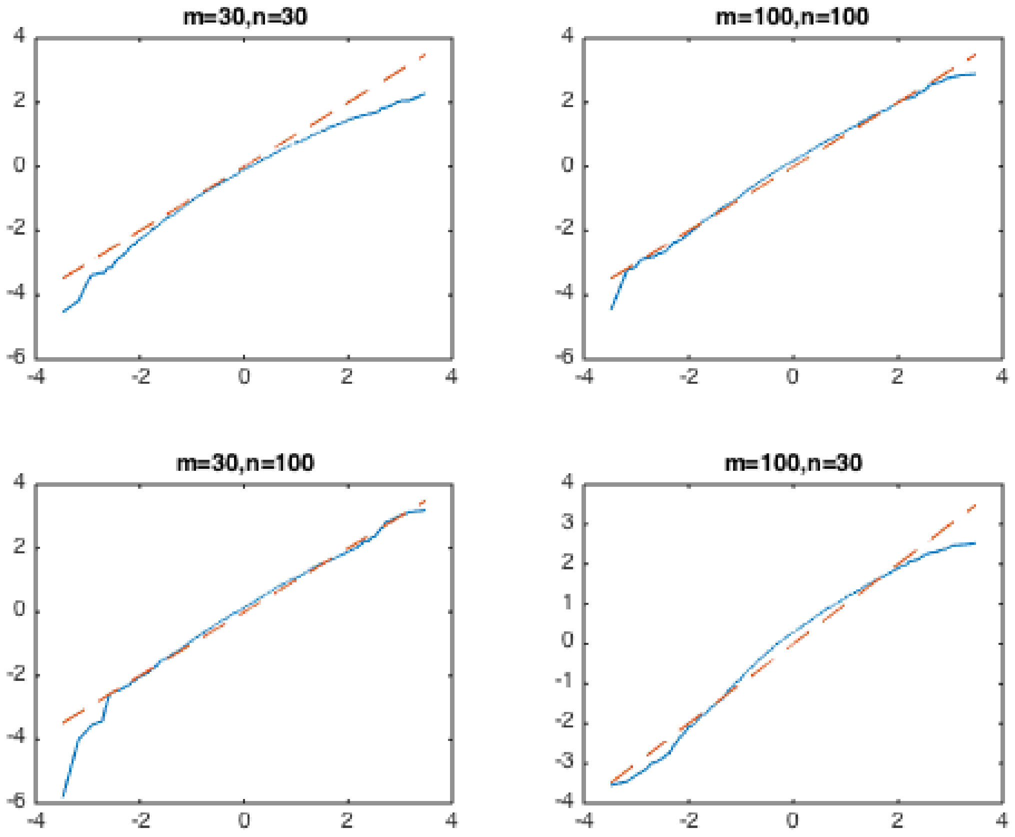

Figure 4.1: QQ Plot of Null Distribution of T_j

are set to be equal: a matrix with entries 1 in the center $t \times t$ voxels, and 0 elsewhere, where t ranges from 2 to 10, representing different signal strength. The actual false discovery rates with different settings are shown in Table 4.1. As we can see, the traditional approaches gives inflated FDR estimates, while knockoff filters gives more reasonable estimates of FDR. The potential reason for the inflated FDR of traditional approaches is that due to the co linearity of the covariates, there are spurious correlations between some of the covariates, which brings in false discover results. On the other hand, since knockoffs generate a set of covariates with the same correlation structures as the original covariates, the spurious correlations will also appear in the knockoff variables, thus limiting the probability of false discover due to spurious correlations.

Table 4.1: False Discovery Rate Control using Knockoff, Benjamini–Yekutieli, and Benjamini–Hochberg

	q=0.05			q=0.10			q=0.25		
Size of Signal	Knockoff	BH	BY	Knockoff	BH	BY	Knockoff	BH	BY
2	0.083	0.391	0.129	0.117	0.578	0.232	0.319	0.622	0.388
4	0.075	0.408	0.179	0.136	0.715	0.112	0.231	0.740	0.287
6	0.088	0.319	0.101	0.128	0.620	0.171	0.291	0.705	0.410
8	0.106	0.442	0.237	0.168	0.502	0.199	0.180	0.544	0.225
10	0.092	0.29	0.126	0.135	0.418	0.210	0.309	0.569	0.316

4.4.3 Multivariate Regression

The performance of the multivariate regression based on coordinate descend algorithm is studied.

We first examined the coordinate descend approach separately by including only the set of true covariates. We used the setting $I = 1,000, p = s = 20, m = 50, n = 70$. Figure 4.2 shows the true coefficient matrices and the estimated coefficient matrices. We compared our coordinate descend approach with several other approaches including L2RM Kong et al. (2020), voxel-wise regression, and voxel-wise regression with L_0 penalty. We can see coordinate descend approach gives similar results as L2RM approach, both of which can recover the true signal well. The voxel-wise approaches, however, includes some noise in the recovered images, which is mainly due to lack of consideration of the low-rank nature of the true signals.

Figure 4.3 shows the relative error of the predicted response. We can see again L2RM and coordinate descend algorithm outperforms voxel-wise regression approaches, which is mainly due the the flexibility of the voxel-wise models. In addition, the L_0 penalty based voxel-wise regression model performs worse than ordinary linear regression when all the coefficient matrices are the same, which is due to the fact that the sparse assumption is violated in this case.

Next, we also evaluated the overall performance of our proposed marginal screening regression models (MSRM) and compare it with other competitors. Specifically, we compared our approach with L2RM, L2RM with marginal screening (L2RM-MS), voxel-wise regression with L_0 penalty, and voxel-wise regression with L_1 penalty. The RMSE, computational time

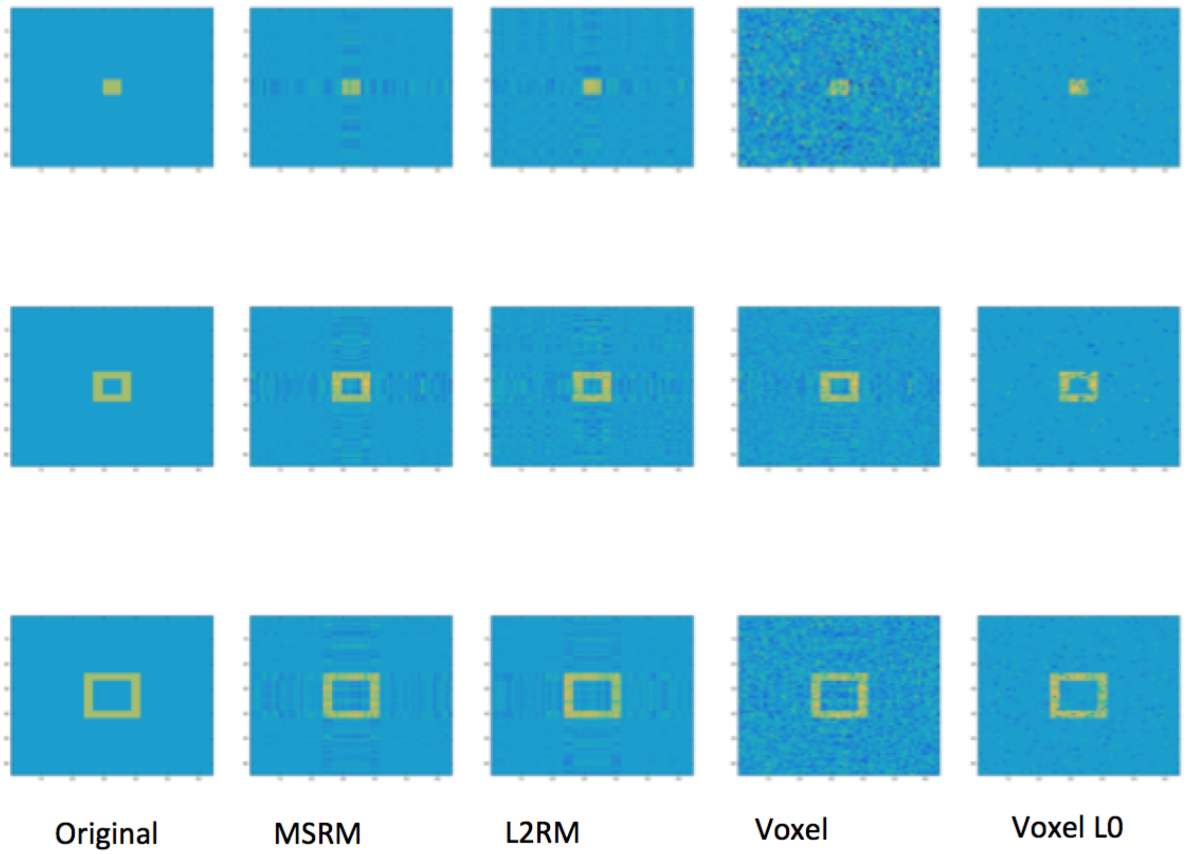

Figure 4.2: Recovered Coefficient Matrices

are listed in Table 4.2. We can see L2RM achieves the best RMSE in all cases, while L2RM and MSRM are comparable in terms of L2RM and slightly worse than L2RM. However, the computational time is much faster for theoretical marginal screening based approaches, because L2RM needs bootstrap test. voxel-wise approaches are fast in terms of computational time, but performs poorly in terms of RMSE.

Table 4.2: Comparison of Relative Error and Computational Speed of Different Methods for Matrix Regression

(m,n)	(30,30)		(30,100)		(100,30)		(100,100)	
Method	RE	Time (s)	RE	Time (s)	RE	Time (s)	RE	Time (s)
MSRM	0.3126	149.6	0.371	408.100	0.374	459.320	0.367	1214.775
L2RM	0.2855	2885.2	0.358	13764.7	0.343	17744.650	0.413	43382.110
L2RM-MS	0.3048	173.7	0.384	419.610	0.361	482.584	0.417	1122.140
Voxel L0	0.4109	65.1	0.485	217.176	0.506	193.602	1.377	629.037
Voxel L1	0.5498	102.3	0.725	309.382	0.857	302.237	1.514	691.181

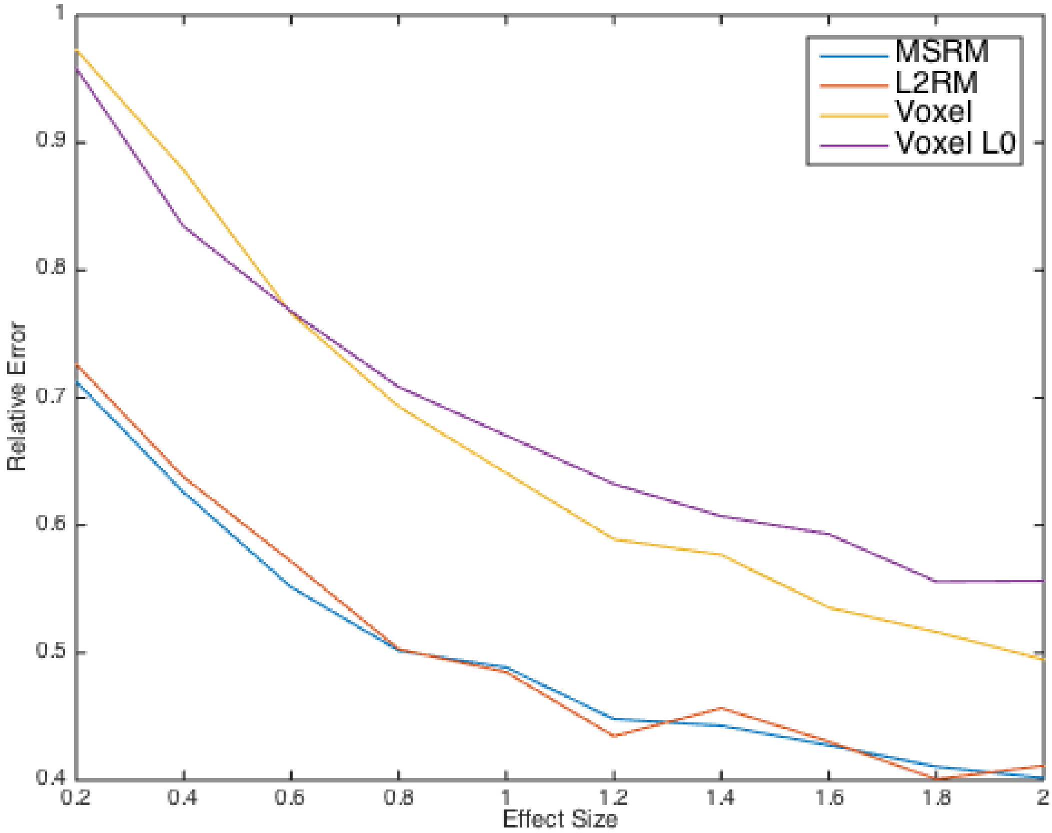

Figure 4.3: Relative Error of Recovered Response Images

4.5 Conclusion

In this paper, we proposed a matrix regression model with high-dimensional imaging outcome and genetic predictors. A marginal screening procedure with asymptotic null distribution was derived. A knockoff procedure was proposed along with the marginal screening to control for the FDR rate at the screening level. Furthermore, a matrix factorization approach was used for efficiently estimating the associations after screening. Numerical studies have shown that the proposed method outperforms existing approaches in terms of speed and screening accuracy. The derived asymptotic distribution is close to empirical distribution and the knockoff procedure controls the FDR rate better than other multiplicity approaches, although the FDR is still inflated, which is a future work for this project.

Moreover, numerical studies have shown that the proposed method is powerful at capturing complex features in high-dimensional associations, and thus providing a powerful tool for imaging genetic studies.

CHAPTER 5: PREDICTION WITH MODALITY IMPUTATION

5.1 Introduction

In imaging genetics studies, missing data is a fundamental problem that most of the researchers have to face. For example, in ADNI study, only half of the participants have complete imaging data (both resting-state fMRI and DTI). The common solution to missing data in imaging genetics is to exclude subjects with missing modality. However, excluding half of the subjects in a data set may lead to significantly reduced power for inference and accuracy for prediction. Therefore, efficient approaches for imputing the high-dimensional missing data is needed. The most straightforward way of imputing the missing data is through regression models, which may lack the power to handle high-dimensional cases. More advanced approaches have been proposed to replace simple linear regression. For example, ridge regression Hilt, Seegrist, Service., and Northeastern Forest Experiment Station (Radnor (Hilt et al.) and LASSO Tibshirani (1996) are alternatives to linear regression to handle high-dimensional problems. The two estimators shrinks the coefficient estimators to provide a more accurate prediction by a trade-off between variance and bias. Other machine learrning alternatives including random forest and SVM are also proposed to replace linear model for imputation. Recently, as the emergence of the powerful tool of deep learning, more and more deep learning techniques are also used for imputation. Among them, the most promising approach is generative adversarial networks (GAN). Two approaches Yoon et al. (2018); Dong et al. (2021) have shown great improvement of imputation accuracy.

However, the ultimate goal of imputation in imaging genetics is to better predict the phenotype of our interest. Therefore, how the imputation approaches could affect the prediction of phenotype remains unknown. In this paper, with a simple example of linear imputation, we first showed that imputation of missing covariates may lead to a biased prediction and worse accuracy. Thus we have to be careful when we decide to use imputation. Next, we compared the performance of multiple imputation approaches, and shown that in high-dimensional cases, the GAN based approach GAIN Dong et al. (2021) outperforms other imputation techniques.

5.2 Model Setup

Assume we have n subjects with responses Y and covariates $X \in \mathbb{R}^p$. Particularly, X has three blocks, $X_0 \in \mathbb{R}^{p_0}$, $X_1 \in \mathbb{R}^{p_1}$, and $X_2 \in \mathbb{R}^{p_2}$. Meanwhile, all the n subjects could be divided into 3 blocks, n_1 of them are missing X_1, n_2 of them are missing X_2, and n_3 of them have complete covariates. Therefore, the full data of the subjects are

$$X = \begin{bmatrix} X_{10} & \color{red}{X_{11}} & X_{12} \\ X_{20} & X_{21} & \color{red}{X_{22}} \\ X_{30} & X_{31} & X_{32} \end{bmatrix}$$

The missing mechanism is assumed to be missing completely at random (MCAR) for simplicity.

Further, we assume the covariance structure of X as

$$Cov(X) = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} & \Sigma_{02} \\ \Sigma_{10} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{20} & \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

The response Y is associated with X via the following lienar model:

$$Y = X\beta + \epsilon \tag{5.1}$$

where $\epsilon{\sim}N(0, \sigma^2 I_n)$ is the random error.

We compare the following 2 methods of estimating beta.

- **Imputation**. We use linear models to impute the missing data. The models for imputation are based on n_3 subjects with complete data. After imputing X_{11} and X_{22}, we get the imputed data matrix $\tilde{X}$, and estimate $\tilde{\beta}$ with Y and $\tilde{X}$.

- **Complete data**. Use n_3 subjects to estimate $\hat{\beta}$.

After some calculation, we can see that the distributions of the estimates are

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{n_3}\Sigma^{-1})$$

and

$$\tilde{\beta} \sim N(S^{-1}T\beta, \sigma^2 S^{-1})$$

where

$$S = \begin{bmatrix} S_{00} & S_{01} & S_{02} \\ S_{10} & S_{11} & S_{12} \\ S_{20} & S_{21} & S_{22} \end{bmatrix}$$

and

$$S_{00} = n\Sigma_{00}$$

$$S_{01} = S'_{10} = n_1 \begin{bmatrix} \Sigma_{00} & \Sigma_{02} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{02} \\ \Sigma_{20} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{01} \\ \Sigma_{21} \end{bmatrix} + (n_2 + n_3)\Sigma_{01}$$

$$S_{02} = S'_{20} = n_2 \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{02} \\ \Sigma_{12} \end{bmatrix} + (n_1 + n_3)\Sigma_{02}$$

$$S_{11} = n_1 \begin{bmatrix} \Sigma_{10} & \Sigma_{12} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{02} \\ \Sigma_{20} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{01} \\ \Sigma_{21} \end{bmatrix} + (n_2 + n_3)\Sigma_{11}$$

$$S_{22} = n_2 \begin{bmatrix} \Sigma_{20} & \Sigma_{21} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{02} \\ \Sigma_{12} \end{bmatrix} + (n_1 + n_3)\Sigma_{22}$$

$$S_{12} = S'_{21} = n_1 \begin{bmatrix} \Sigma_{10} & \Sigma_{12} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{02} \\ \Sigma_{20} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{02} \\ \Sigma_{22} \end{bmatrix}$$

$$+n_2 \begin{bmatrix} \Sigma_{10} & \Sigma_{11} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{02} \\ \Sigma_{12} \end{bmatrix} + n_3\Sigma_{12}$$

From the distribution of $\tilde{\beta}$, we can see it is a biased estimator of β, and thus $X\tilde{\beta}$ is also a biased estimator of the true response $X\beta$.

Besides, we have,

$$T = \begin{bmatrix} T_{00} & T_{01} & T_{02} \\ T_{10} & T_{11} & T_{12} \\ T_{20} & T_{21} & T_{22} \end{bmatrix}$$

and

$$T_{00} = S_{00}, T_{01} = S_{01}, T_{02} = S_{02}, T_{10} = S_{10}, T_{20} = S_{20}, T_{11} = S_{11}, T_{22} = S_{22}$$

$$T_{12} = n_1 \begin{bmatrix} \Sigma_{10} & \Sigma_{12} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{02} \\ \Sigma_{20} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{02} \\ \Sigma_{22} \end{bmatrix} + (n_2 + n_3)\Sigma_{12}$$

$$T_{21} = n_2 \begin{bmatrix} \Sigma_{20} & \Sigma_{21} \end{bmatrix} \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{01} \\ \Sigma_{11} \end{bmatrix} + (n_1 + n_3)\Sigma_{21}$$

As a result, for a new data with the same distribution of X, i.e.,

$$X \sim N(0, \Sigma)$$

The MSE of estimating $X\beta$ with $X\hat{\beta}$ and $X\tilde{\beta}$ are

$$
\begin{aligned}
E||X\tilde{\beta} - X\beta||_2^2 =& E(\tilde{\beta} - \beta)'(X'X)(\tilde{\beta} - \beta) \\
=& Etr[(X'X)(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] \\
=& tr(\Sigma[\sigma^2 S^{-1} + (S^{-1}T - I)\beta\beta'(T'S^{-1} - I)]) \\
=& \sigma^2 tr(\Sigma S^{-1}) + \beta'(T'S^{-1} - I)\Sigma(S^{-1}T - I)\beta
\end{aligned}
$$

and

$$E||X\hat{\beta} - X\beta||_2^2 = \frac{p}{n_3}\sigma^2$$

To further compare the two MSEs, we assume $S = n\Sigma + A$ and $T = n\Sigma + B$. Under the assumption $n1 << n$ and $n_2 << n$, we have

$$E||X\tilde{\beta} - X\beta||_2^2 = \frac{p}{n}\sigma^2 - \frac{1}{n^2}\sigma^2 tr(A\Sigma^{-1}) + \frac{1}{n^2}\beta'(B' - A)\Sigma^{-1}(B - A)\beta + o(n^{-2})$$

and

$$E||X\tilde{\beta} - X\beta||_2^2 = \frac{p}{n}\sigma^2 + \frac{1}{n^2}p(n_1 + n_2)\sigma^2 + o(n^{-2})$$

Comparing the MSE of $\hat{\beta}$ and β, we can conclude that imputation performs better since $\frac{n+n_1+n_2}{n^2} \leq \frac{1}{n_3}$. However, when n_3 is close to n, the improvement may not be significant, and could be negative due to the contribution of the residuals.

5.3 TWAS

The problem of TWAS is formulated as following. There are three groups of subjects, the reference panal, training data, and testing data. We have SNP data X, gene expression data Z, and phenotype Y. In training and testing data, we only observed X_1 and Y_1, and X_2 and Y_2, respectively. In reference panel, we observe X_3 and Z_3.

We assume that

$$Y = X\alpha + Z\beta + \epsilon$$

and

$$Z = X\Lambda + E$$

In training data, we only have X_1 and Y_1, so we have to impute Z_1 if we want to model it. Assume we have obtained a linear estimator of Z based on X, i.e.,

$$\hat{Z} = X\hat{\Lambda}$$

then in the training data, we have the model

$$Y_1 = X_1\beta + \hat{Z}_1\alpha + e^* = X_1 D\gamma + e^*$$

where $D = [I\hat{\Lambda}]$.

This TWAS model is essentially a special case of the model that we described above where $n_2 = 0$ and $p_3 = 0$.

5.4 Conclusion

Imputation is a effective technique for genetic imaging studies to improve the sample size for inference and thus increase statistical power and learning accuracy. We confirmed that imputation of missing data always improve estimation acuracy, especially when the number of subjects with missing data is large.

APPENDIX A: TECHNICAL DETAILS OF CHAPTER 3

Main proofs

The idea of the proof for theorem 1-3 is based on the iteration from space $B^{(k)}$ to space $B^{(k+1)}$. From Algorithm 1, we know $B^{(k+1)}$ is the eigen-space of matrix $A^{(k)} = \sum S_i^{(k)} B^{(k)} B^{(k)\prime} S_i^{(k)}$. Let $a_k = ||\mathbf{Sin}\Theta(B, B^{(k)}||$, we are interested in finding the relationship between a_k and a_{k+1}.

First of all, let $\Sigma_i^0 = B\Lambda_i B'$, we know that the eigen-space of matrix

$$A^0 = \sum \Sigma_i^0 BB' \Sigma_i^0 = \sum B\Lambda_i^2 B'$$

is B. Therefore, we can write $A^{(k)}$ as a perturbation of matrix A^0 as

$$\begin{aligned} A^{(k)} =& A^0 + (A^{(k)} - A^0) \\ =& B(\sum \Lambda_i^2)B' + \sum S_i^{(k)} B^{(k)} B^{(k)\prime} S_i^{(k)} - \sum \Sigma_i^0 BB' \Sigma_i^0 \\ =& X + Z \end{aligned}$$

According to Cai and Zhang (2018), the perturbation bound of $||\mathbf{Sin}\Theta(B, B^{(k)}||$ is bounded by the following quantity

$$||\mathbf{Sin}\Theta(B, B^{(k+1)}|| \leq \frac{(\alpha+\beta)z_{12}}{\alpha^2 - \beta^2 - z_{12}^2}$$

where

$$\alpha = \sigma_{\min}(B'(X+Z)B)$$

$$\beta = ||B_\perp' Z B_\perp||$$

$$z_{12} = ||B' Z B_\perp||$$

For α, it is easy to obtain

$$\alpha = \sigma_{\min}(B'XB) = \sigma_{\min}(\sum \Lambda_i^2)$$

Consider β and z_{12}, we know

$$\begin{aligned} Z =& \sum S_i^{(k)} B^{(k)} B^{(k)\prime} S_i^{(k)} - \sum \Sigma_i^0 BB' \Sigma_i^0 \\ =& \sum (S_i^{(k)} - \Sigma_i^0)(B^{(k)} B^{(k)\prime} - BB')(S_i^{(k)} - \Sigma_i^0) \\ &+ \sum (S_i^{(k)} - \Sigma_i^0) BB' (S_i^{(k)} - \Sigma_i^0) \\ &+ \sum (S_i^{(k)} - \Sigma_i^0) BB' \Sigma_i^0 \end{aligned}$$

Lemma 4. Under sub-Gaussian assumption, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$, $||(S_i^{(k)} - \Sigma_i^0)|| \leq \tau[\sigma(\Lambda_i) + \sigma^2]$ for all $i = 1, ldots, I$, where $\tau = \frac{56}{c}[\frac{p}{n} \log \frac{2I}{\delta} + \frac{r}{p}]$.

Lemma 4 gives a uniform bound of sample covariance matrix estimation. It could be proved using ϵ-net with $\epsilon = 3/7$.

Now we use lemma 4 to get a upper bound for β and z_{12}. First let's look into β.

$$\begin{aligned} \beta =& ||B_\perp' Z B_\perp|| \\ \leq& \sum ||B_\perp' (S_i^{(k)} - \Sigma_i^0)(B^{(k)} B^{(k)\prime} - BB')(S_i^{(k)} - \Sigma_i^0) B_\perp|| \\ &+ \sum ||B_\perp' (S_i^{(k)} - \Sigma_i^0) BB' (S_i^{(k)} - \Sigma_i^0) B_\perp|| \\ &+ \sum ||B_\perp' (S_i^{(k)} - \Sigma_i^0) BB' \Sigma_i^0 B_\perp|| \\ \leq& \sum 2||\mathbf{Sin}\Theta(B, B^{(k)}|| \cdot ||S_i^{(k)} - \Sigma_i^0||^2 \\ &+ \sum ||S_i^{(k)} - \Sigma_i^0||^2 \\ \leq& \tau^2 \sum (\sigma(\Lambda_i) + \sigma^2)^2 + 2\tau^2 a_k \sum (\sigma(\Lambda_i) + \sigma^2)^2 \end{aligned}$$

Similarly we can get the same upper bound for z_{12},

$$z_{12} \leq \tau^2 \sum (\sigma(\Lambda_i) + \sigma^2)^2 + 2\tau^2 a_k \sum (\sigma(\Lambda_i) + \sigma^2)^2$$

For α, we have

$$\begin{aligned}\alpha =&\sigma_{\min}(B'(X+Z)B)\\ \geq&\sigma_{\min}(B'XB) - ||B'ZB||\\ \geq&\sigma_{\min}(\sum \Lambda_i^2) - \tau \sum (\sigma(\sigma(\Lambda_i) + \sigma^2)\\ &- \tau^2 \sum(\sigma(\Lambda_i) + \sigma^2)^2 - 2\tau^2 a_k \sum(\sigma(\Lambda_i) + \sigma^2)^2\end{aligned}$$

Let

$$A_0 = \sigma_{\min}(\sum \Lambda_i^2)$$

$$A_1 = \tau \sum (\sigma(\sigma(\Lambda_i) + \sigma^2)$$

$$A_2 = \tau^2 \sum(\sigma(\Lambda_i) + \sigma^2)^2$$

we have

$$\alpha \geq A_0 - A_1 - A_2 - 2A_2 a_k$$

$$z_{12} \leq A_2 + 2A_2 a_k$$

$$\beta \leq A_2 + 2A_2 a_k$$

Under the condition that $\frac{A_1}{A_0} \leq k_1$ and $\frac{A_2}{A_0} \leq k_2$, we have

$$\begin{aligned}a_{k+1} =&||\mathbf{Sin}\Theta(B, B^{(k+1)}|| \leq \frac{(\alpha+\beta)z_{12}}{\alpha^2 - \beta^2 - z_{12}^2}\\ \leq&\frac{(A_0 - A_1)(A_2 + 2A_2 a_k)}{(A_0 - A_1 - A_2 - 2A_2 a_k)^2 - 2(A_2 + 2A_2 a_k)^2}\\ \leq&\frac{(1-k_1)(k_2 + 2k_2 a_k)}{(1-k_1-3k_2)^2 - 18k_2^2}\\ =&\frac{(1-k_1)k_2}{(1-k_1-3k_2)^2 - 18k_2^2} + \frac{2(1-k_1)k_2}{(1-k_1-3k_2)^2 - 18k_2^2} a_k\end{aligned}$$

Therefore, let $\rho = \frac{2(1-k_1)k_2}{(1-k_1-3k_2)^2 - 18k_2^2}$, we have

$$a_{k+1} \leq \frac{1}{2}\rho + \rho a_k$$

Under the condition $\rho < 1$, we have the convergence of the above series at a exponential rate of ρ. Furthermore, with $k \to \infty$, we have

$$\lim_{k\to\infty} a_k \leq \frac{\rho}{2(1-\rho)}$$

This finishes the proof of Theorem 1 and 2.

The proof of theorem 3 could be based on the results of theorem 2. Note that $\Sigma_i = B\Lambda_i B'$ and $\hat{\Sigma}_i = B^{(k)}B^{(k)\prime}S_i^{(k)}B^{(k)}B^{(k)\prime}$, we have

$$\begin{aligned}
||\hat{\Sigma}_i - \Sigma_i||_F =& ||B^{(k)}B^{(k)\prime}S_i^{(k)}B^{(k)}B^{(k)\prime} - B\Lambda_i B'||_F \\
=& ||B^{(k)}B^{(k)\prime}S_i^{(k)}B^{(k)}B^{(k)\prime} - B^{(k)}B^{(k)\prime}\Sigma_i B^{(k)}B^{(k)\prime} + B^{(k)}B^{(k)\prime}\Sigma_i B^{(k)}B^{(k)\prime} \\
& - BB'\Sigma_i B^{(k)}B^{(k)\prime} + BB'\Sigma_i B^{(k)}B^{(k)\prime} - B\Lambda_i B'||_F \\
\leq& ||B^{(k)}B^{(k)\prime}S_i^{(k)}B^{(k)}B^{(k)\prime}||_F + ||B^{(k)}B^{(k)\prime}\Sigma_i B^{(k)}B^{(k)\prime} - BB'\Sigma_i B^{(k)}B^{(k)\prime}||_F \\
& + ||BB'\Sigma_i B^{(k)}B^{(k)\prime} - B\Lambda_i B'||_F \\
\leq& \tau[\sigma(\Lambda_i) + \sigma^2]\sqrt{r} + 2a_k||\Lambda_i||_F + 2a_k||\Lambda_i||_F \\
\leq& \frac{2\rho}{(1-\rho)}||\Lambda_i||_F + \tau\sqrt{r}[\sigma(\Lambda_i) + \sigma^2]
\end{aligned}$$

as $k \to \infty$.

REFERENCES

Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **26,** 63–72.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., Miller, K. L., and Smith, S. M. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *NeuroImage* **166,** 400–424.

Atasoy, S., Donnelly, I., and Pearson, J. (2016). Human brain networks function in connectome-specific harmonic waves. *Nature Communications* **7,** 10340.

Bassett, D. S., Nelson, B. G., Mueller, B. A., Camchong, J., and Lim, K. O. (2012). Altered resting state complexity in schizophrenia. *NeuroImage* **59,** 2196–2207.

Bullmore, E. T. and Bassett, D. S. (2011). Brain graphs: Graphical models of the human brain connectome. *Annual Review of Clinical Psychology* **7,** 113–140.

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics* **46,** 60–89.

Candès, E. J. and Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. *CoRR* **abs/0903.1476,**.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95,** 759–771.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., and Initiative, A. D. N. (2009). Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. *Hippocampus* **19,** 579–587.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis* **7,** 1–46.

Dong, W., Fong, D. Y. T., Yoon, J.-s., Wan, E. Y. F., Bedford, L. E., Tang, E. H. M., and Lam, C. L. K. (2021). Generative adversarial networks for imputing missing data for big data clinical research. *BMC Medical Research Methodology* **21,** 78.

Durante, D., Dunson, D. B., and Vogelstein, J. T. (2017). Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association* **112,** 1516–1530.

Enciso-Olivera, C. O., Ordóñez-Rubiano, E. G., Casanova-Libreros, R., Rivera, D., Zarate-Ardila, C. J., Rudas, J., Pulido, C., Gómez, F., Martínez, D., Guerrero, N., Hurtado, M. A., Aguilera-Bustos, N., Hernández-Torres, C. P., Hernandez, J., and Marín-Muñoz, J. H. (2021). Structural and functional connectivity of the ascending arousal network for prediction of outcome in patients with acute disorders of consciousness. *Scientific Reports* **11,** 22952.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147,** 186–197.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39,** 3320–3356.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 849–911.

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., and Raichle, M. E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 10046–10051.

Fox, M. D. and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience* **8,** 700–711.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Consortium, G., Nicolae, D. L., Cox, N. J., and Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47,** 1091–1098. 26258848[pmid].

Gao, W., Gilmore, J. H., Giovanello, K. S., Smith, J. K., Shen, D., Zhu, H., and Lin, W. (2011). Temporal and spatial evolution of brain network topology during the first two years of life. *PLOS ONE* **6,** e25278–.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P., and Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48,** 245–252. 26854917[pmid].

He, B., Yang, L., Wilke, C. T., and Yuan, H. (2011). Electrophysiological imaging of brain activity and connectivity—challenges and opportunities. *IEEE Transactions on Biomedical Engineering* **58,** 1918–1931.

Hilt, D. E., Seegrist, D. W., Service., U. S. F., and Northeastern Forest Experiment Station (Radnor, P. *Ridge, a computer program for calculating ridge regression estimates*, volume no.236. Upper Darby, Pa, Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station, 1977. https://www.biodiversitylibrary.org/bibliography/68934.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps.

Huang, J., Jiao, Y., Liu, Y., and Lu, X. (2018). A constructive approach to l_0 penalized regression. *Journal of Machine Learning Research* **19,** 1–37.

Jin, J., Ke, Z. T., and Luo, S. (2017). Estimating network memberships by simplex vertex hunting.

Joenssen, D. W. and Bankhofer, U. (2012). Hot deck methods for imputing missing data. In Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 63–75, Berlin, Heidelberg. Springer Berlin Heidelberg.

Johnson, K. A., Fox, N. C., Sperling, R. A., and Klunk, W. E. (2012). Brain imaging in alzheimer disease. *Cold Spring Harbor perspectives in medicine* **2,** a006213–a006213.

Josse, J. and Husson, F. (2012). Selecting the number of components in pca using cross-validation approximations. *Computational Statistics Data Analysis* **56,**.

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review / Revue Internationale de Statistique* **51,** 175–188.

Ke, Z. T. (2016). A geometrical approach to topic model estimation.

Ke, Z. T. and Wang, M. (2017). A new svd approach to optimal topic estimation.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.

Kong, D., An, B., Zhang, J., and Zhu, H. (2020). L2rm: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association* **115,** 403–424.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature* **412,** 150–157.

Mao, X., Sarkar, P., and Chakrabarti, D. (2018). Overlapping clustering models, and one (class) svm to bind them all.

Medland, S. E., Duffy, D. L., Wright, M. J., Geffen, G. M., Hay, D. A., Levy, F., van Beijsterveldt, C. E. M., Willemsen, G., Townsend, G. C., White, V., Hewitt, A. W., Mackey, D. A., Bailey, J. M., Slutske, W. S., Nyholt, D. R., Treloar, S. A., Martin, N. G., and Boomsma, D. I. (2009). Genetic influences on handedness: Data from 25,732 australian and dutch twin families. *Neuropsychologia* **47,** 330–337.

Nurhayu, W., Nila, S., Widayati, K. A., Rianti, P., Suryobroto, B., and Raymond, M. (2020). Handedness heritability in industrialized and nonindustrialized societies. *Heredity* **124,** 313–324.

O'Donnell, L. J. and Westin, C.-F. (2011). An introduction to diffusion tensor image analysis. *Neurosurgery clinics of North America* **22,** 185–viii.

Patz, J. A., Campbell-Lendrum, D., Holloway, T., and Foley, J. A. (2005). Impact of regional climate change on human health. *Nature* **438,** 310–317.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C R, J., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., and Weiner,

M. W. (2010). Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology* **74,** 201–209.

Rogers, B. P., Morgan, V. L., Newton, A. T., and Gore, J. C. (2007). Assessing functional connectivity in the human brain by fmri. *Magnetic resonance imaging* **25,** 1347–1357.

Sainburg, R. L. (2014). Convergent models of handedness and brain lateralization. *Frontiers in psychology* **5,** 1092–1092.

Scruggs, J. T. and Glabadanidis, P. (2003). Risk premia and the dynamic covariance between stock and bond returns. *Journal of Financial and Quantitative Analysis* **38,** 295–316.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22,** 231–245.

Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L., and Beckmann, C. F. (2014). Group-pca for very large fmri datasets. *NeuroImage* **101,** 738–749.

Stekhoven, D. J. and Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118.

Sun, W. W. and Li, L. (2016). Store: Sparse tensor response regression and neuroimaging analysis.

Teeuw, J., Brouwer, R. M., Guimarães, J. P. O. F. T., Brandner, P., Koenis, M. M. G., Swagerman, S. C., Verwoert, M., Boomsma, D. I., and Hulshoff Pol, H. E. (2019). Genetic and environmental influences on functional connectivity within and between canonical cortical resting-state networks throughout adolescent development in boys and girls. *NeuroImage* **202,** 116073.

Teipel, S. J., Wohlert, A., Metzger, C., Grimmer, T., Sorg, C., Ewers, M., Meisenzahl, E., Klöppel, S., Borchardt, V., Grothe, M. J., Walter, M., and Dyrba, M. (2017). Multicenter stability of resting state fmri in the detection of alzheimer's disease and amnestic mci. *NeuroImage. Clinical* **14,** 183–194.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288.

Tozakidou, M., Wenz, H., Reinhardt, J., Nennig, E., Riffel, K., Blatow, M., and Stippich, C. (2013). Primary motor cortex activation and lateralization in patients with tumors of the central region. *NeuroImage. Clinical* **2,** 221–228.

van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T. T., and Schmidt, R. (2017). Proportional thresholding in resting-state fmri functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations. *NeuroImage* **152,** 437–449.

van den Heuvel, M. P. and Pol, H. E. H. (2010). Exploring the brain network: A review on resting-state fmri functional connectivity. *European Neuropsychopharmacology* **20,** 519 – 534.

van Ettinger-Veenstra, H. M., Ragnehed, M., Hällgren, M., Karlsson, T., Landtblom, A. M., Lundberg, P., and Engström, M. (2010). Right-hemispheric brain activation correlates to language performance. *NeuroImage* **49,** 3481–3488.

Vincent, A., Bien, C. G., Irani, S. R., and Waters, P. (2011). Autoantibodies associated with diseases of the cns: new developments and future challenges. *The Lancet Neurology* **10,** 759 – 772.

Wang, H., Banerjee, A., and Boley, D. (2011). Common component analysis for multiple covariance matrices. pages 956–964.

Wang, S., Arroyo, J., Vogelstein, J. T., and Priebe, C. E. (2019). Joint embedding of graphs.

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* **12,** 99–111.

Wee, C.-Y., Yap, P.-T., Zhang, D., Wang, L., and Shen, D. (2014). Group-constrained sparse fmri connectivity modeling for mild cognitive impairment identification. *Brain structure & function* **219,** 641–656.

Xia, C. H., Ma, Z., Cui, Z., Bzdok, D., Bassett, D. S., Satterthwaite, T. D., Shinohara, R. T., and Witten, D. M. (2019). Multi-scale network regression for brain-phenotype associations. *bioRxiv* .

Yang, B., Janssens, D., Ruan, D., Cools, M., Bellemans, T., and Wets, G. (2012). A data imputation method with support vector machines for activity-based transportation models. In Wang, Y. and Li, T., editors, *Foundations of Intelligent Systems*, pages 249–257, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88,** 76–82.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). GAIN: missing data imputation using generative adversarial nets. *CoRR* **abs/1806.02920,**.

Yu, R., Zhang, H., An, L., Chen, X., Wei, Z., and Shen, D. (2017). Connectivity strength-weighted sparse group representation-based brain network construction for mci classification. *Human brain mapping* **38,** 2370–2383.

Yuan, M. (2016). Degrees of freedom in low rank matrix estimation. *Science China Mathematics* **59,** 2485–2502.

Zhang, J., Sun, W. W., and Li, L. (2019). Mixed-effect time-varying network model and application in brain connectivity analysis. *Journal of the American Statistical Association* page 1–15.

Zhang, T., Liao, Q., Zhang, D., Zhang, C., Yan, J., Ngetich, R., Zhang, J., Jin, Z., and Li, L. (2021). Predicting mci to ad conversation using integrated smri and rs-fmri: Machine learning and graph theory approach. *Frontiers in Aging Neuroscience* **13,**.

Zhang, Y., Zhang, H., Chen, X., Liu, M., Zhu, X., Lee, S.-W., and Shen, D. (2019). Strength and similarity guided group-level brain functional network construction for mci diagnosis. *Pattern Recognition* **88,** 421–430.

Zhou, Z., Chen, X., Zhang, Y., Qiao, L., Yu, R., Pan, G., Zhang, H., and Shen, D. (2019). Brain network construction and classification toolbox (brainnetclass).