2022

Extracting Generalizable Hierarchical Patterns Of Functional Connectivity In The Brain

Dushyant Sahoo
University of Pennsylvania

Extracting Generalizable Hierarchical Patterns Of Functional Connectivity In The Brain

Abstract

The study of the functional organization of the human brain using resting-state functional MRI (rsfMRI) has been of significant interest in cognitive neuroscience for over two decades. The functional organization is characterized by patterns that are believed to be hierarchical in nature. From a clinical context, studying these patterns has become important for understanding various disorders such as Major Depressive Disorder, Autism, Schizophrenia, etc. However, extraction of these interpretable patterns might face challenges in multi-site rsfMRI studies due to variability introduced due to confounding variability introduced by different sites and scanners. This can reduce the predictive power and reproducibility of the patterns, affecting the confidence in using these patterns as biomarkers for assessing and predicting disease. In this thesis, we focus on the problem of robustly extracting hierarchical patterns that can be used as biomarkers for diseases.

We propose a matrix factorization based method to extract interpretable hierarchical decomposition of the rsfRMI data. We couple the method with adversarial learning to improve inter-site robustness in multi-site studies, removing non-biological variability that can result in less interpretable and discriminative biomarkers. Finally, a generative-discriminative model is built on top of the proposed framework to extract robust patterns/biomarkers characterizing Major Depressive Disorder.

Results on large multi-site rsfMRI studies show the effectiveness of our method in uncovering reproducible connectivity patterns across individuals with high predictive power while maintaining clinical interpretability. Our framework robustly identifies brain patterns characterizing MDD and provides an understanding of the manifestation of the disorder from a functional networks perspective which can be crucial for effective diagnosis, treatment and prevention. The results demonstrate the method's utility and facilitate a broader understanding of the human brain from a functional perspective.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Electrical & Systems Engineering

First Advisor

Christos Davatzikos

Subject Categories

Engineering | Neuroscience and Neurobiology | Statistics and Probability

EXTRACTING GENERALIZABLE HIERARCHICAL PATTERNS OF FUNCTIONAL CONNECTIVITY IN THE BRAIN

Dushyant Sahoo

A DISSERTATION

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Christos Davatzikos, Professor of Radiology and Electrical and Systems Engineering

Graduate Group Chairperson

Alejandro Ribeiro, Professor of Electrical and Systems Engineering

Dissertation Committee

Pratik Chaudhari, Assistant Professor of Electrical and Systems Engineering
Ted Satterthwaite, Associate Professor of Psychiatry

ACKNOWLEDGEMENT

Looking back at the past five years, I realize that my Phd has been a very enriching journey. Not only did I learn to become an independent researcher, but I also grew in other aspects of life to become a more mature person. This thesis is a manifestation of that journey, and its completion would not have been possible without the support from a large group of people.

First and foremost, I would like to warmly thank and express my highest appreciation to my adviser, Dr. Christos Davatzikos, for his patience and guidance during the past few years. He provided me with much freedom to explore issues and directions that interest me and always offered extremely useful and thoughtful feedback during the exploration. His profound research insights and broad vision have been foundational in shaping my identity as an independent researcher.

I express my heartfelt gratitude to my dissertation committee members, Dr. Pratik Chaudhari and Dr. Theodore Satterthwaite for giving valuable feedback on my thesis. I would not be able to overcome all the challenges along the journey without the support from my collaborators and all the people in CBICA. I would like to especially thank Dr. Haochang Shou, Dr. Cynthia Fu, Dr. Yong Fan, Dhivya, Dr. Junhao Wen, Dr. Mathilde Antoniades, Dr. Ahmed Abdulkadir and Dr. Guray Erus for their help in research and outside.

I am beyond blessed to have so many amazing friends like family. I want to thank my friends– Aalok, Abha, Arpit, Ishaan, Prathamesh, Rohan, Roopal, Sharath, Soham, and many more, who have always been by my side in tough times.

Finally and most importantly, I would like to thank my parents Prem and Raj Sahoo and my brother Abhimanyu for all the sacrifices they have made for this to be possible. Life is full of ups and downs, but they are always my most vital support and my greatest motivation. The unconditional love that they provided was always the largest comfort for me. Without them, I will not be the person I am today. This thesis is dedicated to them!

ABSTRACT

EXTRACTING GENERALIZABLE HIERARCHICAL PATTERNS OF FUNCTIONAL CONNECTIVITY IN THE BRAIN

Dushyant Sahoo

Christos Davatzikos

The study of the functional organization of the human brain using resting-state functional MRI (rsfMRI) has been of significant interest in cognitive neuroscience for over two decades. The functional organization is characterized by patterns that are believed to be hierarchical in nature. From a clinical context, studying these patterns has become important for understanding various disorders such as Major Depressive Disorder, Autism, Schizophrenia, etc. However, extraction of these interpretable patterns might face challenges in multi-site rsfMRI studies due to variability introduced due to confounding variability introduced by different sites and scanners. This can reduce the predictive power and reproducibility of the patterns, affecting the confidence in using these patterns as biomarkers for assessing and predicting disease. In this thesis, we focus on the problem of robustly extracting hierarchical patterns that can be used as biomarkers for diseases.

We propose a matrix factorization based method to extract interpretable hierarchical decomposition of the rsfRMI data. We couple the method with adversarial learning to improve inter-site robustness in multi-site studies, removing non-biological variability that can result in less interpretable and discriminative biomarkers. Finally, a generative-discriminative model is built on top of the proposed framework to extract robust patterns/biomarkers characterizing Major Depressive Disorder.

Results on large multi-site rsfMRI studies show the effectiveness of our method in uncovering reproducible connectivity patterns across individuals with high predictive power while maintaining clinical interpretability. Our framework robustly identifies brain patterns characterizing MDD and provides an understanding of the manifestation of the disorder from a functional networks perspective which can be crucial for effective diagnosis, treatment and prevention. The results facilitate a broader understanding of the human brain from a functional perspective.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction

1.1 Overview

The human brain is a complex structure that integrates and coordinates information in the human body. It plays a central role in decision making by codifying instructions that control the rest of the body. Anatomically, the brain is made up of a network of cell bodies that process information and axons that relay the information. Functionally, it can be described as a network of regional hubs working together to perform various tasks. Recent technological advancements such as Magnetic Resonance Imaging, Electroencephalogram, etc., have allowed us to study networks to understand the organization of the human brain, revealing communications in the human brain at micro-scale, meso-scale, and macro-scale (Rauschecker and Scott, 2009; Gilbert and Li, 2013; Hirabayashi et al., 2013; Van Kerkoerle et al., 2014). These studies have provided several insights into the functioning of the healthy, aging and diseased human brain. In this thesis, we look at the patterns measured using resting-state functional Magnetic Resonance Imaging (rsfMRI).

Resting-state functional Magnetic Resonance Imaging (rsfMRI) (Biswal et al., 1995; Raichle et al., 2001) is a type of functional Magnetic Resonance Imaging (fMRI) to measure the brain activity of different regions by detecting changes in the blood flow that occurs in the resting state. rsfMRI can be used to measure the amount of co-activation or functional connectivity across different brain regions at rest. Here the underlying assumption is that two brain regions which reliably co-activate are more likely to participate in similar neural processes as opposed to two uncorrelated regions (Fox and Raichle, 2007; Biswal et al., 1995; Buckner et al., 2008). Functional connectivity is commonly estimated by computing the Pearson correlations between the time series recorded at different brain locations. Functional connectivity measure is helpful to explore the organization of the human brain and is used to extract functional networks, a set of distributed brain areas that show synchronous activities (Damoiseaux et al., 2006; Horovitz et al., 2008; Smith et al., 2009). It has been suggested that the networks extracted from fMRI data exhibit hierarchical structure (Guye et al., 2010; Sporns and Betzel, 2016). “This might provide evolutionary and adaptive advantages as different hubs can respond to the evolutionary or environmental pressure without jeopardizing the functioning of the entire brain” (Simon, 1991). Moreover, changes in the representation of functional networks extracted from rsfMRI have been observed in groups suffering from brain diseases and with disorders such as Spilepsy (Rajpoot et al., 2015; Riaz et al., 2013), schizophrenia (Kumari et al., 2009; Koch et al., 2015), Attention Deficit Hyperactivity Disorder (ADHD) (Wang et al., 2013; Riaz et al., 2017), Alzheimer’s disease (Wee et al., 2012), Parkinson’s disease (Díez-Cirarda et al., 2018; Wu et al., 2009) and Major Depressive Disorder (MDD) (Dansereau et al., 2017; Xia et al., 2019), thus can provide with potential biomarkers of illness.

Recently, with the introduction of several open-access neuroimaging data-sharing initiatives providing large samples and high-quality data acquired from multiple sites, the changes in the functional patterns can be investigated in large and varied populations (Alexander et al., 2017; Biswal et al., 2010; Casey et al., 2018; Di Martino et al., 2017). These initiatives come with the benefits of higher statistical power and better validations across different sites, which can help identify reliable patterns of functional brain alterations in disorders such as MDD, ADHD, etc. But pooling of studies results in demographic heterogeneity and non-biological variability, interfering in identifying robust biomarkers depending on the task. In addition, the correlation between site effects and biological predictors can lead to an incorrect inference of non-biological differences as biological. Therefore, there is a need to develop novel methods to extract hierarchical patterns in large populations, which can be reliably used as biomarkers for various diseases and help understand the functioning of the human brain in multiple settings.

1.2 Aims of this thesis

This thesis aims to develop advanced machine learning methods based on matrix factorization and adversarial learning to extract reproducible and interpretable patterns using fMRI data that can be used as biomarkers of various diseases. This goal is divided into four parts as described below:

Aim 1: Extracting hierarchical Sparse Connectivity Patterns

Functional networks of the human brain are typically extracted at a single scale using various methods, including Independent Component Analysis (Smith et al., 2009), Non-Negative Matrix Factorization (Potluru and Calhoun, 2008) and Sparse Dictionary Learning (Lee et al., 2010; Eavani et al., 2015a). However, since numerous studies have suggested that the brain's functional organization is hierarchical (Meunier et al., 2009; Park and Friston, 2013), hierarchical decompositions might better capture functional connectivity patterns. Moreover, hierarchical decompositions can efficiently reduce the very high dimensionality of functional connectivity data and help analyze the data at different scales.

Several multi-scale community detection methods have been developed to understand the hierarchical organization of the human brain (Ferrarini et al., 2009; Al-Sharoa et al., 2018; Ashourvan et al., 2019b). In addition to traditional methods, recently, deep learning based methods have been introduced to estimate functional networks (Huang et al., 2017; Hu et al., 2018; Dong et al., 2019; Zhang et al., 2020b,b). These methods have shown promising results in terms of prediction performances. Still, there are one or more disadvantages: 1) removal of negative edge links in the method because the negative links are treated as repulsions, 2) the inability to capture a subject-specific representation of the patterns, and 3) "black-box" results due to non-linearity of the deep learning model causing loss of interpretability.

To overcome these limitations, we develop a method to extract group-level interpretable hierarchical patterns in the functional connectivity data while capturing heterogeneity in the population.

Aim 2: Adversarial learning for hSCP

Robust and reproducible estimation of hSCP can be challenged by inter-scanner variations, rsfMRI noise, irrelevant fluctuations, and other confounding variables. This can considerably reduce these components' reproducibility and hence their utility as biomarkers of diseases that disrupt functional connectivity. This can cause unreliable measurements of disease biomarkers.

To address this limitation, we introduce adversarial learning aiming to estimate hierarchical components that are robust to such confounding variations.

Aim3: Robust to site effects hSCPs

Multi-site fMRI studies have recently garnered much attention due to improved generalizability and replicability of brain patterns and evaluating a hypothesis in multiple sites/settings. However, they face the challenge that the pooling introduces systematic non-biological site-specific variance due to hardware, software, and environment. The non-biological variability introduced can affect the biomarkers or common features extracted from fMRI data (Yu et al., 2018; Shinohara et al., 2017). It can considerably reduce these features' reproducibility across different datasets and their utility as biomarkers for diseases that disrupt functional connectivity.

One of the common methods to remove site effects is the harmonization of data. Harmonization of fMRI data especially derived measures, is very nascent, even though it is much needed with the growing number of multi-site data sets (Adhikari et al., 2019). Many existing methods to reduce site effects are based on an empirical Bayes method ComBat (Johnson et al., 2007), has been applied for harmonizing different measures derived from functional MRI (Yu et al., 2018). However, ComBat and its variants such as ComBat-GAM (Pomponio et al., 2020) can not be directly applied to connectivity matrices since it can alter the structure of the connectivity matrix.

We tackle the above challenges by developing a matrix factorization and adversarial learning based framework to reduce the effects of non-biological variations introduced due to pooling data in multi-site studies.

Aim 4: Robust Hierarchical Patterns for identifying MDD patients

MDD is one of the most widespread psychiatric disorders characterized by abruptions in the connectivity between functional networks (Wu et al., 2011; Zhu et al., 2012). Several studies have used Multi Variate Pattern Analysis (MVPA) using the whole-brain functional connectivity matrix to find significant patterns of altered connectivity between functional networks (Zhong et al., 2017; Nakano et al., 2020; Yan et al., 2020). These methods are primarily two step procedures separating feature extraction from classification (Ravishankar et al., 2016; Hong et al., 2018). The first step consists of capturing interpretable patterns and the second step uses these patterns as input for prediction. Moreover, the methods focus on classification results at the expense of clinical interpretability.

Notably, consistent conclusions about the reproducible alternations of functional networks in MDD patients are limited. This can be attributed to small sample size, heterogeneity in data due to varied age and sex, and different hardware and software parameters when data is acquired from multiple sites (Button et al., 2013; Gong and He, 2015). Demographic heterogeneity and non-biological variability can interfere in identifying robust biomarkers depending on the task. Current multi-site fMRI studies do not address the issue of heterogeneity of the data and its effect on the reproducibility of extracted patterns.

Toward addressing the above limitations, we aim to reproducibly extract clinically relevant patterns of brain activity characterizing MDD in a large multi-site rsfMRI data.

1.3 Contributions

Contributions of this thesis are summarized as follows:

1. **Hierarchical Sparse Connectivity Patterns**: We introduced a deep matrix factorization based learning model to extract sparse, hierarchical, low-rank patterns. The method is termed as hierarchical Sparse Connectivity Patterns (hSCPS) and is formulated as a large-scale non-convex problem. The method is an extension of Sparse Connectivity Patterns (SCPs) by Eavani et al. (2015b). We utilize advancement in optimization algorithms, and use adaptive gradient descent along with alternating minimization to estimate the patterns. The method decomposes the functional connectivity matrix of each subject into a positively weighted set of sparse components, thus ensuring the positive semi-definite property of the input matrix. The sparse components are shared across all subjects, and weights are calculated for each individual, capturing heterogeneity in the data.

2. **Adversarial Learning**: We use the concept of attack and defense of adversarial learning

framework, traditionally used in supervised learning (Lowd and Meek, 2005; Farnia et al., 2018) to improve the robustness of the hSCPs. We introduce an unsupervised adversarial learning framework in our matrix factorization based approach. We model it as a two-player game where the attacker is trying to force the model to deviate from the optimal solution, and in defense, the model is trying to learn a robust solution. This method is validated using simulated data and a real dataset showing improvement in the reproducibility of the extracted components.

3. **Reduce site effects in hSCPs**: We introduce a novel model to reduce site effects in hSCPs. For this, the method learns site-specific features and global space, storing the information about the scanner and site, and uses these features to reduce site effects in the components. We also use an adversarial learning approach on top of our method to improve the reproducibility and generalizability of the components across components from the same site. We formulate the method as a non-convex optimization problem and solve it using adaptive stochastic gradient descent. Experiments on simulated and real datasets show that our method can improve the reproducibility of the components while retaining age-related biological variability in the data, thus capturing informative heterogeneity.

4. **Robust Generative-Discriminative hSCPs**: We introduce a novel method to extract robust patterns characterizing MDD patients in a large heterogeneous dataset. The proposed method consists of three building blocks: 1) a generative term to extract sparse patterns, 2) a discriminative term to guide the model to find patterns that are associated with MDD, and 3) an adversarial learning based term to reduce the effects of age, sex and site, thus reducing the heterogeneity. Our generative discriminative method has good classification results without compromising reproducibility and clinical interpretability of the patterns. This is ensured through the generative loss function that encourages good reconstruction and discriminative loss, which selects a subset of the patterns for classification. The findings of the method could expand our understanding of MDD from a functional network perspective. To the best of our knowledge, we are the first to propose a method that uses adversarial learning in a matrix

factorization framework to reliably and robustly extract changes in functional patterns in MDD from a purely data-driven perspective.

1.4 Organization of the Thesis

The two main methodological contributions of this thesis are described in Chapter 2, and 3 and 4, and application of the methods in Chapter 5. Chapter 2 discusses current methods used for finding functional patterns, describes hierarchical Sparse Connectivity Patterns (hSCPs) method and examines components extracted from the real dataset. In Chapter 3, we describe how adversarial learning can be used in our matrix factorization based unsupervised setting and show improvement in reproducibility using simulated and real datasets. Chapter 4 describes adversarial learning based matrix factorization method to reduce site/scanner effects in hSCPs. The method's effectiveness is demonstrated by showing more reproducible and interpretable components than the vanilla approach. In Chapter 5, we extract human brain patterns characterizing MDD and robust to effects of age, sex and site in a large multi-site dataset. Chapter 6 summarizes the contribution of the thesis, and discusses limitations and some interesting future directions.

CHAPTER 2

Extraction of Hierarchical Sparse Connectivity Patterns

In this chapter, we will introduce our framework to extract hierarchical Sparse Connectivity Patterns (hSCPS). We will show that our method can extract sparse interpretable patterns of the human brain with high reproducibility using real datasets and capturing heterogeneity in the data, which can be useful in understanding varied populations.

2.1 Introduction

The hierarchical organization has been observed in large-scale computer architectures (Ozaktas, 1992), communication systems (Akyildiz et al., 2005), and social networks (Nickel and Kiela, 2017). Such an organization provides a unique solution to balancing information within a group at a single scale and between groups at multiple scales. It also promotes optimal and efficient information processing and transmission in real-world information processing systems (Kinouchi and Copelli, 2006). The hierarchical organization is also seen in natural information processing systems such as the human brain, where this organization is present both in space (Bassett et al., 2010) and time (Chaudhuri et al., 2014).

It has been known that the human brain consists of spatially different regions which are functionally connected to form networks (Sporns, 2010). In addition, these networks are thought to be hierarchically organized in the brain (Guye et al., 2010; Meunier et al., 2009; Ferrarini et al., 2009; Doucet et al., 2011; Park and Friston, 2013; Sporns and Betzel, 2016). However, our understanding of the hierarchical nature of these networks is limited due to their complex nature.

Outline: We start with literature review of the existing functional connectivity based approaches in Section 2.2. In Section 2.3, we present the method for the extraction of hSCPs shared between rs-fMRI scans. Section 2.4 presents experimental results for validation of the method on simulated datasets and the effectiveness on the rs-fMRI scans of the 100 unrelated HCP subjects (Van Essen et al., 2013) and 969 subjects from the Philadelphia Neurodevelopmental Cohort (PNC) data set (Satterthwaite et al., 2014). We conclude with a discussion.

2.2 Literature Review

Seed based analysis: Correlation is a widely used statistical measure to estimate functional connectivity by calculating a correlation matrix storing correlation between time series of different "seed" Regions of Interests (ROIs). A high correlation between two seed regions would imply a strong functional connection. A hypothesis test is usually performed to reveal significant brain connections (Raichle, 2011). However, these approaches do not estimate the network structure since they can't cluster similar functional regions (Zhang et al., 2012). Another limitation is that it requires prior knowledge of seed which can bias the finding towards specific structures (Buckner et al., 2008).

Community detection approaches: Several interesting multi-scale community detection methods have been developed for estimating the underlying hierarchical organization of human brain connectivity (Ferrarini et al., 2009; Al-Sharoa et al., 2018; Akiki and Abdallah, 2019; Ashourvan et al., 2019b). Ashourvan et al. (2019a) proposed agglomerative ("bottom-up") type Hierarchical Community Detection, where the method begins by regarding each element as a separate network and then merging them into larger networks successively. Some approaches assume that the communities are independent (Betzel and Bassett, 2017; Puxeddu et al., 2020; Betzel et al., 2015) where they have investigated multi-scale brain networks and conducted multi-scale community detection by manipulating the number of communities. But, this is not the case in the human brain, where it is known that certain brain regions interact with multiple networks, i.e., the networks overlap (Xu et al., 2016).

Moreover, negative edge links are treated as repulsions in community detection approaches. Previously, most approaches used thresholds before their analysis and estimated networks using sparse graphs. The reason for thresholding was that the strong edges contain the most relevant information leading to the removal of negative edges. In contrast, in resting fMRI, a negative edge link carries essential information on functional co-variation with the opposing phase (Rubinov and Sporns, 2011) and has a substantial physiological basis (Zhan et al., 2017; Fox et al., 2005). These relations may play an important role in neuropsychiatric disorders and cognitive differentiation (Fitzpatrick et al., 2007). Some studies have recently shown that the weak network edges contain unique information that can not be revealed by analysis of just strong edges (Goulas et al., 2015; Santarnecchi et al., 2014). Assigning anti-correlated and correlated regions to the same component can reveal more details about the organization of the human brain patterns (Eavani et al., 2015a), as long as interpreted correctly. The major limitations of the community detection approaches are one or more than one of the following: 1) the assumption of independent components, 2) not capturing heterogeneity in the data, and 3) inability to detect weights while estimating links.

Matrix factorization methods: These approaches are also widely used along with community detection methods. Most prevalent among them are Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF) and Principle Component Analysis (PCA)(Beckmann et al., 2005; Calhoun et al., 2009; Anderson et al., 2014; Zhou et al., 2009). ICA is increasingly used to estimate spatio/temporal patterns directly using fMRI time series (Calhoun et al., 2009). The main idea behind ICA is to extract independent patterns using higher-order moments. Spatial ICA is widely utilized to obtain spatially independent patterns, commonly called "Intrinsic Connectivity Networks (ICNs)" (Calhoun et al., 2003). In some cases, PCA is combined with ICA to extract these independent patterns (Zhou et al., 2009; Allen et al., 2014). Smith et al. (2012) extended the above method to find overlapping brain patterns by applying temporal ICA, thus overcoming the limitation of non-overlapping patterns recovered by spatial ICA. Even though there is an improvement, these patterns are still based on the notion of independent temporal structure, contrasting to the idea that different brain regions take part in multiple complex functions.

Sparse representation and dictionary learning are another set of methods that have gained attention in performing fMRI data analysis. Numerous novel approaches have been developed focussing on finding sparse interpretable representation of fMRI data (Aharon et al., 2006; Lee et al., 2010; Lv et al., 2014, 2017). One of the most prominent examples is sparse connectivity patterns (SCPs) by Eavani et al. (2015a) which is the motivation of our work. It extracts sparse low-rank patterns using matrix factorization of functional connectivity matrices while capturing subject-specific information. Various tensor decomposition works have been introduced based on the similar idea of factorizing connectivity matrices into a lower-dimensional subspace (Wang et al., 2011; Hamdi et al., 2018; Noroozi and Rezghi, 2020; Zhang et al., 2020c).

Non-negative Matrix Factorization (NMF) (Yang and Leskovec, 2013; Anderson et al., 2014) is another common matrix decomposition approach that many researchers use for obtaining information about community structure by analyzing low dimensional matrix. Recently, (Li et al., 2018a) used Deep Semi Non-negative Matrix Factorization (Trigeorgis et al., 2017) for estimating hierarchical, potentially overlapping, functional networks. The model given by (Li et al., 2018a) could only find networks containing regions with a positive correlation between them as the method is based on non-negative matrix factorization, thus limiting the model to only use positive matrices.

Deep learning based methods: The rise of deep learning has led researchers to move towards a non-linear model with high representational and prediction power. Many studies have proposed deep learning based framework for reconstructing functional networks, e.g., the Deep Convolutional Auto Encoder (DCAE), Deep Belief Network (DBN) and Convolutional Neural Network (CNN), Restricted Boltzmann Machine (RBM) (Hu et al., 2018; Huang et al., 2017; Dong et al., 2019; Zhang et al., 2020b). These methods have reported the meaningful hierarchical temporal organization of fMRI time series in the task-evoked fMRI data. Deep learning models have also found great utility for disease prediction using functional connectivity. Most of the models use deep learning as a black-box model and focus on disease classification, for example, Multi-Layered Perceptrons (Heinsfeld et al., 2018), Deep Belief Networks (Akhavan Aghdam et al., 2018), Convolutional Neural Networks (Khosla et al., 2018), Graph Convolutional Network (Wang et al., 2021). Still, there are one or more disadvantages: 1) requirement of large training samples, 2) large computational resources (GPUs or TPUs), 3) considerable training time; 4) low reproducibility of the extracted patterns, 5) nonexistence of positively and negatively correlated nodes in a component, 6) inability to capture heterogeneity in the data, and 7) “black box” results lacking explainability mainly due to non-linearity in the hierarchical associations.

2.3 Method

Our method aims to find Hierarchical Sparse Connectivity Patterns (hSCPs) by jointly decomposing correlation matrices into multiple components having different ranks using a cascaded framework for matrix factorization. hSCP addresses aforementioned limitations by modeling the fMRI data to capture essential properties of the network, namely- 1) Sparsity: only a small subset of nodes interact with other nodes in a given network; 2) Heterogeneity: some networks might be more prominent in particular individuals as compared to others; 3) Existence of positively and negatively correlated nodes in a network; 4) Overlapping networks, which is likely to reflect true brain organization, as brain networks might share certain regional components; and 5) Hierarchy: By adding extra layers of abstraction we can learn latent attributes and the hierarchy in the networks. Our method is built upon Sparse Connectivity Patterns (SCPs) (Eavani et al., 2015a) which can be considered a symmetric CP decomposition for which an indirect fitting procedure makes the model structure equivalent to the PARAFAC2 model representation considered in (Madsen et al., 2017) with the addition of sparsity rather than orthogonality.

Notations and Conventions: Lowercase boldface letters are used for vectors, and for matrices, we use capital boldface letters. An element of a vector x is denoted by x_i, and an element of a matrix A is denoted by $A_{i,j}$. The set of symmetric positive definite matrices of size $P \times P$ is denoted by $\mathbb{S}_{++}^{P\times P}$. Matrix $\mathbf{A}$ with all the elements greater than or equal to 0 is denoted by $\mathbf{A} \geq 0$. $\mathbf{J}_P$ denotes $P \times P$ matrix with all elements equal to one. $P \times P$ identity matrix is denoted by $\mathbf{I}_P$ and element-wise product between two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \circ \mathbf{B}$. We will be using the same notations and conventions in the upcoming chapters.

2.3.1 Sparse Connectivity Patterns

Let $\mathbf{X}^n \in \mathbb{R}^{P\times T}$ be the fMRI data of the n^{th} subject having P regions and T time points, and $\mathbf{\Theta}^n \in \mathbb{S}_{++}^{P\times P}$ is the correlation matrix where $\mathbf{\Theta}^n_{m,o} = \mathrm{corr}(\mathbf{x}^i_m, \mathbf{x}^i_o)$ is the correlation between time series of m^{th} and o^{th} node. We first define the model for estimating the Sparse Connectivity Patterns (SCPs) (Eavani et al., 2015a) in the fMRI data which decomposes the correlation matrices into

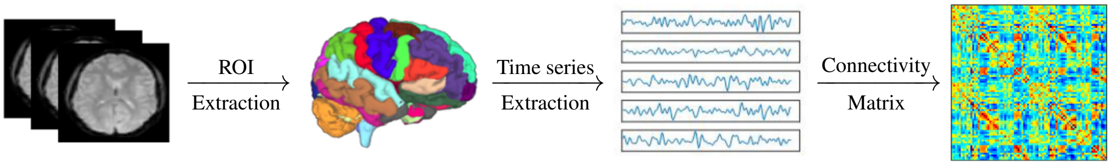


Figure 1: Schematic illustrating extraction of functional connectivity. We first extract region averaged time series using a parcellation scheme and then extract the functional connectome storing pairwise correlation in the average time courses.

non-negative linear combination of sparse low rank components such that for all $n = 1, \dots, N$ we have $\mathbf{\Theta}^n \approx \mathbf{W}\mathbf{\Lambda}^n\mathbf{W}^T$ where $\mathbf{W} \in \mathbb{R}^{P \times k}$ is a set of shared patterns across all subjects, $k < P$ and $\mathbf{\Lambda}^n \succeq 0$ is a diagonal matrix storing the subject specific information about the strength of each of the components. Let $\mathbf{w}_l \in \mathbb{R}^P$ be the l^{th} column of $\mathbf{W}$ such that $-1 \preceq \mathbf{w}_l \preceq 1$ and let $w_{l,s}$ be the s^{th} element of $\mathbf{w}_l$ vector, then $\mathbf{w}_l$ represents a component which reflects the weights of the nodes in the component and if $w_{l,s}$ is zero then s^{th} node does not belong to l^{th} component. If the sign of weights of any two nodes in a component is same then they are positively correlated else they have anti-correlation. To make the patterns sparse, each column of $\mathbf{W}$ was subjected to L_1 penalty and the below optimization is solved to obtain the SCPs:

$$\begin{aligned} \underset{\mathbf{W},\mathbf{\Lambda}}{\text{minimize}} \quad & \sum_{n=1}^{N} ||\mathbf{\Theta}^i - \mathbf{W}\mathbf{\Lambda}^n\mathbf{W}^T||_F^2 \\ \text{subject to} \quad & \|\mathbf{w}_l\|_1 \leq \lambda, l = 1, \dots, k, \\ & \|\mathbf{w}_l\|_\infty \leq 1, l = 1, \dots, k, \\ & \mathbf{\Lambda}^i \succeq 0, n = 1, \dots, N, \end{aligned} \tag{2.1}$$

where S is the total number of subjects and λ controls the sparsity of the components.

2.3.2 Hierarchical Sparse Connectivity Patterns

We have extended the above work and introduced Hierarchical Sparse Connectivity Patterns (hSCPs) to estimate hierarchical sparse low rank patterns in the correlation matrices. In our model, a

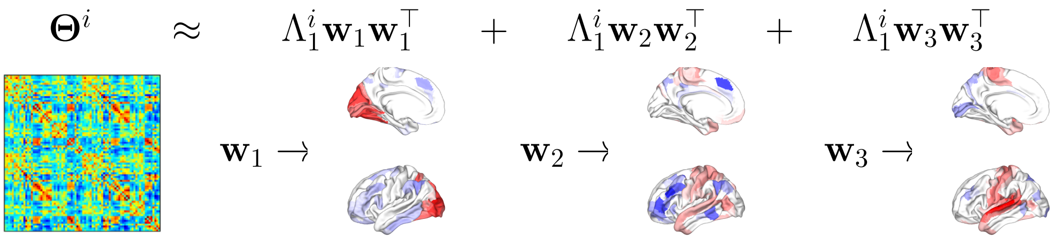

$\boldsymbol{\Theta}^i$
$\approx$
$\Lambda_1^i \mathbf{w}_1 \mathbf{w}_1^\top$
$+$
$\Lambda_1^i \mathbf{w}_2 \mathbf{w}_2^\top$
$+$
$\Lambda_1^i \mathbf{w}_3 \mathbf{w}_3^\top$
$\mathbf{w}_1 \rightarrow$
$\mathbf{w}_2 \rightarrow$
$\mathbf{w}_3 \rightarrow$

Figure 2: Schematic illustrating SCP where $\boldsymbol{\Theta}^i$ is a subject specific correlation matrix which is approximated by non-negative sum of sparse rank one matrices.

correlation matrix is decomposed into K levels as -

$$\begin{aligned}
\boldsymbol{\Theta}^n &\approx \mathbf{W}_1 \boldsymbol{\Lambda}_1^n \mathbf{W}_1^T, \\
\boldsymbol{\Theta}^n &\approx \mathbf{W}_1 \mathbf{W}_2 \boldsymbol{\Lambda}_2^n \mathbf{W}_2^T \mathbf{W}_1^T, \\
&\vdots \\
\boldsymbol{\Theta}^n &\approx \mathbf{W}_1 \mathbf{W}_2 \ldots \mathbf{W}_K \boldsymbol{\Lambda}_K^n \mathbf{W}_K^T \mathbf{W}_{K-1}^T \ldots \mathbf{W}_1^T,
\end{aligned} \tag{2.2}$$

where $\mathbf{W}_1 \in \mathbb{R}^{P \times k_1}$ and $\mathbf{W}_q \in \mathbb{R}^{k_{q-1} \times k_q}$, $\boldsymbol{\Lambda}_q^i \in \mathbb{R}^{k_q \times k_q}$ is a diagonal matrix storing subject specific information of the patterns, $P \gg k_1 > k_2 > \ldots > k_K$, $P \gg K$ and $\mathbf{W}^T$ is the transpose of $\mathbf{W}$. In the above formulation, $\mathbf{W}_1 \in \mathbb{R}^{P \times k_1}$ stores k_1 components at the bottom most level, and each successive multiplication by $\mathbf{W}_2$, $\mathbf{W}_3$, $\ldots$, $\mathbf{W}_K$ linearly transforms to a lower dimensional space of k_2, k_3, $\ldots$, k_K dimension. Here k_r is the number of components at the r^{th} level, note that k_1 is the number of components at the lower most level of the hierarchy. If we consider 2 layer hierarchical representation of a given correlation matrix then we can define $\mathbf{Z}_1 = \mathbf{W}_1 \mathbf{W}_2$ to be a $P \times k_2$ matrix, then $\mathbf{Z}_1$ is a coarse network which consist of weighted linear combination of $\mathbf{W}_1$ which are fine level components where weights are stored in $\mathbf{W}_2$.

For better interpretability, for noise reduction in the model, but also because of our hypothesis that brain subnetworks are relatively sparse (Achard and Bullmore, 2007), we have introduced sparsity constraints on the $\mathbf{W}$ matrices. By making $\mathbf{W}_1$ sparse we are forcing the components to contain few number of nodes and by forcing rest of the $\mathbf{W}$s to be sparse, we are forcing that the components at each of the next level are sparse linear combination of previous components. The hierarchical

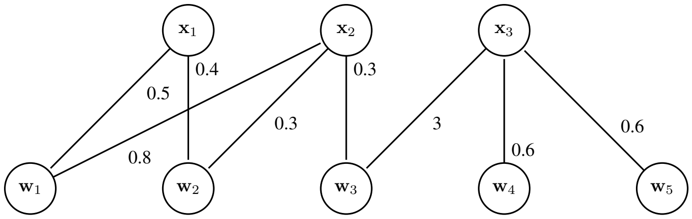


Figure 3: Example of 2-layer hierarchical structure

networks can be estimated by solving the below minimization procedures simultaneously under the constraints mentioned above

$$\begin{aligned}
&\min_{W_1,\Lambda_1} \sum_{n=1}^{N} \|\boldsymbol{\Theta}^n - \mathbf{W}_1 \boldsymbol{\Lambda}_1^n \mathbf{W}_1^T\|_F^2, \\
&\min_{W_1,W_2,\Lambda_2} \sum_{n=1}^{N} \|\boldsymbol{\Theta}^n - \mathbf{W}_1 \mathbf{W}_2 \boldsymbol{\Lambda}_2^n \mathbf{W}_2^T \mathbf{W}_1^T\|_F^2, \\
&\vdots \\
&\min_{\mathcal{W},\Lambda_K} \sum_{n=1}^{N} \|\boldsymbol{\Theta}^i - \mathbf{W}_1 \mathbf{W}_2 \ldots \mathbf{W}_K \boldsymbol{\Lambda}_K^i \mathbf{W}_K^T \mathbf{W}_{K-1}^T \ldots \mathbf{W}_1^T\|_F^2,
\end{aligned} \tag{2.3}$$

where $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_K\}$ is the set storing sparse components shared across all subjects. As the above minimization procedures are inter-dependent, we need to solve them jointly. Let $\mathcal{D} = \{\boldsymbol{\Lambda}_r^n \mid r = 1, \ldots, K; n = 1, \ldots, N\}$ be set storing subject specific diagonal matrix with $\boldsymbol{\Lambda}_r^n \geq 0$ and $\mathcal{C} = \{\boldsymbol{\Theta}^n \mid n = 1, \ldots, N\}$ be the set storing correlation matrix for all subjects. The hierarchical components are estimated by solving the below optimization problem:

$$\begin{aligned}
\underset{\mathcal{W},\mathcal{D}}{\text{minimize}} \quad & H(\mathcal{W}, \mathcal{D}, \mathcal{C}) = \sum_{n=1}^{N} \sum_{r=1}^{K} \|\boldsymbol{\Theta}^n - (\prod_{j=1}^{r} \mathbf{W}_j) \boldsymbol{\Lambda}_r^n (\prod_{j=1}^{r} \mathbf{W}_j)^\top\|_F^2 \\
\text{subject to} \quad & \|\mathbf{w}_l^r\|_1 < \lambda_r, l = 1, \ldots, k_r \quad \text{and} \quad r = 1, \ldots, K, \\
& \|\mathbf{w}_l^r\|_\infty \leq 1, l = 1, \ldots, k_r \quad \text{and} \quad r = 1, \ldots, K, \\
& \mathbf{W}_j \geq 0, j = 2, \ldots, K, \\
& \boldsymbol{\Lambda}_r^n \succeq 0, n = 1, \ldots, N \quad \text{and} \quad r = 1, \ldots, K, \\
& \text{trace}(\boldsymbol{\Lambda}_r^i) = 1, n = 1, \ldots, N \quad \text{and} \quad r = 1, \ldots, K,
\end{aligned} \tag{2.4}$$

where trace operator calculates sum of diagonal elements of a matrix. In the above minimization procedure, the sum of diagonal values of Λ^n is fixed to be 1 such that the sparsity of W is not trivially minimized. We will be denoting above constraint set as $\Omega_{\mathcal{W}} = \{\mathbf{W} \mid \|\mathbf{w}_l^r\|_1 < \tau_r,\ \|\mathbf{w}_l^r\|_\infty \leq 1, \mathbf{W}_j \geq 0,\ j = 2, \dots, K\}$ and $\Psi = \{\mathbf{\Lambda} \mid \mathrm{trace}(\mathbf{\Lambda}_r^n) = 1, \mathbf{\Lambda}_r^n \geq 0\}$.

Note In the above formulation, the last level has the highest number of components k_1, and in the level after that we have k_2 number of components which are linear combination of components at previous level, so on and so forth. In this way, we have built up a hierarchical model where each component is made up of linear combination of components at the previous hierarchy. Note that we can not just use the last decomposition in the above architecture to get the hierarchy as different layers have different ranks and different approximations, hence we will need all the approximations to build the hierarchical structure. In addition, one would expect $\mathbf{W}_2$ and $\mathbf{W}$s to be degenerate, but that would be the case only when $\mathbf{W}_1$ is orthogonal matrix. Consider the case where we have a two level hierarchy, we can have better approximation by taking a linear combination of columns of $\mathbf{W}_1$ which we have also observed empirically.

2.3.3 Alternating Minimization

The optimization problem defined in 2.4 is a non-convex problem which we solved using alternating minimization. Algorithm 1 describes the complete alternating minimization procedure where $\mathrm{proj}_1(\mathbf{W}, \lambda)$ operator projects each column of $\mathbf{W}$ into intersection of L_1 and L_∞ ball (Podosinnikova et al., 2013), and proj_2 projects a matrix onto $\mathbb{R}_+$ by making all the negative elements in the matrix equal to zero. As the gradients are not globally Lipschitzs, we don't have bounds on the step size for the gradients. For that reason, we have used AMSGrad (Reddi et al., 2019), ADAM (Kingma and Ba, 2014) and NADAM (Dozat, 2016) as gradient descent algorithms which have adaptive step size, and the update rules are defined in Appendix A. $\mathrm{descent}$ function in the Algorithm 1 is the update rule used by different gradient descent techniques. The cost of computing gradients of $\mathbf{\Lambda}$ is $\mathcal{O}(KSP^2k_1)$ and of $\mathbf{W}$ is $\mathcal{O}(KSP^2k_1 + K^2SPk_1^2)$. The overall cost of Algorithm 1 is number of iterations $\times\mathcal{O}(KSP^2k_1 + K^2SPk_1^2)$. From our previous assumption that $P \gg K$, the final cost is number of iterations $\times\mathcal{O}(KSP^2k_1)$.

Algorithm 1 hSCP

```
Input: Data $\mathcal{C}$, number of connectivity patterns $k_1,\ldots,k_K$ and sparsity $\lambda_1,\ldots,\lambda_K$ at different level
$\mathcal{W}$ and $\mathcal{D}$ = Initialization($\mathcal{C}$)
repeat
    for $r = 1$ to $K$ do
        $\mathbf{W}_r \leftarrow \mathrm{descent}(\mathbf{W}_r)$
        if $r == 1$ then
            $\mathbf{W}_r \leftarrow \mathrm{proj}_1(\mathbf{W}_r, \lambda_r)$
        else
            $\mathbf{W}_r \leftarrow \mathrm{proj}_2(\mathbf{W}_r)$
        for $n = 1, \ldots, N$ do
            $\mathbf{\Lambda}_r^n \leftarrow \mathrm{descent}(\mathbf{\Lambda}_r^n)$
            $\mathbf{\Lambda}_r^n \leftarrow \mathrm{proj}_2(\mathbf{\Lambda}_r^n)$
until Stopping criterion is reached
Output: $\mathcal{W}$ and $\mathcal{D}$
```

2.3.4 Gradients

In this section, we define gradients used for alternating gradient descent. Let

$$\tilde{\mathbf{W}}_0 = \mathbf{W}_0 = \mathbf{I}_P, \qquad \mathbf{Y}_r = \prod_{j=0}^{r} \mathbf{W}_j, \qquad \mathbf{T}^r_{m,n} = (\prod_{j=1}^{m-r} \mathbf{W}_j)\mathbf{\Lambda}^n_{m-r}(\prod_{j=1}^{m-r} \mathbf{W}_j)^\top.$$

The gradient of H with respect to $\mathbf{W}_r$ is written as:

$$\frac{\partial H}{\partial \mathbf{W}_r} = \sum_{n=1}^{N}\sum_{j=r}^{K} -4\mathbf{Y}_{r-1}^\top \mathbf{X}_n \mathbf{Y}_{r-1}\mathbf{W}_r\mathbf{T}^r_{j,n} + 4\mathbf{Y}_{r-1}^\top \mathbf{Y}_{r-1}\mathbf{W}_r\mathbf{T}^r_{j,n}\mathbf{W}_r^\top \mathbf{Y}_{r-1}^\top \mathbf{Y}_{r-1}\mathbf{W}_r\mathbf{T}^r_{j,n}.$$

The gradient of H with respect to $\mathbf{\Lambda}_r^n$ is:

$$\frac{\partial H}{\partial \mathbf{\Lambda}_r^i} = (-2\mathbf{Y}_r^T \mathbf{\Theta}_r^i \mathbf{Y}_r + 2\mathbf{Y}_r^T \mathbf{Y}_r \mathbf{\Lambda}_r^i \mathbf{Y}_r^T \mathbf{Y}_r) \circ \mathbf{I}_{k_r}.$$

2.3.5 Initialization procedure for Gradient Descent

Single level matrix decomposition considered in hSCP is structurally similar to Singular Value Decomposition (SVD) but with the dependent components and sparsity added. Hence, we believe

Algorithm 2 Initialization

```
Input: Data $\Theta$
for $r = 1$ to $K$ do
    for $n = 1$ to $N$ do
        if $r == 1$ then
            $\mathbf{U}^n\mathbf{V}^n(\mathbf{U}^i)^T = k_1$- rank SVD($\mathbf{\Theta}^n$)
        else
            $\mathbf{V}^n = k_{r-1}$top values of $\mathbf{\Lambda}^n_{r-1}$
            $\mathbf{U}^n =$ Permutation matrix
        $\mathbf{\Lambda}^n_r = \mathbf{V}^n$
    $\mathbf{W}_r = \frac{1}{N}\sum_{n=1}^{S}(\mathbf{U}^n)$
Output: $\mathcal{W}$ and $\mathcal{D}$
```

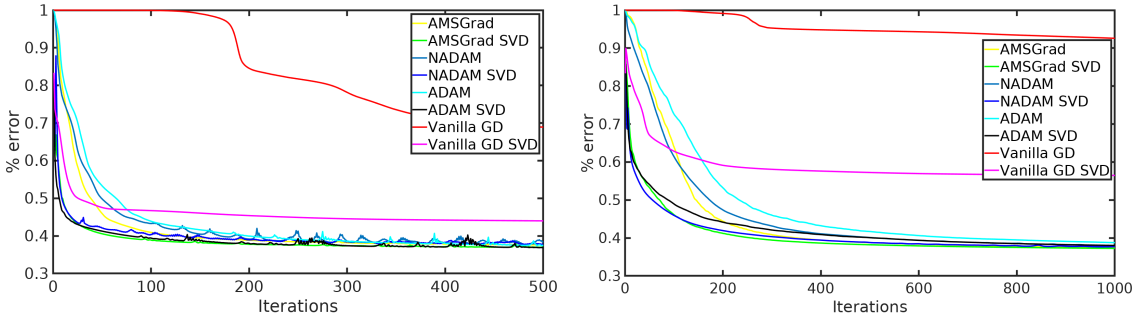
that the final components estimated are a modification of singular vectors. Thus, we have initialized the $\mathcal{W}$ and $\mathcal{D}$ in Algorithm 1 by taking SVD of input data matrix. This helps in making algorithm deterministic. Define $\bar{\mathbf{\Theta}}$ as the sample mean of $\mathbf{\Theta}_n$. We then perform k-rank SVD of $\bar{\mathbf{\Theta}}$ and obtain $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{UVU}^\top =$ k-rank SVD of $\bar{\mathbf{\Theta}}$. We then initialize $\mathbf{W}_1$ by $\mathbf{U}$ and $\mathbf{\Lambda}^n_1$ by $\mathbf{V}^n$ where $\mathbf{V}^n$ can be obtained by taking k-rank SVD of $\mathbf{\Theta}_n$ as described in Algorithm 2. For $r > 1$, $\mathbf{W}_r$ can be initialized as a permutation matrix and $\mathbf{\Lambda}_r$ by top k_r diagonal elements of k_{r-1} so that we don't have to perform SVD at each level. We empirically show in the next section that SVD initialization results in faster convergence.

2.4 Experiments

2.4.1 Dataset

We used two real dataset for demonstrating the effectiveness of the method:

1. HCP- Human Connectome Project (HCP) (Van Essen et al., 2013) dataset is one of the widely used dataset for fMRI analysis containing fMRI scans of 100 unrelated subjects as provided at the HCP 900 subjects data release (Van Essen et al., 2012) which were processed using ICA+FIX pipeline with MSMAll registration (Glasser et al., 2013). Each subject has 4004 time points and the time series were normalized to zero mean and unit L2 norm, averaged over the 360 nodes of the multimodal HCP parcellation (Glasser et al., 2016).



(a) $k_1 = 15, k_2 = 5, \lambda_1 = 9, \lambda_2 = 7.5$

(b) $k_1 = 15, k_2 = 5, \lambda_1 = 3.6, \lambda_2 = 3$

Figure 4: Performance comparison of different gradient descent techniques. SVD corresponds to gradient descent with SVD initialization

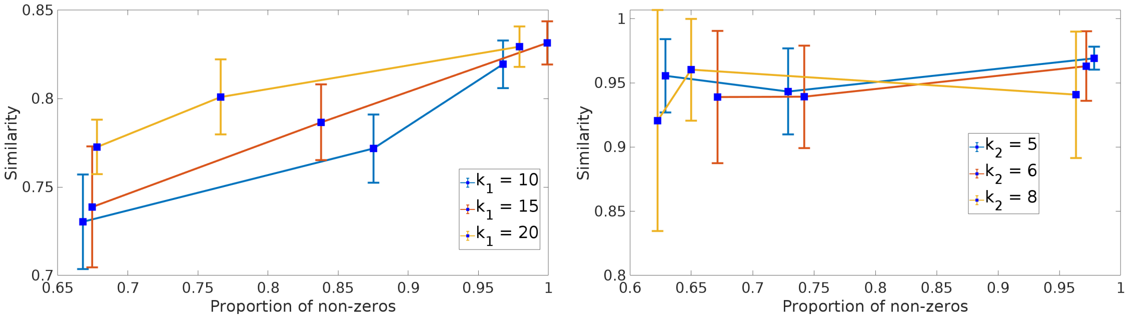
2. PNC- Philadelphia Neuro-developmental Cohort (PNC) (Satterthwaite et al., 2014) dataset contains 969 subjects (ages from 8 to 22) each having 120 time points and 121 nodes described in (Doshi et al., 2016). The data were preprocessed using an optimized procedure (Ciric et al., 2017) which includes slice timing, confound regression, and band-pass filtering.

2.4.2 Convergence Analysis

We compare AMSGrad, ADAM, NADAM and vanilla gradient descent with SVD initialization and random initialization by measuring percentage error which is defined as:

$$\frac{\sum_{n=1}^{N}\sum_{r=1}^{K}||\mathbf{\Theta}^n - (\prod_{j=1}^{r}\mathbf{W}_j)\mathbf{\Lambda}_r^n(\prod_{j=1}^{r}\mathbf{W}_j)^T||_F^2}{\sum_{n=1}^{N}\sum_{r=1}^{K}||\mathbf{\Theta}^n||_F^2}.$$

For fair comparison, we set $\beta_1 = 0.9$ and $\beta_2 = 0.99$ for ADAM, NADAM and AMSGRAD algorithm, where β_1 and β_2 are the hyperparameters used in the update rules of the gradient descent algorithms. These are values are typically used as parameter settings for adaptive gradient descent algorithms (Reddi et al., 2019). Figure 4 shows the convergence of the algorithm on the complete HCP data for two different combinations of sparsity parameters at a particular set of k_1 and k_2. From the Figure 4 we can see that the AMSGrad has the best convergence and SVD initialization gives a better convergence rate. For rest of the experiments we have used AMSGrad algorithm with SVD initialization to perform gradient descent.



(a) Similarity comparison at fine scale

(b) Similarity comparison at coarse scale

Figure 5: Comparison between ground truth and extracted hSCPs components on simulated dataset. X axis corresponds to proportion of non-zeros in the estimated components.

2.4.3 Simulation

To evaluate the performance of the proposed model, we first use synthetic data. We compared the hierarchical components extracted from hSCP to hierarchical overlapping communities obtained using EAGLE (Shen et al., 2009) and OSLOM (Lancichinetti et al., 2011). Implementation of EAGLE and OSLOM was obtained from the authors. We randomly generate $\mathbf{V}_1 \in \mathbb{R}^{p\times k_1}$ with percentage of non-zeros equal to μ_1, $\mathbf{W}_2 \in \mathbb{R}^{k_1\times k_2}$ with percentage of non-zeros equal to μ_2 and $\mathbf{\Lambda}^n \in \mathbb{R}^{k_2\times k_2}$ for $n = 1, \ldots, N$. The goal is to generate $\mathbf{V}_1\mathbf{W}_2\mathbf{\Lambda}^n\mathbf{W}_2^T\mathbf{V}_1^T$ matrices which are close to a correlation matrix. For this, we first take mean of all $\mathbf{\Lambda}^i$ such that $\mathbf{U} = \frac{1}{n}\sum_{n=1}^{N}\mathbf{\Lambda}^n$ and generate $\mathbf{T}$ such that $\mathbf{T} = \mathbf{V}_1\mathbf{W}_2\mathbf{U}\mathbf{W}_2^T\mathbf{V}_1^T$. Now, let $\mathbf{D}$ be a matrix containing diagonal elements of $\mathbf{T}$, to make $\mathbf{T}$ a correlation matrix, we modify $\mathbf{V}_1$ by multiplying it by $\mathbf{D}^{\frac{1}{2}}$. Let $\mathbf{W}_1 = \mathbf{D}^{\frac{1}{2}}\mathbf{V}_1$, then $\mathbf{R} = \mathbf{W}_1\mathbf{W}_2\mathbf{U}\mathbf{W}_2^T\mathbf{W}_1^T$ would be a correlation matrix. Now, we generate correlation matrix for each subject by using the below equation

$$\mathbf{\Theta}^n = \mathbf{W}_1\mathbf{W}_2\mathbf{\Lambda}^n\mathbf{W}_2^T\mathbf{W}_1^T + \mathbf{E}_n,$$

where $\mathbf{E}_n \in \mathbb{R}^{p\times p}$ such that $\mathbf{W}_1\mathbf{W}_2\mathbf{\Lambda}^n\mathbf{W}_2^T\mathbf{W}_1^T$ matrix becomes a correlation. For the experiments, the parameters were set as follows: $n = 300$, $p = 100$ $k_1 = 20$, $k_2 = 6$, $\mu_1 = 0.4$ and $\mu_2 = 0.5$.

We compare components derived from hSCP with $k_1 \in \{10, 15, 20\}$, $k_2 \in \{5, 6, 8\}$, $\lambda_1 \in P \times$

		10	15	20
5	hSCP	0.8293 ± 0.0467	0.8097 ± 0.0728	0.8305 ± 0.0614
	EAGLE	0.4051 ± 0.0304	0.4180 ± 0.0290	0.4068 ± 0.0070
	OSLOM	0.6866 ± 0.0442	0.6955 ± 0.0362	-
6	hSCP	0.8421 ± 0.0585	0.8660 ± 0.0286	0.8497 ± 0.0292
	EAGLE	0.3867 ± 0.0141	0.4855 ± 0.0731	0.4463 ± 0.0334
	OSLOM	0.6249 ± 0.0554	0.7302 ± 0.0431	-
8	hSCP	0.8350 ± 0.0666	0.8457 ± 0.0353	0.8454 ± 0.0385
	EAGLE	0.4408 ± 0.0857	0.5339 ± 0.0900	0.4099 ± 0.0274
	OSLOM	0.6610 ± 0.0540	-	-

Table 1: Similarity comparison (mean±std) on simulated dataset. The rows correspond to values of k_1 and the columns correspond to values of k_2.

$5(10^{[-3:-1]})$ and $\lambda_2 \in k_1 \times 10^{[-3:-1]}$. By varying λ values, we generate components with different sparsity. We first compare fine-scale and coarse-scale components separately to demonstrate the effect of sparsity on the performance. For a fixed k_1 and λ_1, we find k_2 and λ_2 giving the maximum similarity with the ground truth and for a fixed k_2 and λ_2, we find k_1 and λ_1 giving the maximum similarity with the ground truth over 10 runs. Here the similarity is defined as the average correlation between extracted and the ground truth components. Fig. 5 shows the similarity of the fine-scale and coarse-scale components with the ground truth. From the figure, we can see that the hSCP can extract components that are highly similar to the ground truth. Also, as the fine-scale components become sparse, the similarity decreases. Next, we compare hSCP to EAGLE and OSLOM. Hierarchical components and communities with $k_1 \in \{5, 6, 8\}$, $k_2 \in \{10, 15, 20\}$ were extracted from hSCP, EAGLE and OSLOM. The correlation matrix averaged across all the subjects was used as an input to EAGLE and OSLOM. For hSCP, among different values of λ_1 and λ_2, we extract components at level k_1 and k_2, which have maximum similarity with the ground truth. Table 1 shows the similarity of the extracted components with the ground truth. Some cells in the tables are empty as the EAGLE and OSLOM algorithms were not able to generate hierarchical structures for particular values of k_1 and k_2. It can be seen that the hSCP method can extract the components which are closer to ground truth as compared to other methods.

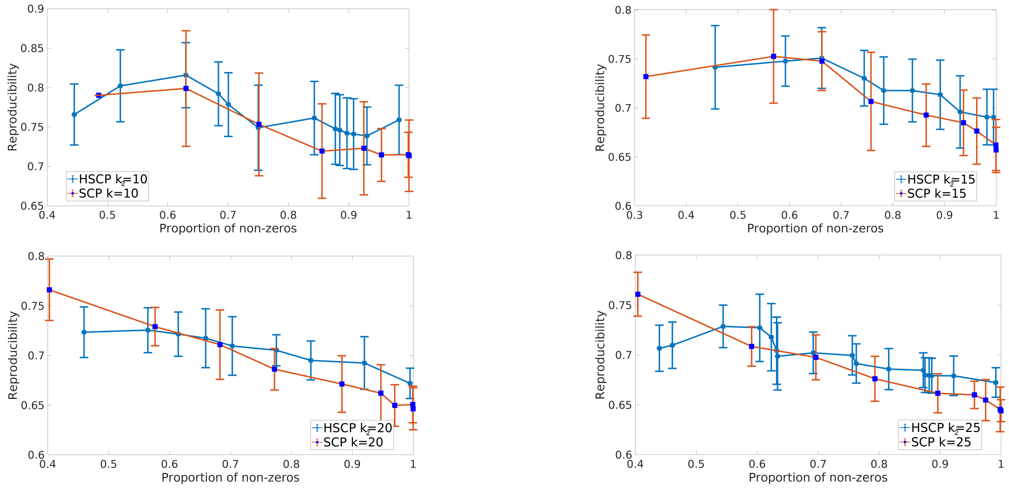


Figure 6: Comparison of single scale (SCP) and hierarchical (HSCP) components on HCP dataset. X axis corresponds to proportion of non-zeros in the components. All the HSCP are second level components.

2.4.4 Comparison with single scale components

We also compared the reproducibility of the shared components extracted from the hierarchical model (hSCP) versus single scale components (SCP). Reproducibility here is defined as the normalized inner product of components derived from the two equal random sub-samples of the data averaged across all the components. We decomposed the correlation matrix into two levels to demonstrate the advantages of hierarchical factorization and show proof of concept. There might not be a single K that best describes the data, and the algorithm allows us to investigate the continuum of functional connectivity patterns at different Ks. We compare components derived from hSCP with $k_1 \in \{10, 15, 20, 25\}$, $k_2 \in \{4, 5, 6, 8\}$, $\lambda_1 \in P \times 5(10^{[-3:-1]})$ and $\lambda_2 \in k_1 \times 10^{[-3:-1]}$, and from SCP with $k \in \{4, 5, 6, 8, 10, 15, 20, 25\}$ at $\lambda \in P \times 5(10^{[-4:-1]})$. At a fixed k_2 and λ_2, we find the optimal k_1 and λ_1 by dividing the data into three equal parts: training, validation, and test data, and choosing the parameters corresponding to maximum mean reproducibility over 20 runs on training and validation set. Figure 6 and Figure 7 show the reproducibility of the components averaged over 20 runs on training and test data. We can see that the same number of components extracted from the second level using hSCP are, on average, more reproducible than the components extracted using SCP.

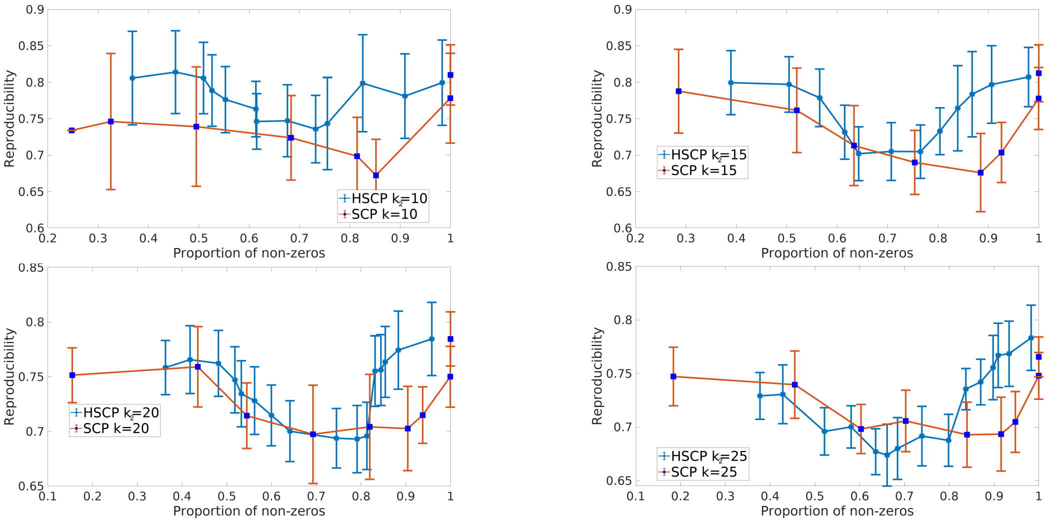


Figure 7: Comparison of single scale (SCP) and hierarchical (HSCP) components on PNC dataset. X axis corresponds to proportion of non-zeros in the components. All the HSCP are second level components.

2.4.5 Comparison of hSCP with existing approaches

We compared the reproducibility of hierarchical components extracted from hSCP to hierarchical overlapping communities obtained using EAGLE (Shen et al., 2009) and OSLOM (Lancichinetti et al., 2011). Implementation of EAGLE and OSLOM was obtained from the authors. Correlation matrix averaged across all the subjects was used as an input to EAGLE and OSLOM. Hierarchical components and communities with $k_1 \in \{4, 5, 6, 8\}$, $k_2 \in \{10, 15, 20, 25\}$ were generated from hSCP, EAGLE and OSLOM. Optimal λ_1 and λ_2 for hSCP were selected by dividing the data into three equal parts: training, validation and test set, and performing the validation procedure as described in Subsection 2.4.4. Reproducibility was computed using training and test for all the methods for all combinations of k_1 and k_2. Table 2 and Table 3 show the reproducibility results on HCP and PNC datasets. For a particular k_1 and k_2, reproducibility table show the average of the two reproducibility values. The results clearly show that the hSCPs have better reproducibility than the communities derived using EAGLE and OSLOM.

		10	15	20
4	hSCP	0.8885 ± 0.0441	0.8351 ± 0.0748	0.8507 ± 0.0635
	EAGLE	0.3077 ± 0.0981	0.4158 ± 0.1321	-
	OSLOM	0.7493 ± 0.0882	-	-
5	hSCP	0.8753 ± 0.0348	0.8356 ± 0.0591	0.8281 ± 0.0656
	EAGLE	0.2908 ± 0.0737	0.2664 ± 0.0333	0.0792 ± 0.1656
	OSLOM	0.6092 ± 0.0733	-	-
6	hSCP	0.8756 ± 0.0375	0.8461 ± 0.0486	0.8224 ± 0.0555
	EAGLE	0.2356 ± 0.0196	0.3209 ± 0.1206	0.3717 ± 0.1698
	OSLOM	0.5791 ± 0.0792	-	-
8	hSCP	0.8781 ± 0.0694	0.8389 ± 0.0479	0.8240 ± 0.0460
	EAGLE	-	-	0.3374 ± 0.1672
	OSLOM	-	-	-

Table 2: Reproducibility comparison (mean±std) on HCP dataset. The rows correspond to values of k_1 and the columns correspond to values of k_2.

		10	15	20
4	hSCP	0.8838 ± 0.0495	0.7998 ± 0.0766	0.8036 ± 0.0599
	EAGLE	0.6287 ± 0.3005	0.6433 ± 0.1321	0.6046 ± 0.2981
	OSLOM	0.6780 ± 0.0537	-	-
5	hSCP	0.8785 ± 0.0675	0.8379 ± 0.0704	0.8099 ± 0.0736
	EAGLE	0.6575 ± 0.1973	0.5327 ± 0.1828	0.5426 ± 0.1656
	OSLOM	0.5867 ± 0.0869	-	-
6	hSCP	0.8655 ± 0.0404	0.8364 ± 0.0649	0.8518 ± 0.0587
	EAGLE	0.7571 ± 0.2366	0.6279 ± 0.1011	0.6244 ± 0.2627
	OSLOM	0.6391 ± 0.1266	-	-
8	hSCP	0.8670 ± 0.0559	0.8347 ± 0.0517	0.8340 ± 0.0657
	EAGLE	-	0.7451 ± 0.0319	0.5933 ± 0.2126
	OSLOM	0.5479 ± 0.0987	-	-

Table 3: Reproducibility comparison (mean±std) on PNC dataset. The rows correspond to values of k_1 and the columns correspond to values of k_2.

2.4.6 Age prediction

We compared the predictability power on the age prediction problem of the hierarchical components extracted from hSCP, EAGLE and OSLOM. Using PNC dataset, we first extracted the components and their strength ($\mathbf{\Lambda}$) for each individual. These strength values were then used to predict age of

		Correlation			
		10	15	20	25
4	hSCP	0.259 ± 0.010	0.301 ± 0.014	0.319 ± 0.012	0.377 ± 0.010
	EAGLE	0.246 ± 0.004	0.298 ± 0.009	0.300 ± 0.004	0.347 ± 0.005
	OSLOM	0.209 ± 0.007	-	-	-
5	hSCP	0.263 ± 0.010	0.327 ± 0.025	0.379 ± 0.028	0.403 ± 0.017
	EAGLE	0.259 ± 0.003	0.298 ± 0.002	0.301 ± 0.006	-
	OSLOM	0.217 ± 0.005	-	-	-
6	hSCP	0.257 ± 0.013	0.342 ± 0.027	0.381 ± 0.021	0.407 ± 0.022
	EAGLE	0.281 ± 0.004	0.308 ± 0.005	0.321 ± 0.007	
	OSLOM	0.236 ± 0.008	-	-	-
8	hSCP	0.278 ± 0.022	0.372 ± 0.026	0.382 ± 0.023	0.409 ± 0.010
	EAGLE	-	0.311 ± 0.003	0.326 ± 0.007	-
	OSLOM	0.264 ± 0.007	-	-	-

Table 4: Prediction performance comparison of hSCP, EAGLE and OSLOM

		MAE (years)			
		10	15	20	25
4	hSCP	3.20 ± 0.06	3.16 ± 0.10	3.13 ± 0.09	3.06 ± 0.14
	EAGLE	3.22 ± 0.01	3.20 ± 0.01	3.15 ± 0.01	3.10 ± 0.01
	OSLOM	3.25 ± 0.01	-	-	-
5	hSCP	3.19 ± 0.07	3.10 ± 0.17	3.06 ± 0.21	3.03 ± 0.13
	EAGLE	3.21 ± 0.01	3.13 ± 0.01	3.09 ± 0.01	-
	OSLOM	3.24 ± 0.01	-	-	-
6	hSCP	3.20 ± 0.08	3.11 ± 0.18	3.06 ± 0.16	3.02 ± 0.18
	EAGLE	3.18 ± 0.01	3.15 ± 0.01	3.14 ± 0.01	-
	OSLOM	3.20 ± 0.01	-	-	-
8	hSCP	3.18 ± 0.14	3.07 ± 0.18	3.05 ± 0.17	3.02 ± 0.13
	EAGLE	-	3.17 ± 0.01	3.13 ± 0.02	-
	OSLOM	3.20 ± 0.01	-	-	-

Table 5: Prediction performance comparison of hSCP, EAGLE and OSLOM

each individual using linear regression. Pearson correlation coefficient and mean absolute error (MAE) between the predicted brain age and the true age was used as the performance measure for comparison. Table 4 and 5 summarizes the result obtained.

To determine if our results are significantly better, the Wilcoxon signed-rank test was performed

		Correlation			
		10	15	20	25
4	EAGLE	0.0034	0.0229	3.6×10^{-4}	4.7×10^{-5}
	OSLOM	4.7×10^{-5}	-	-	-
5	EAGLE	0.0447	1.1×10^{-4}	4.7×10^{-5}	-
	OSLOM	4.7×10^{-5}	-	-	-
6	EAGLE	1	6.2×10^{-4}	4.7×10^{-5}	-
	OSLOM	1.3×10^{-4}	-	-	-
8	EAGLE	-	4.7×10^{-5}	4.7×10^{-5}	-
	OSLOM	0.0175	-	-	-

Table 6: p-value from Wilcoxon signed-rank test on Correlation

as the information about the underlying distribution in case of different performance measures is unknown. As the lower MAE is preferred, we performed a left-tailed hypothesis test when MAE is used as a performance measure. A right-tailed hypothesis test is performed when correlation is used as a performance measure because a higher value of correlation is better. Below is the null hypotheses in the two case:

1. No difference between correlation values obtained from our method compared to other methods.

2. No difference between MAE values obtained from our method compared to other methods.

Table 6 and 7 demonstrates that the prediction model built using hSCPs performed significantly better (p-value < 0.05) better than the model built using EAGLE and OSLOM components in the majority of the cases. This indicates that the hSCPs were more informative for predicting brain age. One of the reasons for the poor performance of EAGLE was that it only estimated if a region is present or not present in a component. In contrast, hSCP can determine the strength of the presence, thus had more degree of freedom resulting in better performance.

2.4.7 Clustering

An extension of the above method is presented below, which estimates hSCPs for better clustering of the data and capturing of heterogeneity. Data clustering is performed using subject specific

			MAE (years)			
		10	15	20	25	
4	EAGLE		0.0159	0.0108	0.0323	0.0351
	OSLOM	0.0012	-	-	-	
5	EAGLE	0.0447	0.1198	0.0108	-	
	OSLOM	0.0413	-	-	-	
6	EAGLE	0.9727	0.0653	0.0145	-	
	OSLOM	1.3×10^{-4}	-	-	-	0.778
8	EAGLE	-	0.0447	0.0209	-	
	OSLOM	0.0563	-	-	-	

Table 7: p-value from Wilcoxon signed-rank test on MAE

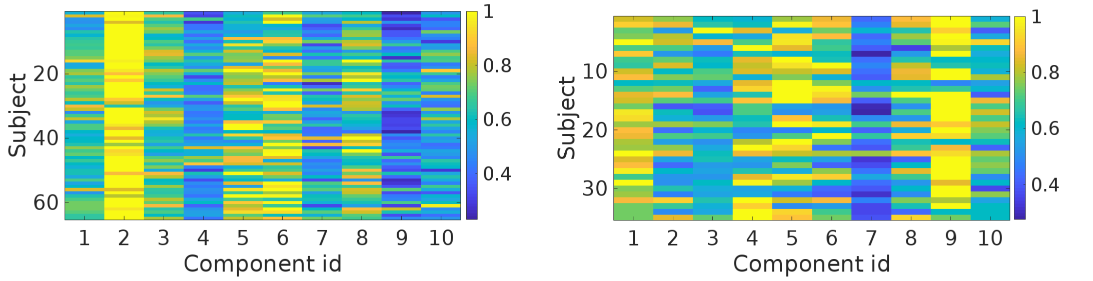
information of the components. We add a penalty term for clustering in the objective function given in problem 2.4. The modified objective function is given in problem 2.5. The joint minimization problem for estimating hSCPs and using their subject specific information for clustering is given below:

$$\begin{aligned}
\underset{\mathcal{W},\mathcal{D},\mathcal{C}}{\text{minimize}} \quad & H(\mathcal{W},\mathcal{D}) + \sum_{r=1}^{K}\sum_{l=1}^{L}\sum_{n\in\mathcal{M}_l}\sum_{d=1}^{k_r} \Big|\Big| \frac{x^n_{d,r}}{\|\mathbf{x}_{d,r}\|} - c^l_{d,r}\Big|\Big|^2 \\
\text{subject to} \quad & \|\mathbf{w}^r_l\|_1 < \lambda_r, l = 1,\ldots,k_r \quad \text{and} \quad r = 1,\ldots,K, \\
& \|\mathbf{w}^r_l\|_\infty \le 1, l = 1,\ldots,k_r \quad \text{and} \quad r = 1,\ldots,K, \\
& \mathbf{W}_j \ge 0, j = 2,\ldots,K, \\
& \mathbf{\Lambda}^n_r \succeq 0, i = 1,\ldots,S \quad \text{and} \quad r = 1,\ldots,K, \\
& \text{trace}(\mathbf{\Lambda}^n_r) = 1, i = 1,\ldots,S \quad \text{and} \quad r = 1,\ldots,K,
\end{aligned} \tag{2.5}$$

where $\mathbf{c}^l_r$ is the l^{th} cluster center at the r^{th} level, L is the number of clusters, $\mathcal{C} = \{\mathbf{c}^l_r\}_{l=1:L,r=1:K}$, $\mathbf{x}^n_r$ stores the diagonal elements of $\mathbf{\Lambda}^n_r$ and $\mathcal{M}_l$ stores the information whether n^{th} subject belongs to l^{th} cluster or not. In the above problem, $||\frac{x^n_{d,r}}{\|\mathbf{x}_{d,r}\|} - c^l_{d,r}||^2$ penalty is used for incorporating clustering by penalizing distance between points in a cluster and cluster center, and $x^n_{d,r}$ is divided by $\|\mathbf{x}_{d,r}\|$ for normalization. The above non-convex problem can be solved in a similar way as the problem 2.4 is solved in Subsection 2.3.2 using alternating minimization. Algorithm 3 provides a complete procedure for solving the problem. In the Algorithm 3, k-means($\mathbf{\Lambda}^n_r$) is used for applying k-means

Algorithm 3 hSCP-clust

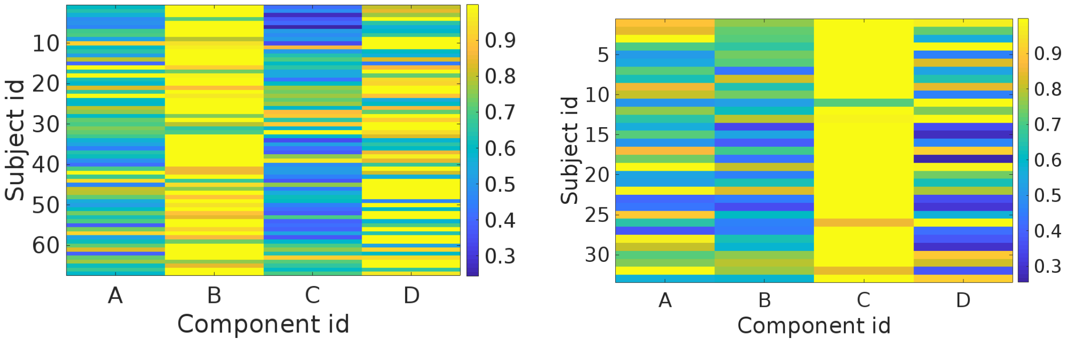
```
Input: Data $\mathcal{C}$, number of connectivity patterns $k_1,\ldots,k_K$ and sparsity $\lambda_1,\ldots,\lambda_K$ at different level
$\mathcal{W}$ and $\mathcal{D}$ = Initialization($\mathcal{C}$)
random initialization for  (uniform sampling in [0, 1])
repeat
    for $r = 1$ to $K$ do
        Step 5-9 from Algorithm 1
        for $n = 1, \ldots, N$ do
            $\mathbf{\Lambda}_r^n \leftarrow \text{descent}(\mathbf{\Lambda}_r^N)$
            $\mathbf{\Lambda}_r^n \leftarrow \text{proj}_2(\mathbf{\Lambda}_r^N)$
        $\mathcal{C}, \{\mathcal{M}_l\}_{l=1:L} \leftarrow$ k-means($\mathbf{\Lambda}_r^n$)
until Stopping criterion is reached
Output: $\mathcal{W}$, $\mathcal{D}$, $\mathcal{C}$ and $\{\mathcal{M}_l\}_{l=1:L}$
```



(a) Heterogeneity captured by subjects of cluster 1 (b) Heterogeneity captured by subjects of cluster 2

Figure 8: Heterogeneity captured by fine scale components in HCP. The color indicates the strength ($\mathbf{\Lambda}_2$) of each component present in a subject. Maximum strength of a component across subjects is fixed to be 1 for comparison purpose.

clustering (Wagstaff et al., 2001) on $\mathbf{\Lambda}_r^n$ and k-means($\mathbf{\Lambda}_r^n$) outputs $\mathcal{C}$ as cluster centers and $\{\mathcal{M}_l\}_{l=1:L}$ as cluster assignments of $\mathcal{L}$. We ran the algorithm on HCP data to extract the components and the clusters in the data. Number of clusters was selected by first extracting the hierarchical components without the penalty term and then clustering the data by using k-means on $\mathcal{L}$. L which is number of clusters was set to 2 by using the elbow method. Number of of coarse scale components was set to be 4 and and fine scale components to be 10 since they exhibited the highest reproducibility between the training and test sets. Figure 8 and Figure 9 show the distribution of fine and coarse components in two clusters. From Figure 9, we can see that component A and C are more prominent in cluster 2 compared to cluster 1, and component B and D are prominent in cluster 1 compared to cluster 2. The



(a) Heterogeneity captured by subjects of cluster 1 (b) Heterogeneity captured by subjects of cluster 2

Figure 9: Heterogeneity captured by coarse scale components in HCP. The color indicates the strength ($\mathbf{\Lambda}_1$) of each component present in a subject. Maximum strength of a component across subjects is fixed to be 1 for comparison purpose.

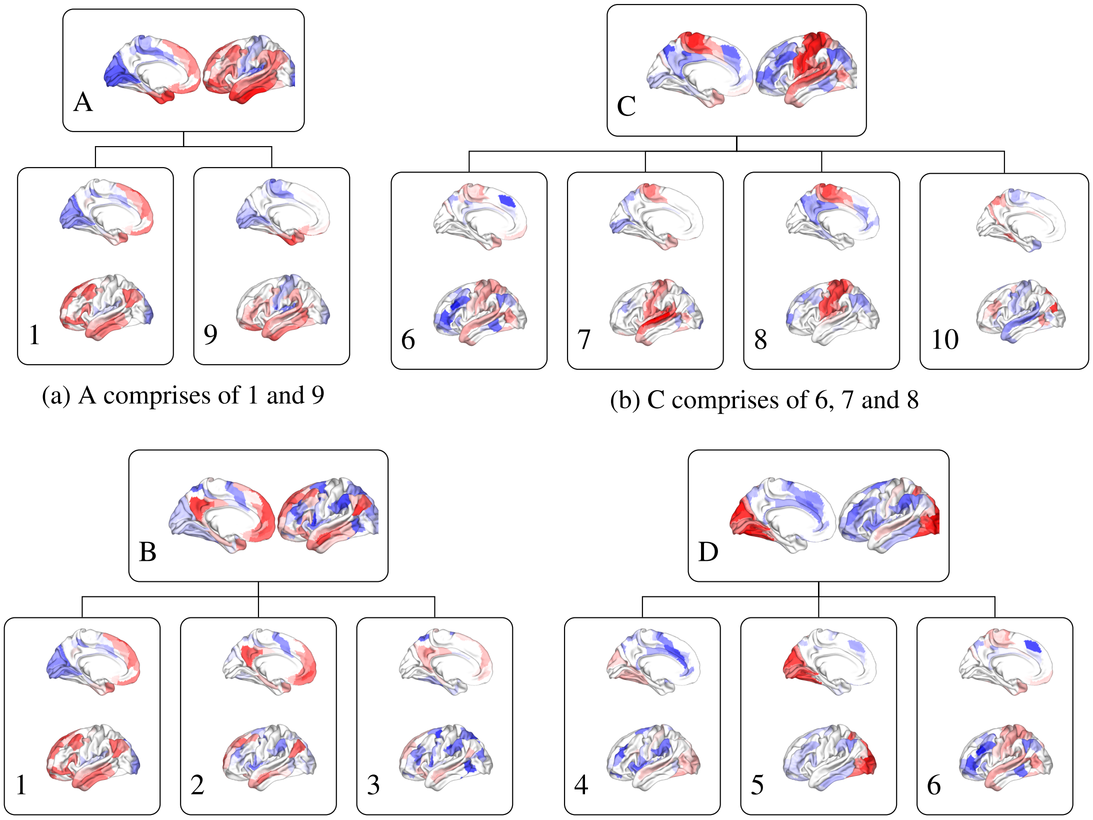
algorithm has forced Component A and its sub-components to have higher weights in one cluster. But for component B, sub-components 2 and 3 are prominent in cluster 1 and sub-component 1 is prominent in cluster 2 which can be seen in Figure 8. From the Figure 8 and 9, it can be seen that our method can reveal heterogeneity in the population by capturing the strength of components' presence in each individual.

2.4.8 Results from resting state fMRI

Figure 10 displays the 10 fine level components, the 4 coarse level components and the hierarchical structure. Nodes with red and blue color are correlated among themselves, but are anitcorrelated with each other. Note that blue color does not need to be necessarily associate with positive or negative correlation because the colors can be flipped without affecting the solution. We will be using similar figures in upcoming chapter where the blue and and red color carry the same meaning. 2 and 3 show different regions of Default Mode anti-correlated with the Dorsal Attention and Cingulo-Opercular system. 8 show different regions of default mode anti-correlated with the sensori-motor areas. 4 and 5 shows different regions of Visual system anti-correlated with Salience and fronto-parietal control systems. It can be clearly seen from Figure 10 that the fine and coarse level components are overlapping and sparse, and coarse components are comprised of a sparse linear combination of fine level components which helps in discovering the relation between different networks at different scales.

The ten fine level components obtained show the relation between different functional networks and are similar to the SCPs extracted in (Eavani et al., 2015a). From Fig. 10, it can be seen that our approach can separate task-positive regions and their associated task-negative regions into separate patterns without using traditional seed-based methods that require knowledge of a seed region of interest. Various studies have found that task-positive regions are positively correlated with each other, and task-negative regions are positively correlated with each other. The regions in the two networks are negatively correlated with each other, which aligns with our results (Raichle, 2015; Yeo et al., 2014). Component 2 covers Default Mode Network and Dorsal Attention Network, which are anti-correlated with each other. This result is a well-known finding, previously described using the seed-based correlation method (Fox et al., 2005). Anti correlations between different brain regions can represent interactions that are dependent on the state of the brain. As our method is not capturing dynamics, it has captured the interactions between different regions in different components. An example of this anti-correlation between the default mode network and the task positive network; these interactions are thought to be facilitated by indirect anatomical connections between the regions of two networks(Buckner et al., 2013). Component 8 shows different regions of DMN anti-correlated with sensorimotor, described in separate study (Karahanoğlu and Van De Ville, 2015).

Component C comprises of three connectivity patterns that involve the sensorimotor areas and its anti-correlations. Component A consists of Visual Network and Ventral Attention Network, which are anti-correlated with each other. From Fig. 10, we can see that part of sensorimotor and emotion networks (Drevets and Raichle, 1998) are anti-correlated with each other. These connections highlighted by our method are corroborated by the fact that these regions have direct anatomical pathways (Vergani et al., 2014). These negatively correlated networks can highlight different interactions in different brain regions, such as suppression, inhibition, and neurofeedback. An extension of this method that estimates dynamic components can help us understand different anti-correlations mechanism between the regions. Future research is needed to understand more about the anti-correlation and the source of these interactions.



(a) A comprises of 1 and 9

(b) C comprises of 6, 7 and 8

(c) B comprises of 1, 2 and 3

(d) D comprises of 4, 5 and 6

Figure 10: Hierarchical components derived from HCP dataset showing the connection between 10 fine scale components ($\mathbf{W}_1$) denoted from 1 to 10 and 4 coarse scale components ($\mathbf{W}_1\mathbf{W}_2$) denoted from A to D.

Our study also finds that compared to the primary sensory cortex, the higher-order association cortex has more has more associations in different components, shown in previous studies (Geranmayeh et al., 2014; Beldzik et al., 2013). Traditional seed-based approaches have been used to show that these regions have functional connectivity with more heterogeneous regions implying that they receive input and send outputs to more diverse brain regions (Katsuki and Constantinidis, 2012; Crossley et al., 2014). Thus, allowing overlapping components and positive and negative correlations within the same components provides additional insights. These features of the method facilitate storing the relation of various overlapping regions within a functional system with other areas by assigning them to different components.

Another important observation is that each of the coarse components comprises fine level components having major functional networks and their relation with other nodes. For instance, coarse component B includes majorly of 2 and 3, which stores the link between regions of Default Mode network and other nodes in the brain. Similarly, coarse component D saves the relationship between visual areas and the rest of the brain regions using 4 and 5. Thus, hSCPs can provide novel insights into the functioning of the brain by jointly uncovering both fine and coarse level components with the coarse components comprised of similarly functioning fine components.

2.5 Conclusion

In this work, we proposed a novel technique for hierarchical extraction of sparse components from correlation matrices, with application to rsfMRI data. The proposed method is a cascaded joint matrix factorization problem where a correlation matrix corresponding to each individual's data is considered an independent observation, thus allowing us to model inter-subject variability. We formulated the problem as non-convex optimization with sparsity constraints. It is important to note that as the decomposition is not by itself unique, the ability to reproducibly recover components hinges on imposing sparsity in the decomposition, which appears to provide useful and reproducible representations. We used an adaptive learning rate based gradient descent algorithm to solve the optimization problem. Compared to the implementation of SCP, which had random initialization, we used SVD initialization which made the complete algorithm both deterministic and faster.

In addition to shared patterns, we are able to extract the 'strength' of these patterns in individual components, thus capturing heterogeneity across data. Experimentally, we showed that our method is able to find sparse, low rank hierarchical decomposition using cascaded matrix factorization, which is highly reproducible across datasets. Experimental results using the PNC dataset demonstrate that the hierarchical components extracted using our model could better predict brain age compared to EAGLE and OSLOM. We also show that our model can capture heterogeneity using the HCP dataset. Our model computationally extracts a set of hierarchical components common across subjects, including resting state networks. At the same time, we capture individual information about subjects as a linear combination of these hierarchical components, making it a useful measure for group studies. Importantly, our work provides a method to uncover hierarchical organization in the functioning of the human brain.

There are several directions for future work. Firstly, it is possible to extend the idea to estimate dynamic hierarchical components similar to (Cai et al., 2017) which can help reveal how the hierarchical networks are varying over time. Secondly, generative-discriminative models can be built on the top of hSCP to find the highly discriminative components of some particular groups. For example, such a model can estimate the hierarchical components which are most discriminative of a neurodegenerative disorder (more on this in Chapter 5). Third, it would be interesting to find the guarantee on the estimation error of the hierarchical components. One possible approach is to adapt the proof techniques of (Yu et al., 2020). Finally, future studies incorporating cognitive, clinical, and genetic data, might elucidate the biological underpinning and clinical significance of the heterogeneity captured by our approach.

CHAPTER 3

Adversarial Learning for Hierarchical Patterns

In this chapter, we extend the hSCP method mentioned in the previous chapter by adding a general adversarial learning framework to make the patterns robust and achieve better reproducibility. This work provides strong evidence that of the adversarial learning framework for extraction of hSCP can improve the robustness and reproducibility of these components.

3.1 Introduction

There has been a lot of research on estimating interpretable components of functional connectivity of the brain. However, these components are often vulnerable to confounding variations, herein referred to as "adversary" using ML language, such as inter-scanner and inter-protocol variations, and rsfMRI noise or irrelevant fluctuations. This can considerably reduce these components' reproducibility and hence their utility as biomarkers of diseases that disrupt functional connectivity. To address this limitation, in this chapter, we introduce adversarial learning aiming to estimate hierarchical components that are robust to such confounding variations. The method is motivated by seminal work of Goodfellow et al. (2014), which showed that the performance of machine learning methods is vulnerable to adversarial attacks on the observed dataset. Adversarial training has been used to mitigate this and improve the generalization and robustness of the machine learning methods (Madry et al., 2018; Tramèr et al., 2018; Sinha et al., 2018; Farnia et al., 2018).

We introduce the Adversarial hSCPs (Adv. hSCPs) method to enhance the sparse component's robustness, improving its generalization performance. We formulate the problem as a bilevel matrix factorization problem and solve it using alternate minimization. Our method is based on recent advances in matrix factorization approaches that have used adversarial training to achieve state-of-the-art performances (He et al., 2018; Luo et al., 2020). In a nutshell, it is a minimax game, where the adversary perturbs model parameters to maximize or deteriorate our objective function, and in defense, we minimize the objective function. Empirical studies performed on simulations and real-world datasets demonstrate that our model can generate more reproducible components than other related methods. We also extract interpretable components using the HCP dataset.

Outline: We start by reviewing adversarial learning. Then, in Section 3.2.1, we present our method Adversarial hSCP. In Section 3.3 that follows, we compare our method against existing methods on simulated and real data.

3.1.1 Adversarial Training

A discriminative model is usually trained by minimizing the empirical expected loss over a function class $\mathcal{F} = \{f_{\mathbf{v}} : \mathbf{v} \in \mathcal{V}\}$ parameterized by $\mathbf{v}$ and parameter space $\mathbf{V}$:

$$\min_{\mathbf{v}\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^{n} l(f_{\mathbf{v}}(\mathbf{x}_i), y_i),$$

where $l(\cdot,\cdot)$ is a loss function, $f_{\mathbf{v}}$ is the output function, $\mathbf{x}_i$ is the feature vector and y_i is the label. Several recent papers (Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2017) have revealed that adding adversarial noise to the sample defined for a sample $(\mathbf{x}, y)$ as:

$$\delta_{\mathbf{v}}^{adv}(\mathbf{x}) := \arg\max_{\|\delta\|\leq\epsilon} l(f_{\mathbf{v}}(\mathbf{x}+\delta), y), \tag{3.1}$$

where $\epsilon > 0$ is the adversarial noise power can drastically decrease model's performance. Adversarial training (Madry et al., 2018) was introduced to provide robustness against the adversaries defined above. The training involves empirical risk minimization over the perturbed samples by solving

$$\min_{\mathbf{v}\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^{n} l(f_{\mathbf{v}}(x_i + \delta_{\mathbf{v}}^{adv}(\mathbf{x})), y_i).$$

The above formulation has a drawback that the accuracy drops drastically. To overcome this problem, Mix-minibatch adversarial training (MAT) (Wong and Kolter, 2018) is performed by solving

$$\min_{\mathbf{v}\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^{n} l(f_{\mathbf{v}}(x_i + \delta_{\mathbf{v}}^{adv}(\mathbf{x})), y_i) + l(f_{\mathbf{v}}(x_i), y_i), \tag{3.2}$$

which balances between accuracy on the clean examples and robustness on the adversarial examples. Motivated by the above methodology, we build adversarial training for learning sparse hierarchical connectivity components. As the above training regime was supervised, we use a different formulation with the same idea for the unsupervised hSCP model.

3.2 Adversarial hierarchical Sparse Connectivity Patterns

3.2.1 Recap

Let there be N number of subjects or participants, and each subject's BOLD fMRI time series has T time points and P nodes representing regions of interest. The input to hSCP are correlation matrices $\mathbf{\Theta}^n \in \mathbb{S}_{++}^{P\times P}$ where ith and jth element of the matrix is the correlation between time series of ith and jth node. hSCP then outputs a set of shared hierarchical patterns following the below equations:

$$\mathbf{\Theta}^n \approx \mathbf{W}_1\mathbf{\Lambda}_1^n\mathbf{W}_1^\top, \quad \ldots \quad \mathbf{\Theta}^n \approx \mathbf{W}_1\mathbf{W}_2\ldots\mathbf{W}_K\mathbf{\Lambda}_K^n\mathbf{W}_K^\top\mathbf{W}_{K-1}^\top\ldots\mathbf{W}_1^\top,$$

where $\mathbf{\Lambda}_k^n$ is a diagonal matrix having positive elements storing relative contribution of the components for the nth subject at kth level, K is the depth of hierarchy and $P > k_1 > \ldots > k_K$.

3.2.2 Adversarial Learning for hSCP

Adversarial learning has shown to achieve state-of-the-art performance of various matrix factorization approaches (He et al., 2018; Luo et al., 2020). The method is based on perturbation of input data $\mathbf{\Theta}^n$ to learn stable components robust to adversaries. First the input data is perturbed to generate new data $\mathbf{\Gamma}^n = \mathbf{\Theta}^n + \sigma\mathbf{J}_P$ where σ is the standard deviation of the data. We consider this perturbation as rank one perturbation which will transform eigenvalues of each subject's correlation matrix differently. Since the addition of ones does not change the symmetric property of the matrix, we just scale the matrix such that the diagonal matrix contains 1 and by Weyl's inequality about perturbation (Stewart, 1998) it will be positive definite. We have experimented with randomly positively scaled rank one perturbation, but the results were worse than the standard hSCP. Also, using rank one perturbation, it is easier to control eigenvalues (and noise added) of the modified matrix than using rank k perturbation. The perturbed set of components $\tilde{\mathbf{W}}_1$ are then estimated using the new data and are ensured to be close to $\mathbf{W}_1$ by solving the below minimization problem: In this section, we demonstrate how to incorporate adversarial learning at one level for the hSCP method, which then can be extended to multiple levels. The idea is to perturb input data $\mathbf{\Theta}^i$ and learn stable components $\mathbf{W}_1$ which are robust to adversaries, such as inter-scanner differences and unwanted rsfMRI noise. There are two parts of the complete learning procedure-

Attack. We first manually perturb the input data to get perturbed data $\mathbf{\Gamma}^n = \mathbf{\Theta}^n + 0.1\sigma\mathbf{1}_P$ where σ is the standard deviation of the data and use it to learn a perturbed weight matrix $\tilde{\mathbf{W}}_1$:

$$\tilde{\mathbf{W}}_1 = \underset{\hat{\mathbf{W}}_1}{\arg\min}\, \alpha\|\hat{\mathbf{W}}_1 - \mathbf{W}_1\|_F^2 + \sum_{n=1}^{N} \|\mathbf{\Gamma}^n - \hat{\mathbf{W}}_1\mathbf{\Lambda}^n\hat{\mathbf{W}}_1^\top\|_F^2. \tag{3.3}$$

In the above equation, the first part is used to estimate $\tilde{\mathbf{W}}_1$, which is close to $\mathbf{W}_1$ in Frobenius norm, thus mimicking the actual components but is learned from the noise-induced data. The second term is used for learning $\tilde{\mathbf{W}}_1$ using a perturbed data matrix. The main goal of the attacker is to learn $\tilde{\mathbf{W}}_1$ for a given $\mathbf{\Gamma}^i$ and fool the model by forcing the model to learn $\mathbf{\Lambda}^i$ from the perturbed data. Our framework does not depend on the perturbations' assumptions; the type of perturbation can be varied depending on different types of practical noises such as site-induced or scanner-induced noise. Finding optimal perturbation is left for future work.

Defense. Aim of the learner is to estimate $\mathbf{\Lambda}^n$ and $\mathbf{W}_1$ by minimizing the below cost function:

$$D(\mathbf{W}, \mathbf{\Lambda}) = \sum_{n=1}^{N} \|\mathbf{\Theta}^n - \tilde{\mathbf{W}}_1\mathbf{\Lambda}^n\tilde{\mathbf{W}}_1^\top\|_F^2 + \beta\sum_{n=1}^{N} \|\mathbf{\Theta}^n - \mathbf{W}_1\mathbf{\Lambda}^n\mathbf{W}_1^\top\|_F^2, \tag{3.4}$$

for a fixed $\tilde{\mathbf{W}}_1$. Learner first estimates subject specific information $\mathbf{\Lambda}^n$ using perturbed weight matrix and then use it to learn $\mathbf{W}_1$. We can now define the optimization problem for the complete adversarial learning at single level using equation 3.3 and 3.4 as:

$$\begin{aligned} &\underset{\mathbf{W}_1, \mathbf{\Lambda}^n\ \forall n}{\text{minimize}} && \sum_{n=1}^{N} \|\mathbf{\Theta}^n - \tilde{\mathbf{W}}_1 \mathbf{\Lambda}^n \tilde{\mathbf{W}}_1^\top\|_F^2 + \beta \sum_{n=1}^{N} \|\mathbf{\Theta}^n - \mathbf{W}_1 \mathbf{\Lambda}^n \mathbf{W}_1^\top\|_F^2 \\ &\text{subject to} && \tilde{\mathbf{W}}_1 = \underset{\hat{\mathbf{W}}_1}{\arg\min}\, \alpha \|\hat{\mathbf{W}}_1 - \mathbf{W}_1\|_F^2 + \sum_{n=1}^{N} \|\mathbf{\Gamma}^n - \hat{\mathbf{W}}_1 \mathbf{\Lambda}^n \hat{\mathbf{W}}_1^\top\|_F^2. \end{aligned} \tag{3.5}$$

The above equation is analogous to discriminate adversarial learning problem defined in equation 3.2. Let $\mathcal{W} = \{\mathbf{W}_r \mid r = 1, \ldots, K\}$, $\mathcal{C} = \{\mathbf{\Theta}^n \mid n = 1, \ldots, N\}$, $\tilde{\mathcal{W}} = \{\tilde{\mathbf{W}}_r \mid r = 1, \ldots, K\}$, $\mathcal{D} = \{\mathbf{\Lambda}_r^n \mid r = 1, \ldots, K; n = 1, \ldots, N\}$, $\hat{\mathcal{W}} = \{\hat{\mathbf{W}}_r \mid r = 1, \ldots, K\}$, $\mathcal{P} = \{\mathbf{\Gamma}^n \mid n = 1, \ldots, N\}$ and

$$H(\mathcal{W}, \mathcal{D}, \mathcal{C}) = \sum_{n=1}^{N} \sum_{r=1}^{K} \|\mathbf{\Theta}^n - (\prod_{j=1}^{r} \mathbf{W}_j) \mathbf{\Lambda}_r^n (\prod_{j=1}^{r} \mathbf{W}_j)^\top\|_F^2.$$

Then the multi level formulation of can be written as:

$$\begin{aligned} &\underset{\mathcal{W}, \mathcal{D}}{\text{minimize}} && J(\tilde{\mathcal{W}}, \mathcal{W}, \mathcal{D}, \mathcal{C}) = H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{C}) + \beta H(\mathcal{W}, \mathcal{D}, \mathcal{C}) \\ &\text{subject to} && \tilde{\mathbf{W}}_r = \underset{\hat{\mathbf{W}}_r}{\arg\min}\, \alpha \|\hat{\mathbf{W}}_r - \mathbf{W}_r\|_F^2 + H(\hat{\mathcal{W}}, \mathcal{D}, \mathcal{P}). \end{aligned} \tag{3.6}$$

3.2.3 Optimization

The complete algorithm to solve the above optimization problem is described in Algorithm 4 (Adv. hSCP). First, the adversarial perturbations are generated by performing gradient descent on $\hat{\mathbf{W}}_r$, and then the model parameters are updated using gradient descent. This process is repeated until the convergence criteria is reached. $\mathrm{descent}$ is the update rules defined by AMSgrad (Reddi et al., 2019) for performing gradient descent. Gradients are defined in the next section. $\mathrm{proj}_1(\mathbf{A})$ (Podosinnikova et al., 2013) function projects each column of $\mathbf{A}$ into intersection of L_1 and L_∞ ball and $\mathrm{proj}_2(\mathbf{A})$ function makes all the negative elements of $\mathbf{A}$ equal to zero. The model parameters are initialized by first optimizing the hSCP model using $\mathrm{svd} - \mathrm{initialization}$ algorithm [Alogirthm 2] in Subsection

Algorithm 4 Adv. hSCP

```
Input: Data $\mathcal{C}$, perturbed data $\mathcal{P}$; $\mathcal{W}$ and $\mathcal{D}$ = hSCP($\mathcal{C}$)
repeat
    for $r = 1$ to $K$ do
        Update adversarial perturbations
        $\hat{\mathbf{W}}_r \leftarrow \mathrm{descent}(\hat{\mathbf{W}}_r)$
        Update model parameters
        $\mathbf{W}_r \leftarrow \mathrm{descent}(\mathbf{W}_r)$
        if $r == 1$ then
            $\mathbf{W}_r \leftarrow \mathrm{proj}_1(\mathbf{W}_r)$
        else
            $\mathbf{W}_r \leftarrow \mathrm{proj}_2(\mathbf{W}_r)$
        $\mathbf{\Lambda}_r^n \leftarrow \mathrm{descent}(\mathbf{\Lambda}_r^n)$; $\mathbf{\Lambda}_r^n \leftarrow \mathrm{proj}_2(\mathbf{\Lambda}_r^n) \quad n = 1, \dots, N$
until Stopping criterion is reached
Output: $\mathbf{W}$ and $\mathbf{\Lambda}$
```

2.3.5, rather than randomly initialized. This makes algorithm deterministic, and the algorithm can start from an optimal point on which adversarial learning can improve if there is overfitting.

3.2.4 Gradients

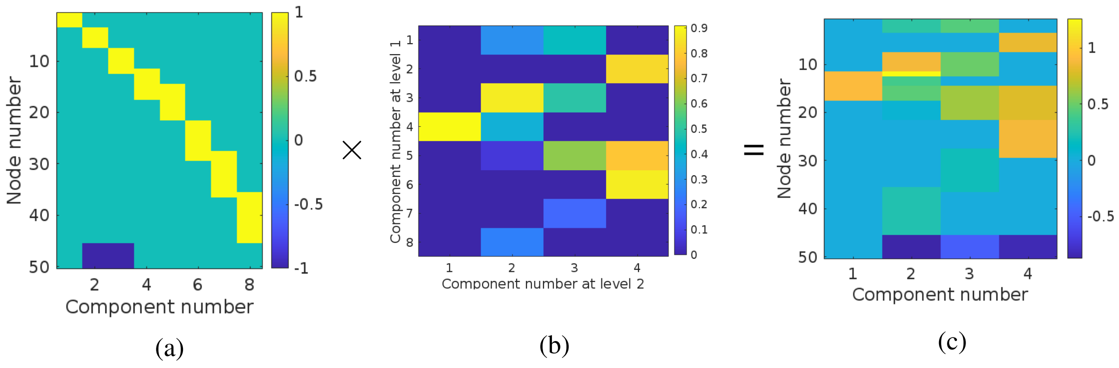
In this section, we define gradients used for alternating gradient descent. Let

$$\tilde{\mathbf{W}}_0 = \mathbf{W}_0 = \mathbf{I}_P, \qquad \mathbf{Y}_r = \prod_{j=0}^{r} \mathbf{W}_j, \qquad \tilde{\mathbf{Y}}_r = \prod_{j=0}^{r} \tilde{\mathbf{W}}_j,$$

$$\mathbf{T}_{m,n}^r = (\prod_{j=1}^{m-r} \mathbf{W}_j)\mathbf{\Lambda}_{m-r}^n(\prod_{j=1}^{m-r} \mathbf{W}_j)^\top, \qquad \tilde{\mathbf{T}}_{m,n}^r = (\prod_{j=1}^{m-r} \tilde{\mathbf{W}}_j)\mathbf{\Lambda}_{m-r}^n(\prod_{j=1}^{m-r} \tilde{\mathbf{W}}_j)^\top.$$

We first define gradient for updating adversarial perturbations $\tilde{\mathbf{W}}_\mathbf{r}$. The objective function is $F = \alpha\|\hat{\mathbf{W}}_r - \mathbf{W}_r\|_F^2 + H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{P})$ and gradient with respect to $\tilde{\mathbf{W}}_\mathbf{r}$ will be

$$\begin{aligned}\frac{F}{\partial \tilde{\mathbf{W}}_\mathbf{r}} &= 2\alpha(\hat{\mathbf{W}}_r - \mathbf{W}_r) + \frac{\partial H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{C})}{\partial \tilde{\mathbf{W}}_r} \\ &= 2\alpha(\hat{\mathbf{W}}_r - \mathbf{W}_r) + \sum_{n=1}^{N}\sum_{j=r}^{K}\big(-4\tilde{\mathbf{Y}}_{r-1}^\top \mathbf{\Gamma}^n \tilde{\mathbf{Y}}_{r-1}\tilde{\mathbf{W}}_r\tilde{\mathbf{T}}_{j,n}^r \\ &+ 4\tilde{\mathbf{Y}}_{r-1}^\top \tilde{\mathbf{Y}}_{r-1}\tilde{\mathbf{W}}_r\tilde{\mathbf{T}}_{j,n}^r\tilde{\mathbf{W}}_r^\top\tilde{\mathbf{Y}}_{r-1}^\top\tilde{\mathbf{Y}}_{r-1}\tilde{\mathbf{W}}_r\tilde{\mathbf{T}}_{j,n}^r\big).\end{aligned}$$



(a) (b) (c)

Figure 11: (a) Visualization of ground truth components at level 1. (b) Weight matrix used to generate components at level 2. (c) Visualization of ground truth components at level 2.

We now define gradients for updating model parameters. The gradient of objective function J with respect to $\mathbf{\Lambda}_r^n$ is:

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{\Lambda}_r^n} &= \frac{\partial H(\tilde{\mathcal{W}},\mathcal{D},\mathcal{C})}{\partial \mathbf{\Lambda}_r^n} + \beta \frac{\partial H(\mathcal{W},\mathcal{D},\mathcal{C})}{\partial \mathbf{\Lambda}_r^n} \\
&= \Big[(-2\tilde{\mathbf{Y}}_r^\top \mathbf{X}_r^n \tilde{\mathbf{Y}}_r + 2\tilde{\mathbf{Y}}_r^\top \tilde{\mathbf{Y}}_r \mathbf{\Lambda}_r^n \tilde{\mathbf{Y}}_r^\top \tilde{\mathbf{Y}}_r) + \beta(-2\mathbf{Y}_r^\top \mathbf{X}_r^i \mathbf{Y}_r + 2\mathbf{Y}_r^\top \mathbf{Y}_r \mathbf{\Lambda}_r^i \mathbf{Y}_r^\top \mathbf{Y}_r)\Big] \circ \mathbf{I}_{k_r}.
\end{aligned}
$$

The gradient of J with respect to $\mathbf{W}_r$ is:

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{W}_r} &= \frac{\partial H(\mathcal{W},\mathcal{D},\mathcal{C})}{\partial \mathbf{W}_r} \\
&= \sum_{n=1}^{N} \sum_{j=r}^{K} -4\mathbf{Y}_{r-1}^\top \mathbf{X}_n \mathbf{Y}_{r-1} \mathbf{W}_r \mathbf{T}_{j,n}^r + 4\mathbf{Y}_{r-1}^\top \mathbf{Y}_{r-1} \mathbf{W}_r \mathbf{T}_{j,n}^r \mathbf{W}_r^\top \mathbf{Y}_{r-1}^\top \mathbf{Y}_{r-1} \mathbf{W}_r \mathbf{T}_{j,n}^r.
\end{aligned}
$$

3.3 Experiments

3.3.1 Simulated dataset

We first use a simulated dataset to evaluate the performance of our model against SCP (Eavani et al., 2015a), NMF (Potluru and Calhoun, 2008), adv. NMF (Luo et al., 2020) and ICA (Smith et al., 2009). The values of α and β are set to be 10^{-3} and 0.5 respectively throughout this chapter. We generate sparse components $\mathbf{S}_1 \in \mathbb{R}^{P \times k_1}$ with $P = 50$ and $k_1 = 8$ and generate network structure from it.

This network is then used as input to NetSim (Smith et al., 2011) with TR equal to 3 seconds to generate time-series data of 100 subjects, each having 300 time-points. NetSim also adds Gaussian noise to the time series of each node. We also add Poisson noise with a mean equal to 0.4 to check how different methods perform in a high noise scenario.

We compare components/factors derived from all the models with $k_1 \in \{6, 8, 10, 12\}$. Accuracy is used as a performance measure defined as a normalized inner product between ground truth components and estimated factors derived from various algorithms. The optimal sparsity parameter λ_1 in the hSCP is selected from $P \times 10^{[-2:1]}$ having the highest average split-sample reproducibility in 20 runs. Split-sample reproducibility of components is computed by randomly dividing the data into two equal parts and then calculating normalized inner product between components extracted from each sample. A high reproducibility value implies that the same component can be extracted from multiple samples. Table 8 shows the comparison of the accuracy of different methods averaged over 20 runs. As the hSCP method is deterministic, the output remains the same in every run. From the table, it can be seen that adversarial training can significantly improve the accuracy of hSCP. An important thing to note here is that adversarial training has also improved the accuracy of NMF, but it remains less than that of hSCP.

We next generated a two-level hierarchy using the components defined above as the first layer. We used linear operator for projection to lower dimensional space to get coarse components with $P = 50$ and $k_2 = 4$. Visualization of the components is in Figure 11. Time-series data were then generated under the same settings presented above. Table 9 shows the components' average accuracy at two-level over 20 runs for hSCP and Adv. hSCP with Adv. hSCP method giving better results. We did not use ICA and NMF as they can only generate components at only one level.

Resting state fMRI data

We used 100 unrelated subjects released within the 900 subjects data release from the publicly available Human Connectome Project (HCP) (Van Essen et al., 2012) dataset for comparing different methods. ICA+FIX pipeline (Glasser et al., 2013) is used to process the complete data. Each subject has 4 scans, with each scan comprising 1001 time points and 360 nodes.

Method	$k_1 = 6$	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$
hSCP	0.801	0.829	0.818	0.814
Adv. hSCP	$\mathbf{0.832}$	$\mathbf{0.847}$	$\mathbf{0.867}$	$\mathbf{0.864}$
ICA	0.656 ± 0.004	0.696 ± 0.022	0.734 ± 0.025	0.748 ± 0.011
NMF	0.650 ± 0.104	0.701 ± 0.071	0.708 ± 0.118	0.712 ± 0.085
Adv. NMF	0.695 ± 0.047	0.718 ± 0.069	0.720 ± 0.091	0.723 ± 0.114

Table 8: Accuracy on simulated dataset

		$k_1 = 6$	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$
$k_2 = 4$	hSCP	0.821	0.837	0.827	0.819
	Adv. hSCP	$\mathbf{0.859}$	$\mathbf{0.864}$	$\mathbf{0.846}$	$\mathbf{0.823}$
$k_2 = 6$	hSCP	0.816	0.819	0.813	0.805
	Adv hSCP	$\mathbf{0.848}$	$\mathbf{0.849}$	$\mathbf{0.826}$	$\mathbf{0.814}$

Table 9: Accuracy on simulated data with two level hierarchy

Method	$k_1 = 6$	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$
hSCP	0.798	$\mathbf{0.779}$	0.739	0.724
Adv. hSCP	$\mathbf{0.804}$	0.771	$\mathbf{0.818}$	$\mathbf{0.843}$
ICA	0.637 ± 0.015	0.671 ± 0.034	0.715 ± 0.027	0.738 ± 0.012
NMF	0.640 ± 0.101	0.655 ± 0.109	0.703 ± 0.079	0.704 ± 0.128
Adv. NMF	0.690 ± 0.079	0.681 ± 0.071	0.694 ± 0.080	0.682 ± 0.088

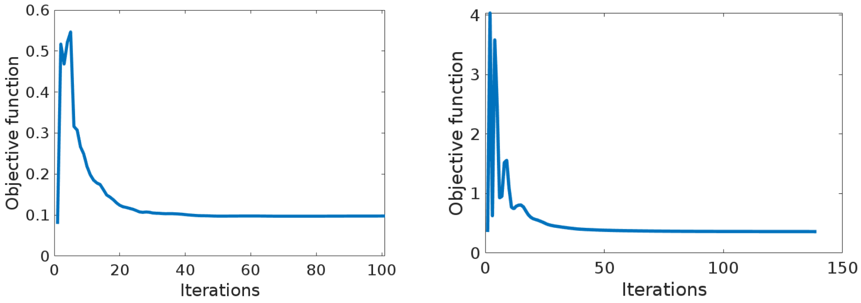
Table 10: Accuracy on simulated dataset with Pois(0.4) noise added

3.3.2 Convergence Analysis

We empirically validate the convergence using the reconstruction error:

$$\frac{\sum_{n=1}^{N}\sum_{r=1}^{K}||\mathbf{\Theta}^n - (\prod_{j=1}^{r}\mathbf{W}_j)\mathbf{\Lambda}_r^n(\prod_{j=1}^{r}\mathbf{W}_j)^T||_F^2}{\sum_{n=1}^{N}\sum_{r=1}^{K}||\mathbf{\Theta}^n||_F^2}.$$

In the figure 12, it can be seen that initially, the function value is low because the initial value is an optimal value of $\mathbf{W}$ and $\mathbf{\Lambda}$ returned using the hSCP algorithm. As the adversarial attack begins, the objective function value starts to fluctuate because of the minimax game, where the adversarial perturbation tries to deviate the result from the optimal value. In defense, we try to minimize the objective function. The algorithm converges when the optimal value becomes robust to perturbations.



(a) $k = 10$ (b) $k = 20$

Figure 12: Convergence of Adv. hSCP algorithm using HCP dataset for different values of k.

Method	10	15	20	25
hSCP	0.749 ± 0.045	0.750 ± 0.046	0.712 ± 0.026	0.701 ± 0.019
Adv. hSCP	$\mathbf{0.787 \pm 0.052}$	$\mathbf{0.765 \pm 0.059}$	$\mathbf{0.716 \pm 0.020}$	$\mathbf{0.721 \pm 0.016}$
ICA	0.695 ± 0.067	0.638 ± 0.046	0.581 ± 0.039	0.523 ± 0.027
NMF	0.689 ± 0.038	0.657 ± 0.067	0.635 ± 0.053	0.629 ± 0.020
Adv. NMF	0.709 ± 0.073	0.659 ± 0.043	0.653 ± 0.026	0.633 ± 0.032

Table 11: Reproducibility on HCP dataset

3.3.3 Results from rsfMRI data

We compare components/factors derived from all the models with $k_1 \in \{5, 10, 15, 20\}$. As the ground truth is not known, we use split-sample reproducibility as a performance measure. We first find the optimal value of λ_1 from $P \times 10^{[-2:1]}$ by dividing the data into three equal parts: training, validation, and test data, and choosing the parameters corresponding to maximum mean reproducibility over 20 runs on training and validation set. After the selecting the optimal parameter, we compare results using training and test. Table 11 shows reproducibility calculated from training and test data. It can be seen that the hSCP method can extract components with high reproducibility. We have similar results presented in Table 12 for a two-level decomposition.

We extract 10 components at level 1, and 4 components at level 2 using Adv. hSCP learning from the HCP dataset. Figure 13 shows two hierarchical components. Component 1 stores anti-correlation information between Default Mode Network and Dorsal Attention Network previously studied using seed-based correlation method (Fox et al., 2005). Component 2 stores anti-correlation between

		$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
$k_2 = 4$	hSCP	0.872 ± 0.044	0.853 ± 0.064	0.831 ± 0.075	0.826 ± 0.091
	Adv. hSCP	$\mathbf{0.895 \pm 0.030}$	$\mathbf{0.866 \pm 0.029}$	$\mathbf{0.848 \pm 0.056}$	$\mathbf{0.830 \pm 0.061}$
$k_2 = 6$	hSCP	0.856 ± 0.070	0.842 ± 0.062	0.828 ± 0.031	0.824 ± 0.035
	Adv hSCP	$\mathbf{0.877 \pm 0.076}$	$\mathbf{0.864 \pm 0.067}$	$\mathbf{0.843 \pm 0.045}$	$\mathbf{0.834 \pm 0.048}$

Table 12: Reproducibility on HCP dataset with two level hierarchy

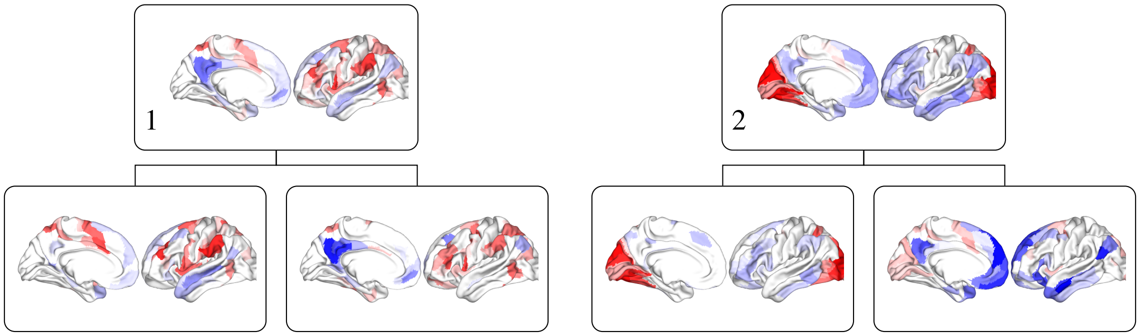


Figure 13: Hierarchical components estimated using Adv. hSCP. Red and blue color are used for showing negative correlations between regions in a component.

Default Mode Network and extrastriate visual areas, which is another well-known finding (Uddin et al., 2009). A more thorough discussion is needed for examining the differences and similarities between the components derived from hSCP and Adv. hSCP, which we have left for future work.

3.4 Conclusion

In this chapter, we used adversarial learning to enhance the hSCP method by increasing the hierarchical components' reproducibility. We formulate the problem as a bilevel optimization problem and used adaptive gradient descent to solve it. Experimental results based on simulated data show that Adv hSCP can extract components accurately compared to other methods. Results using real-world rsfMRI data demonstrate the adversarial learning can improve the reproducibility of the components. We also discuss the interpretability of the components extracted from the HCP dataset.

There are several applications of this work. Improved reproducibility of the components can increase accuracy and confidence when applied to clinical applications such as age prediction, disease diagnosis, etc. Adversarial learning can be extended to other matrix factorization approaches used for the analysis of fMRI data, such as dynamic sparse connectivity patterns (Cai et al., 2017), sparse granger causality patterns (Sahoo et al., 2018), deep non-negative matrix factorization (Li et al., 2018a), etc. It would be interesting to assess the impact of the method in characterizing activity in terms of task-induced activations.

CHAPTER 4

Robust Hierarchical Patterns in Multi-Site fMRI Studies

In the previous chapter, we saw an adversarial learning-based framework to improve the robustness of the factors without knowing the type of noise in the dataset. In this chapter, we build an adversarial learning-based model to reduce the variance introduced by pooling datasets from multiple sites and use the previous chapter's algorithm to estimate cleaner and robust patterns.

4.1 Introduction

Multi-site fMRI studies have gained a lot of interest over the last decade (Noble et al., 2017; Di Martino et al., 2014). One reason for this is the necessity to evaluate a hypothesis in multiple settings/sites and make the hypothesis result generalizable to a diverse population. Also, the pooling of data is essential when studying rare disorders or neurological conditions where the aim is to generalize the results to diverse populations (Dansereau et al., 2017; Keshavan et al., 2016). However, the data pooling often results in the introduction of non-biological systematic variance due to differences in scanner hardware and imaging acquisition parameters (Shinohara et al., 2017). This additional variability can lead to spurious results and a decrease in statistical power. The variability can also hinder in the estimation of true biological changes or in inferring non-biological differences as biological because of the correlation between site effects and biological predictors. Many studies working with multi-site data fMRI have reported considerable variability due site or scanner effects (Abraham et al., 2017; Jovicich et al., 2016; Noble et al., 2017).

The non-biological variability introduced due to inter-scanner and inter-protocol variations can affect the estimation of the common features derived from fMRI (Yu et al., 2018), such as functional connectivity (Shinohara et al., 2017) or sparse hierarchical factors (this work). These features were used for the study of the brain's function during aging (Raichle, 2015), of various neurological disorders (Fornito et al., 2016; Stam, 2014), and tasks (Cook et al., 2007). The non-biological variability can considerably reduce these features' reproducibility across different datasets and hence their utility as biomarkers for diseases that disrupt functional connectivity. Thus the removal of non-biological variance introduced by pooling of the data is essential for many neuroimaging studies. In this chapter, we focus on robust estimation of hSCPs in a multi-site regime.

One of the first investigations of batch effects in rs-fMRI was performed by Olivetti et al. (2012) using extremely randomized trees along with dissimilarity representation. One of the common methods to remove site effects is the harmonization of data. Harmonization of fMRI data especially derived measures, is very nascent, even though it is much needed with the growing number of multi-site data sets (Adhikari et al., 2019). Many existing methods to reduce site effects are based on an empirical Bayes method ComBat (Johnson et al., 2007), which was developed to remove 'batch effects' in genetics and has been applied for harmonizing different measures derived from structural (Pomponio et al., 2020; Fortin et al., 2017) and functional MRI (Yu et al., 2018). However, ComBat and its variants such as ComBat-GAM (Pomponio et al., 2020) can not be directly applied to connectivity matrices since it can destroy the structure of the connectivity matrix and semi definiteness of the connectivity matrix (more details in Section 4.2). A similar difficulty arises when applying ComBat based harmonization to other structured data. Another approach is not to remove site effects, but directly use site information for downstream analysis such as age prediction, finding associations with various clinical variables, etc. Kia et al. (2020); Bayer et al. (2021) used normative modeling for the age prediction task while keeping the site as one of the predictors. One limitation of the method is that without removing the site effects, the biomarkers can not be used for downstream analysis by the clinician, psychiatrist, etc., directly, which is one of the goals of the hSCP.

Recent work by Vega and Greiner (2018) analyzed the impact of covariate analysis, z-score normalization, and whitening on batch effects. Domain adaption has also been introduced in removing batch effects in rs-fMRI data. Domain adaption techniques aim to learn from multiple sources and generalize the model to perform well on a new related target site. Extensive work has been done on unsupervised domain adaptation approaches (Gholami et al., 2020; Zhao et al., 2019). Several methods have been introduced for domain adaptation such as Multi-source Domain Adversarial Networks (Zhao et al., 2018), Multi-Domain Matching Networks (Li et al., 2018b), Moment Matching (Peng et al., 2019), etc. Readers can refer to the detailed survey by Zhao et al. (2020b). In multi-site fMRI data, Wang et al. (2019) introduced a low-rank domain to remove batch effects. Other recent approached include transport-based joint distribution alignment (Zhang et al., 2020a) and federated learning (Li et al., 2020b) for fMRI data.

We develop a new model that is robust to site-effects in the estimation of sparse hierarchical connectivity pattern components (rshSCP). For this, the method learns site-specific features and global space, storing the information about the scanner and site, and uses these features to reduced site effects in the components. We also use adversarial learning approach defined in the previous chapter on top of our method to improve the reproducibility and generalizability of the components across components from the same site. We formulate the method as a non-convex optimization problem which is solved using stochastic gradient descent. Experiments on simulated and real datasets show that our method can improve the split-sample and leave one site reproducibility of the components while retaining age-related biological variability in the data, thus capturing informative heterogeneity.

Outline: We start by reviewing hierarchial Sparse Connectivity Patterns (hSCPs) and adversarial learning. Then, in Section 4.2, we present our method, extracting interpretable hSCPs which are robust to site effects. In Section 4.4.1, we demonstrate the effectiveness of the method on simulated. In Section 4.4.2, we show using a large multi-site dataset that the proposed method can extract more reproducible and cleaner patterns compared to the standard approach.

4.2 Method

4.2.1 Recap

Let there be N number of subjects or participants, and each subject's BOLD fMRI time series has T time points and P nodes representing regions of interest. The input to hSCP are correlation matrices $\mathbf{\Theta}^n \in \mathbb{S}_{++}^{P \times P}$ where ith and jth element of the matrix is the correlation between time series of ith and

jth node. hSCP then outputs a set of shared hierarchical patterns following the below equations:

$$\mathbf{\Theta}^n \approx \mathbf{W}_1 \mathbf{\Lambda}_1^n \mathbf{W}_1^\top, \quad \ldots \quad \mathbf{\Theta}^n \approx \mathbf{W}_1 \mathbf{W}_2 \ldots \mathbf{W}_K \mathbf{\Lambda}_K^n \mathbf{W}_K^\top \mathbf{W}_{K-1}^\top \ldots \mathbf{W}_1^\top,$$

where $\mathbf{\Lambda}_k^n$ is a diagonal matrix having positive elements storing relative contribution of the components for the nth subject at kth level, K is the depth of hierarchy and $P > k_1 > \ldots > k_K$.

4.2.2 Can we use standard harmonization approaches?

These methods reduce site effects by adjusting for additive and multiplicative effects for each feature in data separately and use emperical Bayes estimates the model parameters. These methods can be used in the case of hSCP in two ways. First, harmonization can be directly applied to each element of the correlation matrices, which is the input of hSCP. This will reduce site effects from each element of the correlation matrix, thus from the complete input, but the final matrix that does not necessarily follow the essential property of a correlation matrix i.e., positive definiteness. For similar reasons, COMBAT can not be directly applied to time series; if applied, it can change the inference derived from the correlation matrix. Second, harmonization can be directly applied to $\mathbf{\Lambda}$ to remove site effects. To understand this, we look at the hSCP formulation at one level:

$$\mathbf{\Theta}^n \approx \sum_{l=1}^{k} d_l^n \mathbf{w}_l \mathbf{w}_l^\top \approx \mathbf{W} \mathbf{\Lambda}^n \mathbf{W}^\top,$$

where d_l^n are non-zero elements storing the subject-specific information, which can be affected by the variability introduced by the site. In this model, harmonizing each feature across different sites will change the relative contribution of the components in each subject's functional structure, which is not desirable. Instead, a two step optimization procedure can be used to incorporate ComBat with hSCP (ComBat hSCP). We first run hSCP and use ComBat on the extracted $\mathbf{\Lambda}^n$ to get harmonized subject specific information $\mathbf{\Delta}^n \in \mathbb{R}^{k_1 \times k_1}$ for each subject. We then re-fitted $\mathbf{W}$ using the below decomposition-

$$\mathbf{\Theta}^n \approx \mathbf{W}(\mathbf{\Delta}^n + \mathbf{S})\mathbf{W}^\top.$$

We added a diagonal shift matrix $\mathbf{S} \in \mathbb{R}^{k_1 \times k_1}$ such that $\mathbf{\Delta}^n + \mathbf{S}$ is positive for each subject and performed the optimization to estimate $\mathbf{W}$ and $\mathbf{S}$. We show through experiments that this baseline two step optimization procedure is not optimal and performs worse than standard hSCP.

4.2.3 Robust to site hSCP

Estimating hSCPs in multi-site data can introduce non-biological variances in the components and the subject-specific information. One of the typical approaches would be to use harmonization methods mentioned previously, but it would lead to a loss in the structure of these features, which in turn will lose interpretability. Instead of removing site effects after estimating the components, we jointly model the sparse components and the site effects, and estimate robust to site hSCP (rshSCP). We first look at the case when there is only one hierarchy level, which can then be extended to multiple levels. Let there be total S sites, $\mathcal{I}_s$ be the set storing the labels of subjects from the site s and $\mathbf{y} \in \mathbb{R}^{N \times S}$ be the one-hot encoded site labels. We hypothesize that there is a space $\mathbf{V} \in \mathbb{R}^{P \times P}$ storing site and scanner information for all the possible available data, and for each site s, we have space $\mathbf{U}^s \in \mathbb{R}^{P \times P}$ storing site-specific information for $s = 1, \ldots, S$. Based on the above hypothesis, we decompose the correlation matrix $\mathbf{\Theta}^n$ of $n \in \mathcal{I}_s$ to jointly estimate the hSCPs, $\mathbf{U}^s$ and $\mathbf{V}$ as:

$$\mathbf{\Theta}^n \approx \underbrace{\mathbf{W}\mathbf{\Lambda}^n\mathbf{W}^\top}_{\substack{\text{decomposition of}\\ \text{subject components}}} + \underbrace{\mathbf{U}^s\mathbf{V}}_{\substack{\text{decomposition of}\\ \text{site components}}} , \tag{4.1}$$

where $\mathbf{U}^s$ is constrained to be a diagonal matrix and L_1 sparsity constraint is used for $\mathbf{V}$ to prevent overfitting. In addition to estimating the site effects, we modify subject-specific information $\mathcal{D}$ such that the predictive power to predict site is reduced, which can assist in removing site information. There are two parts of this procedure. In the first part, we train a differentiable classification model $F(\zeta, \mathcal{D})$ parameterized by ζ with input $\mathbf{\Lambda}^n$ that return site predictions $\hat{\mathbf{y}} \in \mathbb{R}^{N \times S}$. These predictions indicate the probabilities that each of N inputs belongs to each of S site labels. The classification model is trained by optimizing for ζ such that the cross-entropy loss $\mathcal{L}(\zeta, \mathcal{D}, \mathbf{y})$ between the predictions $\hat{\mathbf{y}}$ and the true site labels $\mathbf{y}$ is minimized:

$$\zeta^* = \arg\min_{\zeta} -\frac{1}{N}\sum_{n=1}^{N}\sum_{s=1}^{S} y_{n,s}\log \hat{y}_{n,s}. \tag{4.2}$$

In the second part, using this classification model, we modify $\mathbf{\Lambda}^n$ such that its predictability power reduces. We achieve this by maximizing the above loss with respect to $\mathcal{D}$. This will result in a minimax game, where the classifier learns to minimize the cross-entropy or the surrogate classification loss, and $\mathcal{D}$ is adjusted to maximize the loss. The joint optimization problem can be written as:

$$\begin{aligned} \max_{\zeta}\ \min_{\mathbf{W},\mathcal{D},\mathcal{U},\mathbf{V}} \quad & \sum_{s=1}^{S}\sum_{n\in\mathcal{I}_s} \|\mathbf{\Theta}^n - \mathbf{W}\mathbf{\Lambda}^n\mathbf{W}^\top - \mathbf{U}^s\mathbf{V}\|_F^2 - \gamma\mathcal{L}(\zeta,\mathcal{D},\mathbf{y}) \\ s.t. \quad & \mathbf{W}\in\Omega, \quad \mathcal{D}\in\Psi, \quad \|\mathbf{v}_p\|_1 < \mu,\ p=1,\ldots,P, \end{aligned} \tag{4.3}$$

where $\mathcal{U} = \{U_s | s = 1,\ldots,S\}$, $\mathbf{v}_p$ is the pth column of $\mathbf{V}$, $\mathcal{W} = \{\mathbf{W}_r \mid r = 1,\ldots,K\}$ be the set storing sparse components shared across all subjects and $\mathcal{D} = \{\mathbf{\Lambda}_r^n \mid r = 1,\ldots,K; n = 1\ldots,N\}$ be set storing subject specific diagonal matrix with $\mathbf{\Lambda}_r^n \geq 0$.

4.2.4 Complete Model

We can combine the above formulation (4.3) at multi level with the adversarial learning (3.6) to jointly model hSCPS and site effects. Let

$$G(\mathcal{W},\mathcal{D},\mathcal{C}) = \sum_{s=1}^{S}\sum_{n\in\mathcal{I}_s}\sum_{r=1}^{K}\|\mathbf{\Theta}^n - (\prod_{j=1}^{r}\mathbf{W}_j)\mathbf{\Lambda}_r^n(\prod_{n=1}^{r}\mathbf{W}_n)^\top - \mathbf{U}_r^s\mathbf{V}_r\|_F^2, \tag{4.4}$$

where $\mathcal{C} = \{\mathbf{\Theta}^n \mid n = 1,\ldots,N\}$, then the joint optimization problem can be written as:

$$\begin{aligned} \max_{\zeta}\ \min_{\mathcal{W},\mathcal{D},\mathcal{U},\mathbf{V}} \quad & J(\tilde{\mathcal{W}},\mathcal{W},\mathcal{D},\mathcal{C}) = G(\tilde{\mathcal{W}},\mathcal{D},\mathcal{C}) + \beta G(\mathcal{W},\mathcal{D},\mathcal{C}) + \gamma\mathcal{L}(\zeta,\mathcal{D},\mathbf{y}) \\ \text{s.t.} \quad & \tilde{\mathbf{W}}_r = \arg\min_{\hat{\mathbf{W}}_r} \alpha\|\hat{\mathbf{W}}_r - \mathbf{W}_r\|_F^2 + G(\tilde{\mathcal{W}},\mathcal{D},\mathcal{P}) \quad r=1,\ldots,K \\ & \tilde{\mathbf{W}}_r,\mathbf{W}_r\in\Omega \quad \mathcal{D}\in\Psi, \quad \|\mathbf{v}_p\|_1 < \mu,\ p=1,\ldots,P. \end{aligned} \tag{4.5}$$

The optimization problem defined above is a non-convex problem that we solve using alternating minimization. Complete algorithm and the details about the optimization are described in Section 4.3. Note that the random initialization of the variables can result in a very different final solution that might be far from the ground truth. One such solution for $\mathbf{U}$ and $\mathbf{V}$ would be the identity matrix since all the correlation matrices have one as their diagonal element, which can drastically change the final components. It might also be possible that $\mathbf{V}$ might store highly reproducible components since they are present in most individuals, leading to a decrease in reproducibility of hSCPs. We prevent these cases by using $\mathrm{svd} - \mathrm{initialization}$ algorithm 2 for $\mathcal{W}$ and $\mathcal{D}$ defined in Subsection 2.3.5, where, in the starting, most of the variability associated with data is stored in $\mathcal{W}$ and $\mathcal{D}$. In this way, we can prevent $\mathbf{V}$ from storing highly reproducible components during initial iterations. We initialize $\mathbf{U}^s$ and $\mathbf{V}$ using the below equation:

$$\begin{aligned} \mathbf{U}_r^s &= \left[\frac{1}{|\mathcal{I}_s|} \left(\sum_{n \in \mathcal{I}_s} \mathbf{\Theta}^n - (\prod_{j=1}^{r} \mathbf{W}_j) \mathbf{\Lambda}_r^n (\prod_{n=1}^{r} \mathbf{W}_n)^\top \right) \mathbf{J}_p \right] \circ \mathbf{I}_p, \\ \mathbf{V}_r &= \frac{1}{P} \mathbf{J}_P. \end{aligned} \tag{4.6}$$

This complete initialization procedure ensures that the algorithm starts with the majority of variability in the data stored in $\mathcal{W}$ and $\mathcal{D}$, and $\mathbf{U}^s$ start from the residual variance left in site s after the $\mathrm{svd} - \mathrm{initialization}$ procedure. We show in the next sections that this simple strategy, though sub-optimal, can help estimate reproducible components with diminished site effects.

4.3 Algorithm

4.3.1 Alternating Minimization

Algorithm 5 describes the complete alternating minimization procedure. $\mathcal{W}$ and $\mathcal{D}$ are initialized using $\mathrm{svd} - \mathrm{initialization}$ algorithm [Alogirthm 2] in Subsection 2.3.5, and $\mathcal{U}$ and $\mathbf{V}$ according to the equation 4.6. $\mathrm{proj}_1(\mathbf{W}, \tau)$ and proj_2 operator are defined in Subsection 2.3.3 and proj_3 operator is used for projection onto L_1 ball. We use AMSGrad (Reddi et al., 2019) denoted as $\mathrm{descent}$ in the algorithm with gradients defined in the next section for performing gradient descent for all the

Algorithm 5 rshSCP

```
Input: Data $\mathcal{C}$, number of connectivity patterns $k_1, \ldots, k_K$ and sparsity $\tau_1, \ldots, \tau_K$ at different level, hyperparameters $\alpha$, $\beta$, $\gamma$ and $\mu$.
Initialize $\mathcal{W}$ and $\mathcal{D}$ using $\mathrm{svd-initialization}$
Initialize $\mathcal{U}$ and $\mathbf{V}$ using equation 4.6
repeat
    for $r = 1$ to $K$ do
        if Starting criterion is met then
            Update adversarial perturbations
            $\hat{\mathbf{W}}_r \leftarrow \mathrm{descent}(\hat{\mathbf{W}}_r, \alpha)$
            $\mathbf{W}_r \leftarrow \mathrm{descent}(\mathbf{W}_r)$
        if $r == 1$ then
            $\mathbf{W}_r \leftarrow \mathrm{proj}_1(\mathbf{W}_r, \tau_r)$
        else
            $\mathbf{W}_r \leftarrow \mathrm{proj}_2(\mathbf{W}_r)$
        for $n = 1, .., N$ do
            $\mathbf{\Lambda}_r^n \leftarrow \mathrm{descent}(\mathbf{\Lambda}_r^n, \beta, \gamma)$
            $\mathbf{\Lambda}_r^n \leftarrow \mathrm{proj}_2(\mathbf{\Lambda}_r^n)$
    for $s = 1$ to $S$ do
        $\mathbf{U}^s \leftarrow \mathrm{descent}(\mathbf{U}^s)$
    $\mathbf{V} \leftarrow \mathrm{descent}(\mathbf{V})$
    $\mathbf{V} \leftarrow \mathrm{proj}_3(\mathbf{V}, \mu)$
until Stopping criterion is reached
Output: $\mathcal{W}$ and $\mathcal{D}$
```

variables. β_1 and β_2 are kept to be 0.9 and 0.999 in AMSGrad. We start adversarial training only after the convergence of all the variables. We found that the algorithm uses 200 iterations to reach convergence initially, as shown in Figure 15. The reason being that adversarial learning can start from an optimal point on which it can improve upon if there is overfitting. When the adversarial learning starts, first, the adversarial perturbations are generated by performing gradient descent on $\hat{\mathbf{W}}_r$, and then the model parameters are updated using gradient descent. This process is repeated until the convergence criteria is met.

4.3.2 Gradient Calculations

In this section, we define gradients used for alternating gradient descent. Let

$$\tilde{\mathbf{W}}_0 = \mathbf{W}_0 = \mathbf{I}_P, \qquad \mathbf{Y}_r = \prod_{j=0}^{r} \mathbf{W}_j, \qquad \tilde{\mathbf{Y}}_r = \prod_{j=0}^{r} \tilde{\mathbf{W}}_j,$$

$$\mathbf{T}^r_{m,n} = (\prod_{j=1}^{m-r} \mathbf{W}_j)\boldsymbol{\Lambda}^n_{m-r}(\prod_{j=1}^{m-r} \mathbf{W}_j)^\top, \qquad \tilde{\mathbf{T}}^r_{m,n} = (\prod_{j=1}^{m-r} \tilde{\mathbf{W}}_j)\boldsymbol{\Lambda}^n_{m-r}(\prod_{j=1}^{m-r} \tilde{\mathbf{W}}_j)^\top,$$

$$\mathbf{X}^n_r = \boldsymbol{\Theta}^n - \mathbf{U}^s_r\mathbf{V}_r, \qquad \mathbf{Z}^n_r = \boldsymbol{\Theta}^n - (\prod_{j=1}^{r} \mathbf{W}_j)\boldsymbol{\Lambda}^n_r(\prod_{n=1}^{r} \mathbf{W}_n)^\top,$$

where $n \in \mathcal{I}_s$, $\mathbf{X}^n_r$ stores the information after removing site effects from $\boldsymbol{\Theta}^n$ and $\mathbf{Z}^n_r$ stores the information after removing subject-wise and shared component information at the rth level. We first define gradient for updating adversarial perturbations $\tilde{\mathbf{W}}_\mathbf{r}$. The gradient of classifier loss with respect to $\mathcal{D}$ is calculated using automatic differentiation provided by MATLAB. The objective function is $F = \alpha\|\hat{\mathbf{W}}_r - \mathbf{W}_r\|^2_F + H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{P})$ and gradient with respect to $\tilde{\mathbf{W}}_\mathbf{r}$ will be

$$\begin{aligned}
\frac{F}{\partial\tilde{\mathbf{W}}_\mathbf{r}} &= 2\alpha(\hat{\mathbf{W}}_r - \mathbf{W}_r) + \frac{\partial H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{C})}{\partial\tilde{\mathbf{W}}_r} \\
&= 2\alpha(\hat{\mathbf{W}}_r - \mathbf{W}_r) + \sum_{n=1}^{N}\sum_{j=r}^{K}\big(-4\tilde{\mathbf{Y}}^\top_{r-1}\boldsymbol{\Gamma}^n\tilde{\mathbf{Y}}_{r-1}\tilde{\mathbf{W}}_r\tilde{\mathbf{T}}^r_{j,n} \\
&+ 4\tilde{\mathbf{Y}}^\top_{r-1}\tilde{\mathbf{Y}}_{r-1}\tilde{\mathbf{W}}_r\tilde{\mathbf{T}}^r_{j,n}\tilde{\mathbf{W}}^\top_r\tilde{\mathbf{Y}}^\top_{r-1}\tilde{\mathbf{Y}}_{r-1}\tilde{\mathbf{W}}_r\tilde{\mathbf{T}}^r_{j,n}\big).
\end{aligned}$$

We now define gradients for updating model parameters. The gradient of objective function J with respect to $\boldsymbol{\Lambda}^n_r$ is:

$$\begin{aligned}
&\frac{\partial J}{\partial\boldsymbol{\Lambda}^n_r} = \frac{\partial H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{C})}{\partial\boldsymbol{\Lambda}^n_r} + \beta\frac{\partial H(\mathcal{W}, \mathcal{D}, \mathcal{C})}{\partial\boldsymbol{\Lambda}^n_r} + \gamma\frac{\partial\mathcal{L}(\zeta, \mathcal{D}, \mathbf{y})}{\partial\boldsymbol{\Lambda}^n_r} \\
&= \Big[(-2\tilde{\mathbf{Y}}^\top_r\mathbf{X}^n_r\tilde{\mathbf{Y}}_r + 2\tilde{\mathbf{Y}}^\top_r\tilde{\mathbf{Y}}_r\boldsymbol{\Lambda}^n_r\tilde{\mathbf{Y}}^\top_r\tilde{\mathbf{Y}}_r) + \beta(-2\mathbf{Y}^\top_r\mathbf{X}^i_r\mathbf{Y}_r + 2\mathbf{Y}^\top_r\mathbf{Y}_r\boldsymbol{\Lambda}^i_r\mathbf{Y}^\top_r\mathbf{Y}_r)\Big] \circ \mathbf{I}_{k_r} + \gamma\mathbf{F},
\end{aligned}$$

where is $\mathbf{F}$ i.e $\frac{\partial \mathcal{L}(\zeta,\mathcal{D},\mathbf{y})}{\partial \mathbf{\Lambda}_r^i}$ is calcualted using automatic differentiation toolbox in MATLAB. The gradient of J with respect to $\mathbf{W}_r$ is:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{W}_r} &= \frac{\partial H(\mathcal{W},\mathcal{D},\mathcal{C})}{\partial \mathbf{W}_r}\\ &= \sum_{n=1}^{N}\sum_{j=r}^{K} -4\mathbf{Y}_{r-1}^{\top}\mathbf{X}_n\mathbf{Y}_{r-1}\mathbf{W}_r\mathbf{T}_{j,n}^{r} + 4\mathbf{Y}_{r-1}^{\top}\mathbf{Y}_{r-1}\mathbf{W}_r\mathbf{T}_{j,n}^{r}\mathbf{W}_r^{\top}\mathbf{Y}_{r-1}^{\top}\mathbf{Y}_{r-1}\mathbf{W}_r\mathbf{T}_{j,n}^{r}.\end{aligned}$$

The gradient J with respect to$\mathbf{U}^s$ and $\mathbf{V}$ are:

$$\frac{\partial J}{\partial \mathbf{U}^s} = \left(\sum_{n=\mathcal{I}_s} \left(\mathbf{Z}^n - \mathbf{U}^s\mathbf{V}\right)\mathbf{V}^{\top}\right) \circ \mathbf{I}_p,$$

$$\frac{\partial J}{\partial \mathbf{V}} = \sum_{s=1}^{S}\sum_{n\in\mathcal{I}_s} \mathbf{U}^s\left(\mathbf{Z}^n - \mathbf{U}^s\mathbf{V}\right).$$

4.4 Experiments

4.4.1 Simulated Dataset

One level. We first generate simulated dataset at one level to evaluate the performance of our model against the standard hSCP. We simulate data with $p = 50$, $k_1 = 10$, $S = 4$ with 200, 300, 400 and 500 number of participants in each site. We generated sparse shared components $\mathbf{W}_1$ with percentage of non-zeros equal to 60% and each element sampled from $\mathcal{N}(0,1)$. We then generate correlation matrix for nth subject belonging to sth site using:

$$\mathbf{\Theta}^n = \left(\mathbf{W}_1 + \mathbf{E}_1^n\right)\mathbf{\Lambda}^n\left(\mathbf{W}_1 + \mathbf{E}_2^n\right)^{\top} + \mathbf{U}^s\mathbf{V} + \mathbf{E}_2^n, \tag{4.7}$$

where $\mathbf{U}^s$ is a diagonal matrix with positive elements sampled from $\mathcal{N}(1,.1)$, $\mathbf{V}$ is a random matrix sampled from wishart distribution, each element of $\mathbf{\Lambda}^n$ is sampled from $\mathcal{N}(4,1)$ and $\mathbf{E}_1^n$ is the noise matrix added to the components whose each element is sampled from $\mathcal{N}(0,.1)$ and $\mathbf{E}_2^n$ is added to

Method	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
hSCP	0.789	0.787	0.745	0.736
ComBat hSCP	0.763	0.759	0.731	0.718
Adv. hSCP	0.869	0.875	0.862	0.854
rshSCP	0.873	0.865	0.843	0.867
Adv. rshCP	$\mathbf{0.903}$	$\mathbf{0.910}$	$\mathbf{0.902}$	$\mathbf{0.908}$
rshSCP w/ rand.	0.856 ± 0.039	0.834 ± 0.055	0.824 ± 0.031	0.818 ± 0.036
Adv. rshSCP. w/ rand.	0.897 ± 0.030	0.895 ± 0.036	0.892 ± 0.023	0.886 ± 0.034

Table 13: Accuracy of the components on simulated dataset at one level.

ensure that the final matrix is positive definite. However the diagonal elements of $\mathbf{\Theta}^n$ are not equal to 1. To make them 1, we extract diagonal elements $\mathbf{D}$ of $\mathbf{\Theta}^n$ and get the new correlation matrix as $\mathbf{D}^{1/2}\mathbf{\Theta}^n\mathbf{D}^{1/2}$. We used a feed-forward neural network for the classification model with two hidden layers. The networks contain the following layers: a fully connected layer with 50 hidden unites, dropout layer with rate 0.2, ReLU, a fully-connected layer with 4 hidden units and a softmax layer. Optimal value of hyperparameters α, β, μ and τ_1 are selected from $[0.1, 1]$, $[1, 5]$, $[0.1, 0.5, 1]$ and $10^{[-2:2]}$. The criterion for choosing the best hyperparameter is maximum split-sample reproducibility. The split sample reproducibility is the normalized inner product between the components estimated on two random equal splits of the data. Split sample reproducibility tries to answer the question of whether the components are generalizable across subjects from the same sites or not.

We compared different methods for estimation of hierarhical components- hSCP, ComBat hSCP, hSCP with adversarial learning (Adv. hSCP), rshSCP, rshSCP with adversarial learning (Adv. rshSCP), rshSCP and Adv. rshSCP with random initialization (rshSCP w/ rand. and Adv. rshSCP w/ rand.). Table 14 shows the reproducibility of the components generated from different methods. It is computed over 15 runs in all the experiments. We used accuracy of the estimated components as a performance measure. It is defined as the normalized inner product between ground truth components and estimated components. All the experiments were run on a four i7-6700HQ CPU cores single ubuntu machine.

Accuracy. Table 13 displays the accuracy of different methods on the simulated dataset. Here, accuracy is defined as the average correlation between estimated components and the ground truth

Method	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
hSCP	0.769 ± 0.052	0.798 ± 0.047	0.739 ± 0.053	0.734 ± 0.047
ComBat hSCP	0.749 ± 0.040	0.750 ± 0.052	0.724 ± 0.049	0.719 ± 0.052
Adv. hSCP	0.781 ± 0.037	0.818 ± 0.031	0.780 ± 0.034	0.750 ± 0.031
rshSCP	0.825 ± 0.039	0.845 ± 0.030	$\mathbf{0.826 \pm 0.039}$	0.779 ± 0.036
Adv. rshSCP	$\mathbf{0.840 \pm 0.044}$	$\mathbf{0.869 \pm 0.034}$	0.815 ± 0.035	$\mathbf{0.802 \pm 0.030}$
rshSCP w/ rand.	0.804 ± 0.085	0.818 ± 0.086	0.780 ± 0.068	0.758 ± 0.071
Adv. rshSCP w/ rand.	0.826 ± 0.069	0.833 ± 0.081	0.801 ± 0.074	0.782 ± 0.077

Table 14: Split sample reproducbility on simulated dataset at one level.

μ	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
0.1	0.799	0.756	0.761	0.782
0.5	0.873	0.865	0.843	0.867
1	0.801	0.789	0.767	0.754

Table 15: Change in accuracy of rshSCP with sparsity parameter (μ) of $\mathbf{V}$ on simulated dataset at one level.

components. From the results, we can see that the rshSCP with adversarial learning can significantly improve the components' accuracy and the reproducibility of the components. The baseline (ComBat hSCP) performs worse than standard hSCP. One reason for this might be that the harmonized $\mathbf{\Lambda}$ extracted using ComBat might not necessarily result in optimal highly reproducible $\mathbf{W}$. This result bolsters our method that we need a joint optimization procedure to obtain $\mathbf{W}$ and $\mathbf{\Lambda}$ with reduced site effects. The results using random initialization instead of using the initialization strategy mentioned in the previous section indicates that random initialization brings significant variability to performance. On average, it performs worse than our strategy, but there might be instances where the random initialization can perform better, which might suggest that there might be some better strategy for initialization. Also, for $\mathbf{V}$, there is an optimal sparsity value, which achieves the best result. If $\mathbf{V}$ is dense, then it might remove essential information that might reduce reproducibility, and if it is too sparse, then we might not have desired effects to make the model robust. The results showing the variation in the accuracy with the sparsity of $\mathbf{V}$ are in Table 15.

Site prediction. To check if the estimated subject information ($\mathbf{\Lambda}$) has reduced predictive power to predict the site to which the subject belonged, we performed a 5 fold cross-validation using SVM

	$k_2 = 4$				$k_2 = 6$			
Method \ k_1	8	10	12	14	8	10	12	14
hSCP	0.806	0.801	0.783	0.777	0.797	0.790	0.773	0.766
ComBat hSCP	0.788	0.776	0.743	0.729	0.779	0.766	0.747	0.734
Adv. hSCP	0.875	0.872	0.870	0.864	0.863	0.859	0.849	0.851
rshSCP	0.881	0.876	0.860	0.862	0.874	0.871	0.852	0.859
Adv. rshSCP	**0.904**	**0.909**	**0.904**	**0.907**	**0.902**	**0.903**	**0.904**	**0.902**

Table 16: Accuracy of the components on hierarchical simulated dataset.

with RBF kernel. We also ran our experiment using two different feed forward networks with two different architectures:

1. a fully connected layer with 50 hidden units, dropout layer with the rate 0.2, ReLU, a fully-connected layer with 4 hidden units and a softmax layer and

2. a fully connected layer with 50 hidden units, dropout layer with the rate 0.2, ReLU, a fully-connected layer with 20 hidden units, dropout layer with rate 0.2, ReLU, a fully-connected layer with 4 hidden units and a softmax layer.

Our model leads to a decrease in average cross-validation accuracy from 97.6% to 67% for SVM, 98.1% to 67.3% for neural network with architecture 1 and 98.2% to 66.9% for neural network with architecture 2. This suggests that our model can reduce the prediction capability to predict site.

Two level. Under the same settings as defined above, we generate correlation matrix from two level components with $k_2 = 4$ using:

$$\begin{aligned} \mathbf{\Theta}^n &= \tilde{\mathbf{W}}_1\tilde{\mathbf{W}}_2\mathbf{\Lambda}^n\tilde{\mathbf{W}}_2^\top\tilde{\mathbf{W}}_1^\top + \mathbf{U}^s\mathbf{V} + \mathbf{E}_3^n, \\ \tilde{\mathbf{W}}_1 &= \mathbf{W}_1\mathbf{E}_1^n, \quad \tilde{\mathbf{W}}_2 = \mathbf{W}_2 + \mathbf{E}_2^n, \end{aligned} \tag{4.8}$$

where each element of $\mathbf{W}_2$ is sampled from $\mathcal{N}(0, 1)$, the percentage of non-zeros equal to 40%, $\mathbf{E}_1^n$ and $\mathbf{E}_2^n$ is the noise added to the components whose each element is sampled from $\mathcal{N}(0, .1)$ and $\mathbf{E}_3^n$ is added to ensure that the final matrix is positive definite. Table 16 shows the accuracy for different values of k_1 and k_2. Selection of hyparameter is same as in the previous paragraph. We can see that

Method	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
hSCP	97.3 ± 0.3	98.1 ± 0.4	97.1 ± 0.3	97.9 ± 0.2
Adv. rshSCP	65.5 ± 0.6	67.2 ± 0.5	67.5 ± 0.5	68.1 ± 0.7

Table 17: 5 fold cross validation accuracy (%) on simulated dataset at one level.

Method	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
hSCP	0.801 ± 0.037	0.805 ± 0.042	0.787 ± 0.041	0.772 ± 0.037
ComBat hSCP	0.776 ± 0.041	0.756 ± 0.044	0.753 ± 0.045	0.745 ± 0.038
Adv. hSCP	0.808 ± 0.036	0.824 ± 0.034	0.799 ± 0.030	0.783 ± 0.038
rshSCP	0.850 ± 0.037	0.853 ± 0.031	0.839 ± 0.035	0.805 ± 0.034
Adv. rshSCP	$\mathbf{0.852 \pm 0.036}$	$\mathbf{0.861 \pm 0.038}$	$\mathbf{0.842 \pm 0.043}$	$\mathbf{0.813 \pm 0.035}$

Table 18: Reproducbility on simulated dataset ($k_2 = 4$).

Method	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
hSCP	0.786 ± 0.041	0.801 ± 0.039	0.771 ± 0.042	0.769 ± 0.038
ComBat hSCP	0.779 ± 0.044	0.770 ± 0.041	0.742 ± 0.040	0.734 ± 0.045
Adv. hSCP	0.793 ± 0.036	0.828 ± 0.038	0.789 ± 0.035	0.762 ± 0.036
rshSCP	0.833 ± 0.038	0.846 ± 0.031	0.831 ± 0.032	0.795 ± 0.034
Adv. rshSCP	$\mathbf{0.841 \pm 0.039}$	$\mathbf{0.851 \pm 0.035}$	$\mathbf{0.835 \pm 0.033}$	$\mathbf{0.808 \pm 0.039}$

Table 19: Reproducbility on simulated dataset ($k_2 = 6$).

the proposed method estimates most accurate ground truth components.

4.4.2 Real Dataset

Data

We collected functional MRI data from 5 different multi-center imaging studies- 1) Baltimore Longitudinal Study of Aging (BLSA) (Armstrong et al., 2019; Resnick et al., 2003), the Coronary Artery Risk Development in Young Adults study (CARDIA) (Friedman et al., 1988), UK BioBank (UKBB) (Sudlow et al., 2015), Open access series of imaging studies (OASIS) (Marcus et al., 2007) and Aging Brain Cohort Study (ABC) from Penn Memory Center (Pluta et al., 2012). Although UK Biobank has more than 20000 scans, we only used 2023 randomly selected scans to avoid estimating the results that would be heavily influenced by the UK Biobank. We projected the data into a lower-dimensional space such that the number of nodes in each subject's data was 100. Table

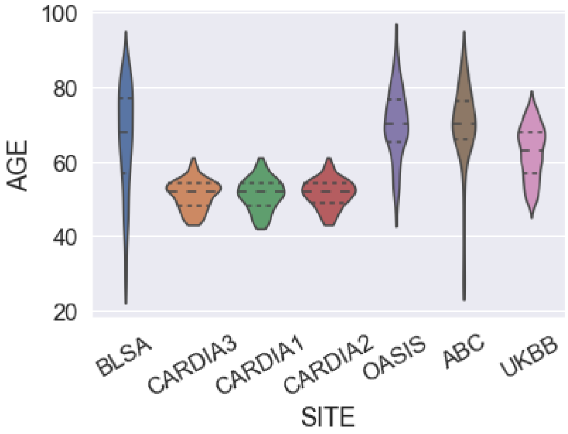


Figure 14: Violin plot of age for different sites.

Data Sites	Participants	% of Females	Age Range (Median)	Scanner
BLSA-3T	784	56.5	$[22, 95](68)$	3T Philips
CARDIA1	199	55.7	$[42, 61](52)$	3T Siemens Tim Trio
CARDIA2	321	51.4	$[43, 61](52)$	3T Philips Achieva
CARDIA3	278	55.3	$[43, 62](52)$	3T Philips Achieva
UKBB	2023	55.2	$[45, 79](63)$	3T Siemens Skyra
OASIS	847	56.0	$[42.6, 97](70)$	1.5T Siemens Vision
ABC	279	59.1	$[23, 95](70)$	3T Siemens Tim Trio

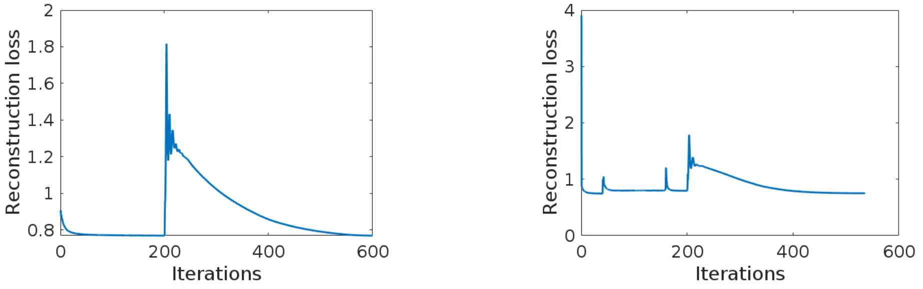
Table 20: Summary characteristics of the real dataset.

20 summarizes the number of participants in each site and age distribution. CARDIA data is divided into three parts because of the acquisition at three different sites.

Data Preprocessing

The pooled dataset included scans of participants with absence of any known diagnosis of a neurological or psychiatric disorder. FMRIB Software (Jenkinson et al., 2012) is used for initial pre-processing as a part of the UK Biobank pipeline. The steps included the removal of the first five volumes, head movement correction using FSL's MCFLIRT (Jenkinson et al., 2012), global 4D mean intensity normalization, and temporal high-pass filtering (> 0.01 Hz).

After standard pre-processing steps, we applied FIX (FMRIB's ICA-based Xnoiseifier) (Salimi-Khorshidi et al., 2014; Griffanti et al., 2014) to remove structured artefacts. In the next step, functional images were co-registered to T1 using FLIRT with BBR as the cost function, and T1-weighted images



(a) $k_1 = 10, k_2 = 4$

(b) $k_1 = 10, k_2 = 6$

Figure 15: Convergence of rshSCP algorithm using the complete dataset for different values of k_2.

were registered to the MNI152 template using FSL's FNIRT (non-linear registration). We projected the data into a lower-dimensional space by extracting a set of group Independent Components (Smith et al., 2014) having dimension 100 from individual subjects. These ICA maps can be considered "parcellations" but contain a continuous range of values and not binary masks. For a given IC map, the group IC spatial maps were mapped onto each subject's resting fMRI time series to derive one representative time series per IC component using Group Information Guided ICA(GIGICA) (Du and Fan, 2013). Metrics used for quality control are defined in Appendix B.

Convergence results We empirically validate the convergence of Algorithm 5 using the reconstruction error:

$$\frac{\sum_{s=1}^{S}\sum_{n\in\mathcal{I}_s}\sum_{r=1}^{K}||\boldsymbol{\Theta}^n - (\prod_{j=1}^{r}\mathbf{W}_j)\boldsymbol{\Lambda}_r^n(\prod_{j=1}^{r}\mathbf{W}_j)^\top - \mathbf{U}^s\mathbf{V}||_F^2}{\sum_{n=1}^{N}\sum_{r=1}^{K}||\boldsymbol{\Theta}^n||_F^2}.$$

Figure 15 shows the convergence of the algorithm on the complete dataset. In the figure, for the first 200 iterations, the algorithm converges without the adversarial perturbations. As the adversarial perturbations are introduced, the loss starts to oscillate where the adversarial perturbations force the algorithm to deviate from the optimal value. In defense, we minimize the objective function until convergence is reached.

Reproducibility. Since we don't have access to ground truth here, we compare the methods based on the split sample and leave one site reproducibility. Leave one site out reproducibility is defined as the similarity between components derived from the site s and all sites except s. Split sample

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	0.713 ± 0.039	0.707 ± 0.038	0.697 ± 0.035	0.683 ± 0.036
ComBat hSCP	0.673 ± 0.049	0.641 ± 0.051	0.639 ± 0.031	0.611 ± 0.038
Adv. hSCP	0.737 ± 0.041	0.719 ± 0.033	0.715 ± 0.037	0.710 ± 0.043
rshSCP	0.806 ± 0.036	0.768 ± 0.032	0.742 ± 0.033	0.743 ± 0.044
Adv. rshSCP	$\mathbf{0.808 \pm 0.030}$	$\mathbf{0.772 \pm 0.036}$	$\mathbf{0.747 \pm 0.034}$	$\mathbf{0.746 \pm 0.036}$

Table 21: Split-sample reproducbility on real dataset ($k_2 = 4$).

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	0.652 ± 0.038	0.618 ± 0.041	0.592 ± 0.033	0.571 ± 0.035
ComBat hSCP	0.614 ± 0.042	0.594 ± 0.035	0.542 ± 0.041	0.528 ± 0.039
Adv. hSCP	0.656 ± 0.035	0.629 ± 0.039	0.601 ± 0.035	0.584 ± 0.034
rshSCP	0.712 ± 0.034	0.701 ± 0.036	0.676 ± 0.038	0.665 ± 0.034
Adv. rshSCP	$\mathbf{0.716 \pm 0.032}$	$\mathbf{0.709 \pm 0.031}$	$\mathbf{0.688 \pm 0.034}$	$\mathbf{0.671 \pm 0.033}$

Table 22: Leave one site out reproducbility on real dataset($k_2 = 4$).

reproducibility tries to answer the question of whether the components are generalizable to other sites or not. For estimating rshSCP with only one site, we used $\mathbf{V}$ estimated from all sites except s since the idea behind $\mathbf{V}$ was to store information about the site/scanner from various sites. This would also help analyze the generalization power of $\mathbf{V}$. The optimum value of the hyperparameters is selected from the range defined in Subsection 4.4.1. τ_1 and τ_2 are selected from $10^{[-2:2]}$ based on maximum split-sample reproducibility. The criterion for choosing the best value is the maximum split sample reproducibility. Table 21 shows the split sample reproducibility for varied values of k_1 and $k_2 = 4$. Leave one site out reproducibility results are shown in Table 22. Table 23 and Table 24 shows split sample reproducibility and leave one site out reproducibility respectively at two-level for $k_2 = 6$. The results demonstrate that the proposed method can significantly improve the split sample reproducibility and leave one site out reproducibility. For the remaining chapter, we focus on the comparison between components learned using adversarial learning from hSCP and Adv. rshSCP.

Site prediction. We performed the same experiment under the same settings as mentioned in the previous section to check $\mathbf{\Lambda}$ has reduced predictive power to predict the site. Using SVM, our model leads to a decrease in average cross-validation accuracy from 51% to 32%. Using the first neural network architecture defined in Subsection 4.4.1, the cross-validation accuracy for hSCP model is

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	0.691 ± 0.034	0.688 ± 0.034	0.677 ± 0.029	0.668 ± 0.032
ComBat hSCP	0.670 ± 0.026	0.664 ± 0.028	0.635 ± 0.030	0.626 ± 0.028
Adv. hSCP	0.701 ± 0.026	0.696 ± 0.029	0.681 ± 0.028	0.679 ± 0.031
rshSCP	0.776 ± 0.027	0.748 ± 0.029	0.722 ± 0.032	0.721 ± 0.024
Adv. rshSCP	$\mathbf{0.779 \pm 0.029}$	$\mathbf{0.751 \pm 0.026}$	$\mathbf{0.731 \pm 0.027}$	$\mathbf{0.732 \pm 0.025}$

Table 23: Split-sample reproducbility on real dataset ($k_2 = 6$).

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	0.637 ± 0.035	0.600 ± 0.036	0.578 ± 0.031	0.560 ± 0.034
ComBat hSCP	0.618 ± 0.037	0.589 ± 0.033	0.543 ± 0.035	0.521 ± 0.032
Adv. hSCP	0.642 ± 0.028	0.608 ± 0.021	0.585 ± 0.023	0.572 ± 0.019
rshSCP	0.701 ± 0.032	0.691 ± 0.034	0.668 ± 0.031	0.659 ± 0.029
Adv. rshSCP	$\mathbf{0.703 \pm 0.033}$	$\mathbf{0.695 \pm 0.035}$	$\mathbf{0.672 \pm 0.030}$	$\mathbf{0.666 \pm 0.031}$

Table 24: Leave one site out reproducbility on real dataset($k_2 = 6$).

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	6.490 ± 1.485	6.468 ± 1.442	6.425 ± 1.412	6.414 ± 1.417
Adv. rshSCP	6.494 ± 1.501	6.467 ± 1.475	6.432 ± 1.483	6.409 ± 1.470

Table 25: Mean absolute error ($k_2 = 4$)

59.3% and for the rshSCP is 33.6%. Using the second architecture, the cross-validation accuracy for hSCP model is 58.7% and for the rshSCP is 33.4%. This suggests that our model can reduce the prediction capability to predict site.

Age prediction. We used subject specific information ($\mathbf{\Lambda}$) having total $k_1 + k_2$ features from the two layers to predict age of each subject. We used Bootstrap-aggregated (bagged) decision trees to perform regression with 400 trees for each site separately. Table 25 and 26 shows the average and standard deviation of 10 fold cross validation mean absolute error (MAE) across site for varied values of k_1 and $k_2 = 4, 6$. We decided to perform age prediction of each site separately because the age is confounded by the site. The correlation between age and site is 0.24 and reduction in site effects would reduce the prediction capability in the pooled setting. From the table, we can see that the proposed method has comparable performance as the hSCP suggesting that it preserves age related biological variance.

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	6.472 ± 1.417	6.440 ± 1.470	6.418 ± 1.484	6.401 ± 1.478
Adv. rshSCP	6.475 ± 1.250	6.439 ± 1.411	6.421 ± 1.454	6.403 ± 1.474

Table 26: Mean absolute error ($k_2 = 6$)

		1	2	3	4	5	6	7	8	9	10
a	ρ	0.07	0.0	0.09	-0.12	-0.02	-0.07	-0.15	-0.04	-0.06	0.02
	$-\ln(p)$	9.28	0.27	15.0	35.3	1.57	9.32	33.3	3.10	5.98	1.27
b	ρ	0.05	-0.02	0.11	-0.07	-0.13	-0.05	-0.03	0.0	-0.02	0.02
	$-\ln(p)$	5.29	0.96	20.0	8.77	28.6	5.27	2.46	0.16	1.67	1.44

Table 27: Spearman correlation (ρ) and p-value of age (> 60) with $\mathbf{\Lambda}_1$ computed from hSCP (a) and Adv. rshSCP (b).

4.4.3 Analysis of components

A robust method should be able to reduce non-biological variability caused by site and scanner while retaining biological variability. In this study, we look at brain aging-related associations and leave analysis with other variables for future work. We also discuss the difference between the components with and reduced site effects. We selected the subjects with age greater than 60 to find an association between brain aging and the components derived from hSCP and rshSCP. We computed spearman correlation of age (> 60) with $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. We then calculated p-values for the hypothesis test of no correlation against the alternative hypothesis of a nonzero correlation and are converted to $-\ln(p)$, where $\ln$ is log base 2. Table 27 and 28 displays spearman correlation and negative log p-value. The total number of subjects with age greater than 60 is 2746. In the case of negative log base 2, if the value is greater than 2.99 then we consider it statistically significant, equivalent to p-value less than 0.05. We derived 10 fine-scale ($1 - 10$) and 4 coarse-scale components (I-IV) because of the high split sample reproducibility and easier interpretation of each component. The correlation and p-values are displayed in Table 27 and 28 for the hSCP and rshSCP.

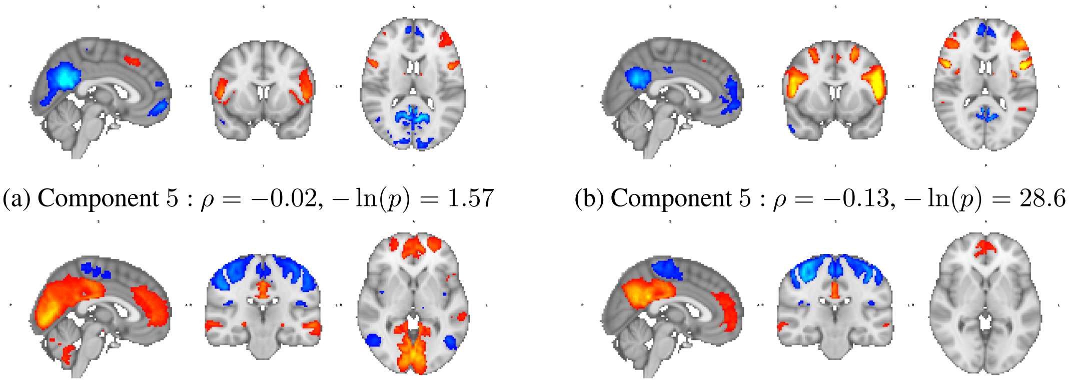
We first compare components from the two methods. Figure 16 shows the components derived from hSCP and the proposed method. Red and blue regions are anti-correlated with each other but are correlated among themselves. The colors are not associated with negative or positive

		I	II	III	IV
hSCP	ρ	0.02	-0.04	-0.08	-0.13
	$-\ln(p)$	1.69	3.68	9.83	24.7
Adv. rshSCP	ρ	0.05	-0.05	-0.06	-0.09
	$-\ln(p)$	4.95	5.34	6.98	13.8

Table 28: Spearman correlation (ρ) and p-value of age (> 60) with $\mathbf{\Lambda}_2$ computed from hSCP and Adv. rshSCP.

correlation since they can be swapped without affecting the final inference. The first row of the figure displays the components with anti-correlation between Default Mode Network (DMN) and Dorsal Attention Network (DAN). The component derived using hSCP has a part of the visual area positively associated with DMN, but the opposite is true, as shown by the previous sparse connectivity patterns (Eavani et al., 2015a). On the other hand, the component with reduced site effects is cleaner since it does not include that relation. This component has a negative correlation with age which has been previously shown in resting-state fMRI and task-based fMRI (Spreng et al., 2016). The magnitude of anti-correlation has been connected to individual differences in task performances in healthy young adults (Keller et al., 2015). However, in the case of older adults, the behavioral implications of reduced anti-correlation remain unclear. The second row of the Figure 16 displays another set of components for comparison. The components stores information about the anti-correlation between DMN and sensorimotor, which aligns with the previous literature (Karahanoğlu and Van De Ville, 2015). But the addition of a positive correlation of DMN with visual areas will cause misleading inference since it contradicts the previous SCPs and studies. Hence making an inference without removing the site effects can be misleading.

Figure 17 displays one of the hierarchical components with coarse-scale component storing relation between different fine-scale components comprising DMN, sensorimotor, and visual areas, previously studied by (Karahanoğlu and Van De Ville, 2015). These findings give evidence that even after removal site effects, the components can have a meaningful interpretation. The results indicate that our approach can extract robust informative patterns without using traditional seed-based methods that are dependent on the knowledge of the seed region of interest.



(a) Component $5 : \rho = -0.02, -\ln(p) = 1.57$

(b) Component $5 : \rho = -0.13, -\ln(p) = 28.6$

(c) Component $6 : \rho = -0.07, -\ln(p) = 9.32$

(d) Component $6 : \rho = -0.05, -\ln(p) = 5.27$

Figure 16: Left column ((a) & (c)) displays the components estimated using hSCP and right column ((b) & (d)) displays the components estimated using rshSCP.

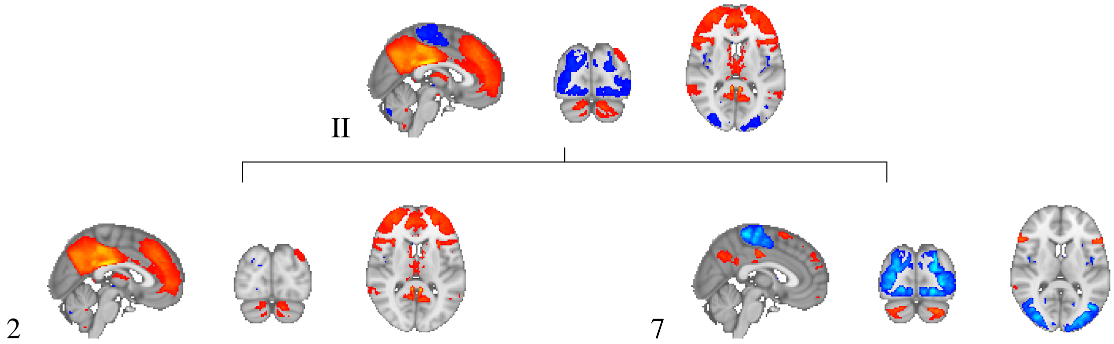
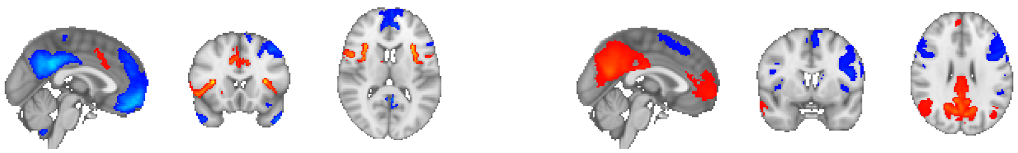


Figure 17: One of the hierarchical components derived from rshSCP comprising of component 2 and 7 at fine scale and component II at coarse scale.

Between-network connectivity in aging. In this part, we discuss related work on changes in between-network connectivity in older adults and connection to our results. Geerligs et al. (2014) published one of the earliest studies on changes in between-network connectivity in older adults using seed-based analysis while participants performed an oddball task. They observed stronger connectivity (or weaker anticorrelations) between distinct functional networks. For example, they found age-related connectivity increases between the DMN, and the somatosensory and the CEN, which aligns with the results of current work. Several other studies reported similar results using different approaches (Ferreira et al., 2016; Geerligs et al., 2015). The DAN and DMN appear to show strong anticorrelations due to their presence in externally directed and internally directed cognition. Spreng et al. (2016) used both resting and task data to show a decrease in anticorrelations between



(a) Component 7 : $\rho = -0.03, -\ln(p) = 2.46$ (b) Component 4 :$\rho = -0.07, -\ln(p) = 8.77$

Figure 18: (a) Anticorrelation between Default Mode Network (DMN) and Salience Network (SN) and (b) Anticorrelation between Default Mode Network (DMN) and Central Executive Network (CEN).

these networks in older compared to younger.

The increase in connectivity between different networks can be thought of as a decrease in the segregation of networks. Previous studies have indicated that this decrease in segregation causes a reduction in the specialization of specialized networks, affecting information processing of the human brain (Schaie and Willis, 2021). Grady et al. (2016) analyzed the connections between DMN, DAN and CEN networks and observed a lower index of segregation in older as compared to young. Our results also indicate a decrease in anticorrelation between various networks, which can be thought of decrease in the segregation of networks, resulting in reorganization of the human brain in old age. From the results, we can see that there is an increase (or decrease in anti-correlation) in connectivity between different networks in the aging brain. This suggests that there is a reorganization of the aging brain aligning with the previous findings (Damoiseaux, 2017). This can serve as a base to explore rshSCP as a biomarker of neurodegenerative diseases.

4.5 Conclusion

In this work, we have presented a method for estimating site effects in hSCP. We formulated the problem as a minimax non-convex optimization problem and solved it using AMSgrad. We also propose a simple initialization procedure to make the optimization procedure deterministic and improve the performance on an average on a simulated and real dataset. Experimentally, using a simulated dataset, we showed that our method accurately estimates the ground truth compared to the standard method with better reproducibility. On the real dataset, we show that the proposed method can capture components with a better split sample and leave one site out reproducibility without losing biological interpretability and information. We also show that without removal of site effects, we can have a noisy estimate of sparse components resulting in misleading downstream analysis.

Below we mention some directions for future research. First, it would be interesting to consider the framework for the analysis of task-induced activity to investigate the extent of site effect and corrections on underlying networks activated by the task. Second, one could look at the changes in the associations of hSCPs with various clinical variables such as Mini-mental score, Digit Span Forward score, etc., after removing site effects. Third, we can also look low-dimensional modeling of $\mathbf{V}$ along with sparse constraints which has been used several robust matrix factorization problems. Since we have only shown age related biological preservation, future studies will focus on whether the proposed method preserves components associated with other demographic, clinical phenotypes, and pathological biomarkers.

There are few weaknesses of our proposed model, which also adds directions for future work. First, our method only captures linear site effects, it would be interesting to see if explicitly capturing non-linear site effects can improve the performance of the model. Second, the result of the optimization algorithm depends on the initialization procedure, which has been shown to perform well on the simulated dataset and real dataset but can be sub-optimal.

CHAPTER 5

Robust Hierarchical Patterns for identifying MDD patients

The previous chapters combined an adversarial learning framework with matrix factorization to develop a robust to site model to estimate hierarchical sparse patterns. In this chapter, we extend the above method to reduce the effects of age, sex and site, and capture robust human brain patterns characterizing Major Depressive Disorder (MDD).

5.1 Introduction

Resting-State functional Magnetic Resonance Imaging (rs-fMRI) is a method of fMRI that can capture patterns of co-activation in the human brain when there is no explicit task performed. These patterns are believed to demonstrate the intrinsic communication between different brain regions (Fox and Raichle, 2007). Consequently, rs-fMRI has been used to characterize neuropsychiatric disorders such as Autism Spectrum Disorder (ASD) (Minshew and Keller, 2010; Heinsfeld et al., 2018; Wolfers et al., 2019), Attention Deficit Hyperactivity Disorder (ADHD) (Bush et al., 2005; Wang et al., 2018; Riaz et al., 2020), anxiety (ANX) (Frick et al., 2014; Liu et al., 2015) and schizophrenia (Niznikiewicz et al., 2003; Rozycki et al., 2018; Yassin et al., 2020). However, researchers have faced challenges to reliably predict clinical manifestations in patients due to the high dimensionality of the data, and inter-patient variability (Benkarim et al., 2021). In addition, biomarkers learned from small and homogenous datasets often result in poor generalization to new or future cohorts, thus posing an issue of replicability of brain patterns valuable for prediction.

Several open-access neuroimaging data-sharing initiatives have been introduced to improve generalizability and replicability of brain patterns and evaluate a hypothesis in multiple sites/settings (Alexander et al., 2017; Biswal et al., 2010; Casey et al., 2018; Di Martino et al., 2017). In these initiatives, data is pooled from multiple sites to capture demographically diverse populations therefore building heterogeneous datasets that are more likely to reflect the wider population. Surprisingly, multi-site studies based on supervised/unsupervised approaches have shown lower classification performance and poor generalization to data from new cohorts compared to single-site studies (Nielsen et al., 2013; Arbabshirani et al., 2017; Munafò et al., 2017; Nosek and Errington, 2017; Dinga et al., 2019; He et al., 2020). Analyzing data from these initiatives poses an inherent challenge due to variability introduced from the diverging backgrounds of the subjects and from site differences in MRI scanner hardware and software (Kostro et al., 2014; Shinohara et al., 2017; Noble et al., 2017). The non-biological variability introduced due to pooling of the data can affect the biomarkers or common features extracted from fMRI data (Yu et al., 2018), these include functional connectivity (Shinohara et al., 2017) and sparse hierarchical factors (Sahoo and Davatzikos, 2021). The non-biological variability can lead to decreased statistical power, spurious results and difficulty in identifying robust biomarkers depending on the task. In addition, the correlation between site effects and biological predictors can lead to an incorrect inference of non-biological differences as biological. Thus, many neuroimaging studies need to develop robust models that remove the non-biological variance and extract biologically relevant information.

This work focuses on extracting functional network-based biomarkers of Major Depressive Disorder (MDD) in a multi-site study. MDD is one of the most widespread psychiatric disorders characterized by persistent sadness, depressed mood, low self-esteem, sleep disturbances, emotional changes, and loss of interest in pleasurable activities, causing disruptions to daily life (Belmaker and Agam, 2008). In addition, MDD causes more than $800,000$ deaths each year globally and is also the leading cause of disability (Otte et al., 2016). Understanding the mechanism of MDD is crucial for effective diagnosis, treatment and prevention, and understanding the functioning of the human brain in a depressive state compared to a healthy one. Considering the breadth of symptoms of this disease, it follows that disruptions within and across multiple brains systems and networks must be at play. Indeed, much work has been done to understand the functional brain changes associated with MDD. However, much remains unknown about the pathophysiology of the disease and the rates of relapse and recurrence remain high (Mueller et al., 1999; Kessler, 2012).

Previous studies have shown that MDD is associated with disruptions in regional functional connectivity and abnormal functional integration of distributed brain regions (Greicius et al., 2007; Liu et al., 2013; Wu et al., 2011; Zhu et al., 2012). More recent approaches using seed-based connectivity, independent component analyses, network homogeneity and graph theory for functional connectivity analyses have revealed similar findings- disruptions in functional networks and in between functional networks across specific region pairs in MDD. The brain networks exhibiting abnormal interactions in MDD include the Default Mode Network (awareness of internal states), Dorsal Attention Network (external awareness), Fronto-Parietal Network (top-down regulation of attention and emotion), Salient Network (salient events) and Affective Network (emotion processing) (Ye et al., 2015; Kaiser et al., 2015; Yan et al., 2019; Mulders et al., 2015; Iwabuchi et al., 2015; Brakowski et al., 2017). Many efforts have been made to build functional connectivity-based predictive models for identifying network-based biomarkers of depression (Craddock et al., 2009; Zeng et al., 2012; Bhaumik et al., 2017; Rosa et al., 2015; Zhao et al., 2020a). Majority of studies are based on multivariate pattern analysis of functional connectivity and have faced challenges generalizing to new site. For instance, Drysdale et al. (2017) created biomarkers from brain networks used in SVM model that achieved more than 82% accuracy from 109 patients from the sites present in the training set and 68.8% accuracy from 16 patients from an independent site. More recently, Yamashita et al. (2021, 2020) used the ComBat harmonization method Johnson et al. (2007); Fortin et al. (2018); Yu et al. (2018) to correct site differences in functional connectivity, but applying ComBat to the functional connectivity matrix destroys its essential property of positive semi-definiteness.

This chapter is aimed to identify the hSCP biomarkers of MDD while reducing the effects of diversity (age, sex, site) common to large pooled datasets. Our work builds on the hSCP model by using the discriminative nature of the subject-specific weights extracted from hSCPs to classify MDD via logistic regression in a large multi-site study. A similar generative-discriminative approach has been used to classify young adults vs. children (Eavani et al., 2014) based on sparse patterns. More recently, D'Souza et al. (2020) used sparse patterns based generative-discriminative model to predict clinically relevant networks characteristic of Autism Spectrum Disorder. To tackle the variability introduced due to pooling of the datasets, we use robust to site hSCP (rshSCP) presented in the previous. The method can capture robust sparse human brain patterns while reducing the site effects improving the generalizability and reproducibility, and capturing aging related patterns. The rshSCP method captures linear site effects and uses adversarial learning to reduce site effects in the subject-specific coefficients. We extend robust to site hSCP and introduce discriminative rshSCP (dis-rshSCP) to extract homogeneous components discriminative of MDD by reducing heterogeneity introduced due to covariates (age, sex and site), which are known to affect neuroimaging analysis (Alfaro-Almagro et al., 2021; Duncan and Northoff, 2013). Experiments on real datasets show that reducing heterogeneity can improve the split-sample and leave one site predictability power of the components while retaining the reproducibility of the components, thus capturing informative heterogeneity. The classification performance on unseen data indicates the generalizability of the model. Our results demonstrate that MDD is associated with increased and decreased representation in patterns associated with various functional networks. The results demonstrate our framework's potential in identifying patient-predictive biomarkers of a MDD.

Outline: We start by presenting our method to extract interpretable hSCPs which are predictive of MDD and are robust to covariates (age, sex and site) in Section 5.2. In Section 5.5, we demonstrate that our method could extract hSCPs with high reproducibility and prediction power on a large multi-site dataset. This is followed by a discussion on the interpretability of the extracted patterns, limitations and future work.

5.2 Method

Problem Setup: The fMRI data of the i^{th} subject having P regions and T time points is denoted by $\mathbf{X}^i \in \mathbb{R}^{P \times T}$ with total N number of subjects or participants. Let $\mathbf{\Theta}^i \in \mathbb{S}_{++}^{P \times P}$ be the correlation matrix where $\mathbf{\Theta}^i_{m,o}$ stores the correlation between time series of m^{th} and o^{th} node. hSCP then outputs a set of shared hierarchical patterns following the below equations:

$$\mathbf{\Theta}^n \approx \mathbf{W}_1 \mathbf{\Lambda}_1^n \mathbf{W}_1^\top, \quad \ldots \quad \mathbf{\Theta}^n \approx \mathbf{W}_1 \mathbf{W}_2 \ldots \mathbf{W}_K \mathbf{\Lambda}_K^n \mathbf{W}_K^\top \mathbf{W}_{K-1}^\top \ldots \mathbf{W}_1^\top,$$

where $\mathbf{\Lambda}_k^n$ is a diagonal matrix having positive elements storing relative contribution of the components for the nth subject at kth level, K is the depth of hierarchy and $P > k_1 > \ldots > k_K$.

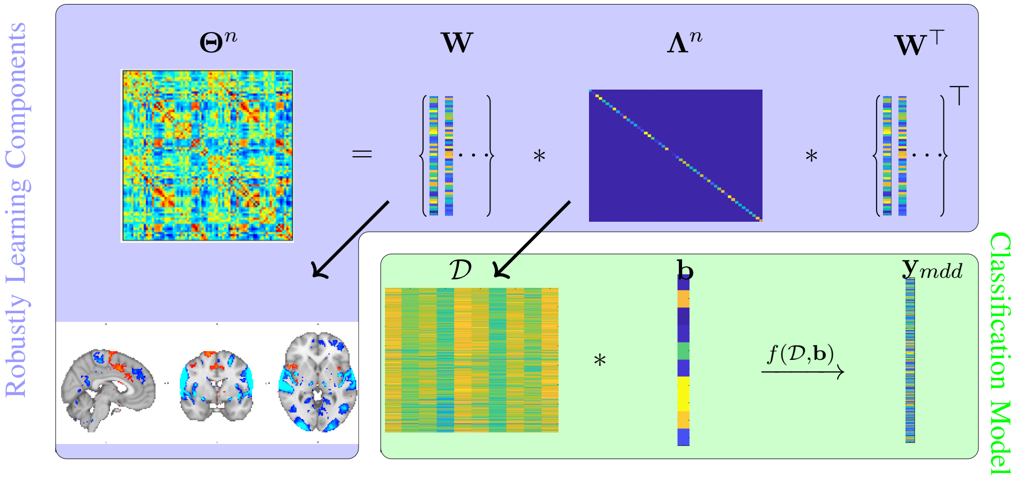


Figure 19: A joint two level modeling for connectivity analysis and prediction. First part is functional data representation depicted in blue box. Here the correlation matrices are decomposed into shared components stored in a basis matrix and subject specific information. We learn this decomposition in a robust manner to reduce variability due to site and demographics. To visualize each component, a column of basis matrix is projected onto the brain. Second part is the prediction of MDD patients depicted in green box.

Let $\mathcal{W} = \{\mathbf{W}_r \mid r = 1, \ldots, K\}$ be the set storing sparse components shared across all subjects and $\mathcal{D} = \{\mathbf{\Lambda}_r^n \mid r = 1, \ldots, K; n = 1 \ldots, N\}$ be set storing subject specific diagonal matrix with $\mathbf{\Lambda}_r^n \geq 0$. Let there be total S sites in the multi-site data and $\mathcal{I}_s$ be the set storing subjects from site s. Let $\mathbf{y}_{site} \in \mathbb{R}^{N \times S}$ be site labels encoded in one-hot manner, and $\mathbf{y}_{age} \in \mathbb{R}^N$, $\mathbf{y}_{sex} \in \mathbb{R}^N$ and $\mathbf{y}_{mdd} \in \mathbb{R}^N$ be the vectors storing information about age, sex and MDD label. We aim to extract set of hierarchical patterns representative of depression using fMRI data with reduced variability introduced due to age, sex and site.

A graphical summary of our model is presented in Fig. 19. The two inputs to our model are the rs-fMRI correlation matrices (upper left) and the binary scores depicting if a person is healthy or has MDD (lower right). The correlation matrices are constructed from the time series data describing the similarity using Pearson's Correlation Coefficient between various nodes of the human brain. The blue box in Fig 19 indicates the generative model estimating the components. Here, we decompose correlation matrices into a set of components capturing co-activation patterns common across the entire cohort and subject-specific information capturing heterogeneity in the data representing the strength of each component in each individual. We capture this information while reducing the effects of demographics and site to improve the generalizability of the components and predictability of the subject-specific information. The green box indicates the discriminative model guiding the components to represent MDD. Here, we leverage the information from the subject-specific coefficients to predict MDD via a classification model for each individual. $f()$ takes subject-specific information and a set of weights as input and maps it to a binary value representing 1 and 0 for MDD and healthy subjects.

5.2.1 Robust to covariates hSCP

In the previous chapter, we saw that how site and scanner effects can be stored using $\mathbf{U}$ and $\mathbf{V}$. In this chapter, in addition to reducing site effects, we also reduce variability due to age and sex. For this, we train a joint model $F(\zeta, \mathcal{D})$ parameterized by ζ with input $\mathbf{\Lambda}^n$ that return age $\hat{\mathbf{y}}_{age} \in \mathbb{R}^N$, sex $\hat{\mathbf{y}}_{sex} \in \mathbb{R}^N$ and site $\hat{\mathbf{y}}_{site} \in \mathbb{R}^{N \times S}$ predictions. The model is trained by optimizing for ζ such that the below loss function $\mathcal{L}_1(\zeta, \mathcal{D}, \mathbf{y}_{site}, \mathbf{y}_{age}, \mathbf{y}_{sex})$ is minimized:

$$\mathcal{L}_1 = \underbrace{\|\hat{\mathbf{y}}_{age} - \mathbf{y}_{age}\|_2^2}_{\substack{\text{preserve age}\\ \text{information}}} - \sum_{s=1}^{S}\sum_{n=1}^{N} \underbrace{y_{site}^{n,s} \log \hat{y}_{site}^{n,s}}_{\substack{\text{preserve site}\\ \text{information}}} - \sum_{n=1}^{N} \underbrace{(y_{sex}^{n} \log \hat{y}_{sex}^{n} + (1 - y_{sex}^{n}) \log(1 - \hat{y}_{sex}^{n}))}_{\substack{\text{preserve sex}\\ \text{information}}} . \tag{5.1}$$

Let $\zeta^* = \underset{\zeta}{\arg\min}\ \mathcal{L}_1$ be optimum value which minimizes $\mathcal{L}_1$. The above problem is the Multi-task learning (MTL) Caruana (1997); Zhang and Yang (2017) problem to learn multiple correlated tasks at the same time. This formulation helps improve the performance of each and reduces the need to introduce multiple models to solve individual tasks. We use the direct sum approach to combine different objectives, a common approach in multi-task learning. We directly minimize the sum of training losses of different tasks and additional constraints. Network architecture is based on the hard parameter sharing strategy in MTL. In this strategy, parameters are shared by the bottom layers among all tasks, while top layers are selected to be task-specific, helping with robustness against overfitting Ruder (2017). It is a commonly used method for designing deep learning models in the literature Long et al. (2017); Ruder et al. (2019); Sener and Koltun (2018). Figure 20 shows MTL

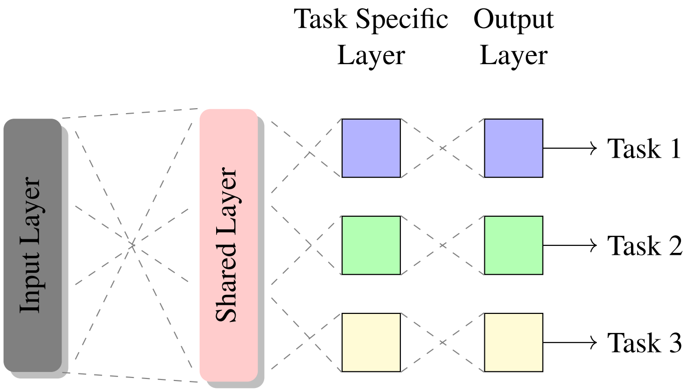


Figure 20: Multi task learning framework

framework used in our problem. The shared layers contain the following layers: a fully connected layer with 20 hidden units, dropout layer with rate 0.2, ReLU, a fully-connected layer with 10 hidden units. Below are the layer detail for each task-

- Age prediction (Task 1)- A fully connected layer with 10 hidden units, ReLU, one output unit
- Site prediction (Task 2)- A fully connected layer with 10 hidden units, ReLU, a softmax layer
- Sex prediction (Task 3)- A fully connected layer with 10 hidden units, ReLU, a softmax layer

Using the loss function $\mathcal{L}_1$, we modify $\mathbf{\Lambda}^n$ such that its predictability power to predict site, age and sex reduces. This can be achieved by maximizing $\mathcal{L}_1$ loss with respect to $\mathbf{\Lambda}^n$. Note that we are trying to solve two problems with one loss function, first is finding optimal ζ which minimizes $\mathcal{L}_1$ and second is finding optimal $\mathbf{\Lambda}^n$ which maximizes $\mathcal{L}_1$. This will result in a minimax game, where the ζ is learned to minimize the cross-entropy and regression loss, and $\mathbf{\Lambda}^n$ is adjusted to maximize the loss. The minimax optimization problem can be written as:

$$\begin{aligned} \max_{\zeta} \min_{\mathbf{W},\mathcal{D},\mathcal{U},\mathbf{V},\mathcal{Z}} \quad & \sum_{s=1}^{S} \sum_{n \in \mathcal{I}_s} \|\mathbf{\Theta}^n - \mathbf{W}\mathbf{Z}^s\mathbf{\Lambda}^n\mathbf{W}^\top - \mathbf{U}^s\mathbf{V}\|_F^2 - \gamma_1 \mathcal{L}_1(\zeta, \mathcal{D}, \mathbf{y}_{site}, \mathbf{y}_{age}, \mathbf{y}_{sex}) \\ s.t. \quad & \mathbf{W} \in \Omega, \quad \mathcal{D} \in \Psi, \quad \|\mathbf{v}_p\|_1 < \mu, \; p = 1, \ldots, P, \end{aligned} \tag{5.2}$$

where $\mathcal{U} = \{\mathbf{U}_s | s = 1, \ldots, S\}$, $\mathcal{Z} = \{\mathbf{Z}_s | s = 1, \ldots, S\}$ and $\mathbf{v}_p$ is the pth column of $\mathbf{V}$. Here $\|\mathbf{\Theta}^n - \mathbf{W}\mathbf{Z}^s\mathbf{\Lambda}^n\mathbf{W}^\top - \mathbf{U}^s\mathbf{V}\|_F^2$ is the total error in the representation of the S subjects and $\mathcal{L}_1$ is the robustness loss, and γ_1 is the tradeoff between representation learning and robustness.

5.2.2 Joint modeling of MDD scores

The aim of the chapter is to learn components which are representative of MDD. For this, we build discriminative rshSCP (dis-rshSCP) to use subject-specific information $\mathbf{\Lambda}_r^n$ to predict whether subject n has MDD or not. Use this model we will subject-specific information which is most predictive of MDD and will give components corresponding to that. We model MDD information using logistic regression framework with parameter $\mathbf{b} \in \mathbb{R}^f$, where $f = \sum_{r=1}^{K} k_r$, subject specific information of all the levels combined $\mathbf{t}^n = diag[\mathbf{\Lambda_1^n}, \ldots, \mathbf{\Lambda_K^n}]$ and the loss function defined below:

$$\begin{aligned} \mathcal{L}_2(\mathbf{b}, \mathcal{D}, \mathbf{y}_{mdd}) = &-\sum_{n=1}^{N} y_{mdd}^n \log\left(\frac{1}{1+\exp\left(-\mathbf{b}^\top \mathbf{t}^n\right)}\right) \\ &- (1 - y_{mdd}^n) \log\left(1 - \frac{1}{1+\exp\left(-\mathbf{b}^\top \mathbf{t}^n\right)}\right). \end{aligned} \tag{5.3}$$

We minimize cross entropy loss $\mathcal{L}_2(\mathbf{b}, \mathcal{D}, \mathbf{y}_{mdd})$ with $\mathbf{b}$ as the parameters to be estimated. Now, let the generative loss be

$$G(\mathcal{C}, \mathcal{W}, \mathcal{D}, \mathcal{U}, \mathbf{V}, \mathcal{Z}) = \sum_{s=1}^{S} \sum_{n \in \mathcal{I}_s} \sum_{r=1}^{K} \|\mathbf{\Theta}^n - (\prod_{j=1}^{r} \mathbf{W}_j) \mathbf{Z}_r^s \mathbf{\Lambda}_r^n (\prod_{j=1}^{r} \mathbf{W}_n)^\top - \mathbf{U}_r^s \mathbf{V}_r\|_F^2, \tag{5.4}$$

which models the components and site information, then the joint optimization problem can be written as:

$$\begin{aligned} \max_{\zeta} \min_{\mathcal{W}, \mathcal{D}, \mathcal{U}, \mathcal{Z}, \mathbf{V}, \mathbf{b}} \quad & \underbrace{G(\mathcal{C}, \mathcal{W}, \mathcal{D}, \mathcal{U}, \mathbf{V}, \mathcal{Z})}_{\substack{\text{learn subject and} \\ \text{site information}}} - \gamma_1 \underbrace{\mathcal{L}_1(\zeta, \mathcal{D}, \mathbf{y}_{site}, \mathbf{y}_{age}, \mathbf{y}_{sex})}_{\substack{\text{reduce age, sex and} \\ \text{site information}}} + \gamma_2 \underbrace{\mathcal{L}_2(\mathbf{b}, \mathcal{D}, \mathbf{y}_{mdd})}_{\substack{\text{preserve MDD} \\ \text{information}}} \\ s.t. \quad & \mathcal{W} \in \Omega, \quad \mathcal{D} \in \Psi, \quad \|\mathbf{v}_p\|_1 < \mu, \; p = 1, \ldots, P. \end{aligned} \tag{5.5}$$

Here, in addition to representation learning and robustness loss, we also have prediction error $\mathcal{L}_2$. γ_1 and γ_2 are the trade-offs between representation learning, robustness, and prediction.

5.2.3 Prediction on unseen data

To estimate $\mathbf{\Lambda}$ for a new subject, we first solve the optimization problem in equation 5.5 to estimate $\mathbf{W}$ computed from the training data. The estimation of the coefficients of unseen subjects are then estimated by solving the below minimization problem where $\mathbf{W}$ is computed from the training data:

$$\begin{aligned} \max_{\zeta} \min_{\mathcal{D},\mathcal{U},\mathcal{Z},\mathbf{V}} \quad & G(\mathcal{C},\mathcal{W},\mathcal{D},\mathcal{U},\mathbf{V},\mathcal{Z}) - \gamma_1 \mathcal{L}_1(\zeta,\mathcal{D},\mathbf{y}_{site},\mathbf{y}_{age},\mathbf{y}_{sex}) \\ s.t. \quad & \mathcal{D} \in \Psi, \quad \|\mathbf{v}_p\|_1 < \mu, \; p = 1,\ldots,P. \end{aligned} \tag{5.6}$$

The estimate for the MDD information for the test subject n is given by:

$$y^n_{mdd} = \begin{cases} 1 & \text{if } \frac{1}{1+\exp\left(-\mathbf{b}^\top \mathbf{t}^n\right)} \geq 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{b}$ and $\mathbf{t}$ is estimated from solving equation 5.5 and 5.6 respectively.

The optimization problems defined in 5.5 and above 5.6 are non-convex problems. We use alternating minimization to solve the optimization procedure.

5.3 Algorithm

5.3.1 Alternating Minimization

We employ the same alternating minimization technique as used in previous chapters for estimating model parameters. Here, we optimize the objective function 5.5 for each variable using adaptive gradient descent (AMSGrad) Reddi et al. (2019) while holding estimates of other variables as constants. β_1 and β_2 value in AMSGrad are kept to be 0.9 and 0.999. The gradients of each variable used in gradient descent is defined in the next section. Algorithm 6 describes the complete alternating minimization procedure to solve equation 5.5. Algorithm 6 can be modified for solving equation 5.6 by commenting out the gradient descent of $\mathbf{b}$. $\mathcal{U}$ and $\mathbf{V}$ are initialized using equation 4.6 $(\mathrm{site-initialization})$ defined in Subsection 4.2.4 and $\mathrm{svd-initialization}$ algorithm 2 in Subsec-

Algorithm 6 dis-rshSCP

```
Input: Data $\mathcal{C}$, number of connectivity patterns $k_1, \dots, k_K$ and sparsity $\lambda_1, \dots, \lambda_K$ at different level, hyperparameters $\mu$, $\gamma_1$ and $\gamma_2$.
Initialize $\mathcal{W}$ and $\mathcal{D}$ using $\mathrm{svd-initialization}$
Initialize $\mathcal{U}$ and $\mathbf{V}$ using $\mathrm{site-initialization}$
repeat
    for $r = 1$ to $K$ do
        if Starting criterion is met then
            $\zeta \leftarrow \mathrm{descent}(\zeta)$
            $\mathbf{b} \leftarrow \mathrm{descent}(\mathbf{b})$
        if $r == 1$ then
            $\mathbf{W}_r \leftarrow \mathrm{proj}_1(\mathbf{W}_r, \lambda_r)$
        else
            $\mathbf{W}_r \leftarrow \mathrm{proj}_2(\mathbf{W}_r)$
        for $n = 1, .., N$ do
            $\mathbf{\Lambda}_r^n \leftarrow \mathrm{descent}(\mathbf{\Lambda}_r^n)$
            $\mathbf{\Lambda}_r^n \leftarrow \mathrm{proj}_2(\mathbf{\Lambda}_r^n)$
        for $s = 1$ to $S$ do
            $\mathbf{U}_r^s \leftarrow \mathrm{descent}(\mathbf{U}_r^s)$
        $\mathbf{V}_r \leftarrow \mathrm{descent}(\mathbf{V}_r)$
        $\mathbf{V}_r \leftarrow \mathrm{proj}_3(\mathbf{V}_r, \mu)$
until Stopping criterion is reached
Output: $\mathcal{W}$, $\mathcal{L}$ and $\mathbf{b}$
```

tion 2.3.5 is used to initialize $\mathcal{W}$ and $\mathcal{D}$. $\mathrm{proj}_1(\mathbf{W}, \lambda)$, proj_2 and proj_3 operators are used directly used from Subsection 2.3.3 .

5.3.2 Gradient Calculations

In this section, we define gradients used for alternating gradient descent, some of which are already defined in previous chapters. Let

$$\tilde{\mathbf{W}}_0 = \mathbf{W}_0 = \mathbf{I}_P, \qquad \mathbf{Y}_r = \prod_{j=0}^{r} \mathbf{W}_j, \qquad \mathbf{Q}_{m,n}^r = (\prod_{j=1}^{m-r} \mathbf{W}_j)\mathbf{Z}_{m-r}^s \mathbf{\Lambda}_{m-r}^n (\prod_{j=1}^{m-r} \mathbf{W}_j)^\top,$$

$$\mathbf{X}_r^n = \mathbf{\Theta}^n - \mathbf{U}_r^s \mathbf{V}_r, \qquad \mathbf{H}_r^n = \mathbf{\Theta}^n - (\prod_{j=1}^{r} \mathbf{W}_j)\mathbf{Z}_{m-r}^s \mathbf{\Lambda}_r^n (\prod_{j=1}^{r} \mathbf{W}_n)^\top,$$

where $n \in \mathcal{I}_s$, $\mathbf{X}_r^n$ stores the information after removing linear site effects from $\mathbf{\Theta}^n$ and $\mathbf{H}_r^n$ stores the information after removing subject-wise and shared component information at the rth level. The gradient of $\mathcal{L}_1$ and $\mathcal{L}_2$ with respect to $\mathcal{D}$ is calculated using automatic differentiation provided by MATLAB. Let J be the objective function defined in 5.5 and then gradient of J with respect to $\mathbf{\Lambda}_r^n$ is:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{\Lambda}_r^n} &= \frac{\partial G(\mathcal{C}, \mathcal{W}, \mathcal{D}, \mathcal{U}, \mathbf{V}, \mathcal{Z})}{\partial \mathbf{\Lambda}_r^n} - \gamma_1 \frac{\partial \mathcal{L}_1(\zeta, \mathcal{D}, \mathbf{y}_{site}, \mathbf{y}_{age}, \mathbf{y}_{sex})}{\partial \mathbf{\Lambda}_r^n} + \gamma_2 \frac{\partial \mathcal{L}_2(\mathbf{b}, \mathcal{D}, \mathbf{y}_{mdd})}{\partial \mathbf{\Lambda}_r^n} \\ &= (-2\mathbf{Y}_r^\top \mathbf{X}_r^n \mathbf{Y}_r + 2\mathbf{Y}_r^\top \mathbf{Y}_r \mathbf{Z}_r^s \mathbf{\Lambda}_r^n \mathbf{Y}_r^\top \mathbf{Y}_r) \circ \mathbf{Z}_r^s + \mathbf{F},\end{aligned}$$

where is $\mathbf{F}$ i.e $-\gamma_1 \frac{\partial \mathcal{L}_1(\zeta, \mathcal{D}, \mathbf{y}_{site}, \mathbf{y}_{age}, \mathbf{y}_{sex})}{\partial \mathbf{\Lambda}_r^n} + \gamma_2 \frac{\partial \mathcal{L}_2(\mathbf{b}, \mathcal{D}, \mathbf{y}_{mdd})}{\partial \mathbf{\Lambda}_r^n}$ is calculated using automatic differentiation. The gradient of J with respect to $\mathbf{W}_r$ is:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{W}_r} &= \frac{\partial G(\mathcal{C}, \mathcal{W}, \mathcal{D}, \mathcal{U}, \mathbf{V}, \mathcal{Z})}{\partial \mathbf{W}_r} \\ &= \sum_{n=1}^{N} \sum_{j=r}^{K} -4\mathbf{Y}_{r-1}^\top \mathbf{X}_n \mathbf{Y}_{r-1} \mathbf{W}_r \mathbf{Q}_{j,n}^r + 4\mathbf{Y}_{r-1}^\top \mathbf{Y}_{r-1} \mathbf{W}_r \mathbf{Q}_{j,n}^r \mathbf{W}_r^\top \mathbf{Y}_{r-1}^\top \mathbf{Y}_{r-1} \mathbf{W}_r \mathbf{Q}_{j,n}^r.\end{aligned}$$

The gradient J with respect to $\mathbf{U}^s$ and $\mathbf{V}$ is:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{U}_r^s} &= \left(\sum_{n=\mathcal{I}_s} (\mathbf{H}_r^n - \mathbf{U}_r^s \mathbf{V}_r) \mathbf{V}_r^\top \right) \circ \mathbf{I}_p \\ \frac{\partial J}{\partial \mathbf{V}_r} &= \sum_{s=1}^{S} \sum_{n \in \mathcal{I}_s} \mathbf{U}_r^s (\mathbf{H}_r^n - \mathbf{U}_r^s \mathbf{V}_r).\end{aligned}$$

The gradient of J with respect to $\mathbf{K}_r^s$ is:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{K}_r^s} &= \frac{\partial G(\mathcal{C}, \mathcal{W}, \mathcal{D}, \mathcal{U}, \mathbf{V}, \mathcal{Z})}{\partial \mathbf{K}_r^s} \\ &= \sum_{n=\mathcal{I}_s} (-2\mathbf{Y}_r^\top \mathbf{X}_r^n \mathbf{Y}_r + 2\mathbf{Y}_r^\top \mathbf{Y}_r \mathbf{Z}_r^s \mathbf{\Lambda}_r^n \mathbf{Y}_r^\top \mathbf{Y}_r) \circ \mathbf{\Lambda}_r^n.\end{aligned}$$

5.4 Materials

5.4.1 Participants

Five worldwide study samples totaling 1657 participants, including 733 with MDD and 924 healthy controls (HC) contributed T1-weighted structural scans and resting-state fMRI data (rs-fMRI) to this study. The included cohorts combine data from the following studies: EMBARC (4 centers across the United States of America, Trivedi et al. (2016)), University of Oxford (United Kingdom, Godlewska et al. (2014, 2018)), Sichuan University Cohort (China, Zhao et al. (2020c)) and STRADL (United Kingdom, Navrady et al. (2018); Stolicyn et al. (2020)). Patient and controls were on average 68.6% (50 − 73.8%) and 54.5% (53 − 70%) female. The mean age across samples was 47 (18 − 78) years in patients and 61 (16 − 84) years for controls. All patients in EMBARC, Oxford and SCU were medication-free, and 15 in SNAP and 170 in STRADL were medicated at the time of scanning and had a primary diagnosis of MDD that was a first episode or recurrent. MDD diagnosis was based on standardized diagnostic criteria: DSM-IV (Oxford) and DSM-IV-TR (EMBARC, Stanford, STRADL, SCU) Frances et al. (1995); First et al. (2004). Table 29 summarizes the number of healthy and MDD participants in each site with their age and sex distribution. We used the same preprocessing pipeline as mentioned in section 4.4.2 of the previous chapter.

5.5 Experiments

5.5.1 Convergence results

Reconstruction error is used to empirically validate the convergence of the Algorithm 6:

$$\frac{\sum_{s=1}^{S}\sum_{n\in\mathcal{I}_s}\sum_{r=1}^{K}||\mathbf{\Theta}^i-(\prod_{j=1}^{r}\mathbf{W}_j)\mathbf{Z}^s\mathbf{\Lambda}_r^i(\prod_{j=1}^{r}\mathbf{W}_j)^\top-\mathbf{U}_r^s\mathbf{V}_r||_F^2}{\sum_{n=1}^{N}K||\mathbf{\Theta}^i||_F^2}.$$

Figure 21 shows the reconstruction loss, training error and test error of the algorithm on one of the folds of 5 fold cross validation. We can see from the result that the algorithm converges after 400 iterations. In the figure, for the first 50 iterations, the algorithm converges without the adversarial and

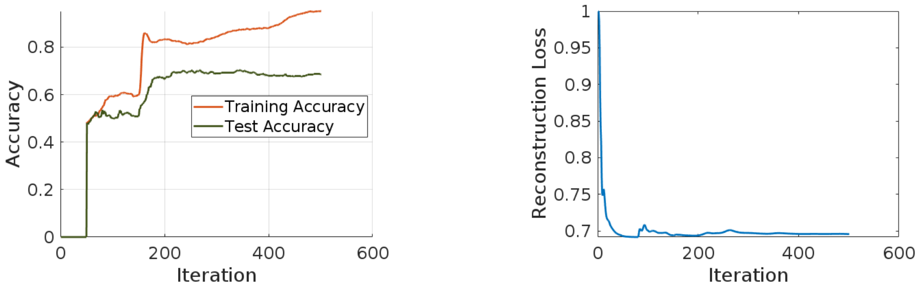
Site	Group	Number	%of F	Age (y)	Medicated	Clinician rating[a]	Comorbid patients	Recurrent
EMBARC-CU	Healthy	12	60.0	$[18, 54](34)$		$0.5(1.0)$		
	Patient	77	68.3	$[18, 64](30)$	0	$18.0(4.1)$	0	76
EMBARC-MG	Healthy	10	66.6	$[18, 65](28)$		$0.6(1.0)$		
	Patient	52	56.2	$[18, 64](28.5)$	0	$19.1(4.0)$	0	52
EMBARC-TX	Healthy	11	50.0	$[23, 57](26)$		$0.6(0.8)$		
	Patient	97	67.7	$[19, 65](44)$	0	$18.6(4.3)$	0	96
EMBARC-UM	Healthy	10	70.0	$[23, 62](41.4)$		$0.7(0.7)$		
	Patient	59	71.1	$[18, 65](31)$	0	$18.5(4.6)$	0	59
Oxford	Healthy	31	58.0	$[19, 58](28)$		$0.4(.8)$		
	Patient	39	60.5	$[20, 61](27)$	0	$22.8(4.4)$	0	14
SCU	Healthy	40	55.0	$[16, 57](26.5)$		N.A.		
	Patient	30	50.0	$[18, 60](30.5)$	0	$22.6(4.5)$	0	N.A.
SNAP	Healthy	55	65.3	$[19, 58](28.8)$		$1.8(2.4)$		
	Patient	55	63.6	$[20, 56](28.2)$	15	$15.1(5.7)$	22	38
STRADL	Healthy	755	53.1	$[26, 84](62)$		$3.5(2.2)$		
	Patient	324	73.8	$[26, 78](60)$	170	$7.0(4.8)$	136	N.A.

Table 29: Demographic characteristics of participants. [a] The 17 item HAMD was used in EMBARC, Oxford and SNAP, the 24 item HAMD was used in SCU and QIDS was used in STRADL

discriminative loss. As the adversarial and discriminative losses are introduced, the reconstruction loss starts to oscillate and then converges to a sub optimal loss as compared to if there were no additional losses were introduced. Whereas the training accuracy keeps on getting better but the test accuracy converges after 200 iteration. This shows that the algorithm can overfit, hence necessary cross-validation is important for selecting the hyperparamters in the loss function 5.5. We see during convergence that the reconstructions loss is a little higher if we hadn't introduced adversarial and discriminative losses. Here, we can expect a trade-off between finding components with optimal reconstruction, adversarial and discriminative loss that depends on γ_1 and γ_2. The optimal value of these hyperparameters is selected using cross-validation, which we explain in the Subsection 5.5.

5.5.2 Evaluating predictive performance

We use two different strategies to evaluate the performance of dis-rshSCP. In the first strategy, we compare five-fold cross-validation accuracy. The goal here is to check how well our model is able to estimate $\mathbf{\Lambda}$ to classify MDD vs. healthy people. We train the model on 80% training set for each fold and test on the remaining. Here training is referred to solving equation 5.5 to estimate $\mathcal{W}$, $\mathcal{D}$, $\mathcal{U}$, $\mathcal{Z}$, $\mathbf{V}$ and $\mathbf{b}$. During the test, we fix $\mathbf{b}$, estimate the rest of the parameters and evaluate the performance of



(a) Training and test accuracy

(b) Reconstruction loss

Figure 21: Convergence of dis-rshSCP algorithm using the complete dataset for $k_1 = 10$ and $k_2 = 4$.

the estimated parameters using $\mathbf{b}$ estimated from training. We use stratification to ensure each fold is representative of all strata (age, sex, site, and MDD) of the data. This is performed to ensure each class is approximately equally represented across each test and training fold. In the second strategy, we compare leave one site out accuracy. We train the model on all sites except one and test the model on the data from the remaining site.

We compare the performance of 4 different versions of the model: 1) standard hSCP without reduction of any covariates, 2) dis-rshCP (demographics) model with reduction of effects of age and sex, 3) dis-rshSCP (site) model with a reduction in site effects and 4) dis-rshSCP (complete) model with a reduction in age, sex and site effects. We fix $k_2 = 4$ for better interpretability and find optimal value of k_1 from the set $\{5, 10, 15, 20\}$. Note that even if we select a large value of k_2, only a few of those components will be used to predict MDD and our experiments show that it is less than 4. Optimal value of hyperparameters μ, γ_1, γ_2 and λ_1 are selected from $[0.1, 1, 5]$, $[.5, 1, 5]$, $[0.1, 1, 5]$ and $[0.1, 1, 10]$.

Table 30 and Table 31 show results of five-fold and leave one site cross-validation. From the results, we can see the baseline method's dull prediction performance, which shows the difficulty of the challenge we are tackling. It can be seen that dis-rshSCP has a better performance than standard hSCP. Out of demographics and site as a covariate, we see that reducing site variability has a major impact on the prediction performance compared to reducing demographics information. The best performance is achieved when we remove both demographics and site information which

Method	$k_1 = 5$	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$
hSCP	0.569 ± 0.024	0.571 ± 0.022	0.576 ± 0.022	0.578 ± 0.019
dis-rshSCP (demogra.)	0.618 ± 0.021	0.626 ± 0.019	0.635 ± 0.014	0.638 ± 0.015
dis-rshSCP (site)	0.667 ± 0.016	0.672 ± 0.020	0.673 ± 0.014	0.687 ± 0.015
dis-rshSCP (complete)	$\mathbf{0.703 \pm 0.018}$	$\mathbf{0.727 \pm 0.015}$	$\mathbf{0.728 \pm 0.016}$	$\mathbf{0.731 \pm 0.013}$

Table 30: Five fold cross validation for $k_2 = 4$ (mean ± standard deviation).

Method	$k_1 = 5$	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$
hSCP	0.579 ± 0.036	0.583 ± 0.029	0.585 ± 0.014	0.593 ± 0.024
dis-rshSCP (demogra.)	0.618 ± 0.038	0.626 ± 0.027	0.635 ± 0.036	0.638 ± 0.032
dis-rshSCP (site)	0.620 ± 0.037	0.632 ± 0.050	0.634 ± 0.038	0.648 ± 0.026
dis-rshSCP (complete)	$\mathbf{0.651 \pm 0.031}$	$\mathbf{0.680 \pm 0.024}$	$\mathbf{0.681 \pm 0.042}$	$\mathbf{0.689 \pm 0.047}$

Table 31: Leave one site accuracy for $k_2 = 4$ (mean ± standard deviation).

suggests that removing heterogeneity from the data can help improve the predictability power of these components, thus giving more reliable components discriminative of MDD.

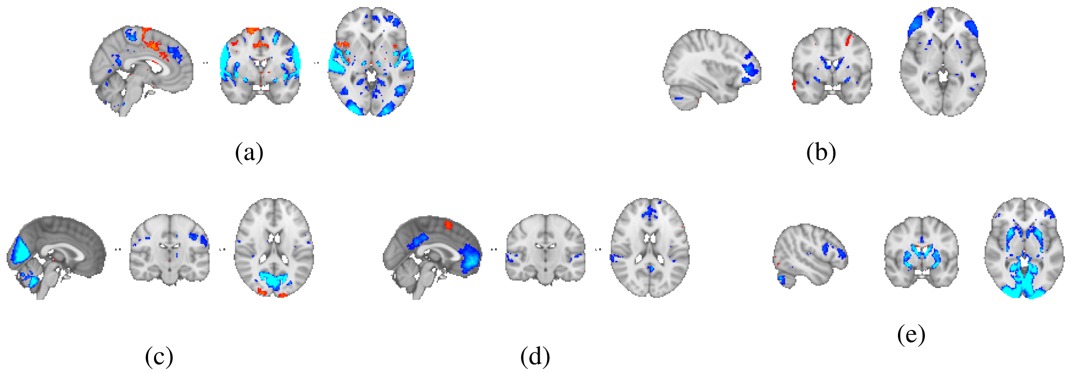
5.5.3 Reproducibility

We have shown that our method can extract components with high predictability power. This section shows that these components are highly reproducible and generalizable, which is important for the future application of our framework. We use split-sample reproducibly to measure the generalizability of the components, which measures how likely a set of components is replicable across the same population. First the optimal parameters are selected based on the highest five fold cross validation, then the split sample reproducibility is computed by dividing the dataset into two random splits with the same stratification and calculating the correlation between components derived from the two splits.

Table 32 shows the split sample reproducibility of the components extracted from the different methods. In all the experiments, results are generated by computing reproducibility over 20 runs. It can be seen that the results are similar to prediction performance results, i.e., reducing demographics and site information helps in improving reproducibility. This suggests even after adding prediction loss, the method can find highly reproducible components.

Method	$k_1 = 5$	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$
hSCP	0.752 ± 0.032	0.713 ± 0.037	0.688 ± 0.039	0.621 ± 0.021
dis-rshSCP (demogra.)	0.781 ± 0.024	0.749 ± 0.045	0.739 ± 0.038	0.688 ± 0.032
dis-rshSCP (site)	0.782 ± 0.031	0.739 ± 0.032	0.728 ± 0.025	0.661 ± 0.028
dis-rshSCP (complete)	$\mathbf{0.805 \pm 0.029}$	$\mathbf{0.761 \pm 0.026}$	$\mathbf{0.725 \pm 0.035}$	$\mathbf{0.692 \pm 0.031}$

Table 32: Split sample reproducibility for $k_2 = 4$ (mean ± standard deviation).


(a)

(b)

(c)

(d)

(e)

Figure 22: (1) SMN, DMN and SN, (2) DMN and CEN, (3) VN, (4) DMN and (5) VN and CEN. We use red and blue colors to represent two different sub-networks which are anti-correlated with each other and regions with a colored part is correlated among themselves.

5.5.4 Analysis of Components

We select k_1 based on optimal tradeoff between classification accuracy and reproducibility, the result of which are provided in Table 30, 31 and 32. We can see from the Table 30 and 31 that classification accuracy starts to plateau at $k_1 = 10$, but if we look at reproducibility, it continues to decrease linearly. Thus we choose k_1 to have good classification accuracy without losing reproducibility of the components. Our aim here is to analyze components that are most predictive of MDD. For this, the components are selected on basis of $\mathbf{b}$ in the logistic regression (5.3). We perform a hypothesis test with the null hypothesis being that the kth component is not discriminative of MDD, i.e., b_k is 0 and p-value < 0.05. If a component has a significant positive value of b_k, then it is likely to have a higher weight in the MDD population than the healthy population and vice versa. Each column of $\mathbf{W}$ contains a component storing information about a set of co-activated regions which can be positively or negatively correlated with each other, and we map normalized values of each column onto corresponding regions. On the basis of hypothesis test, we obtained 5 significant components

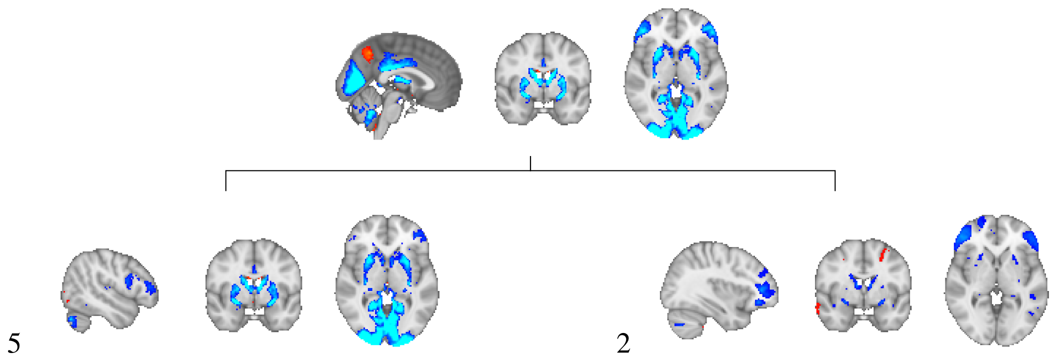
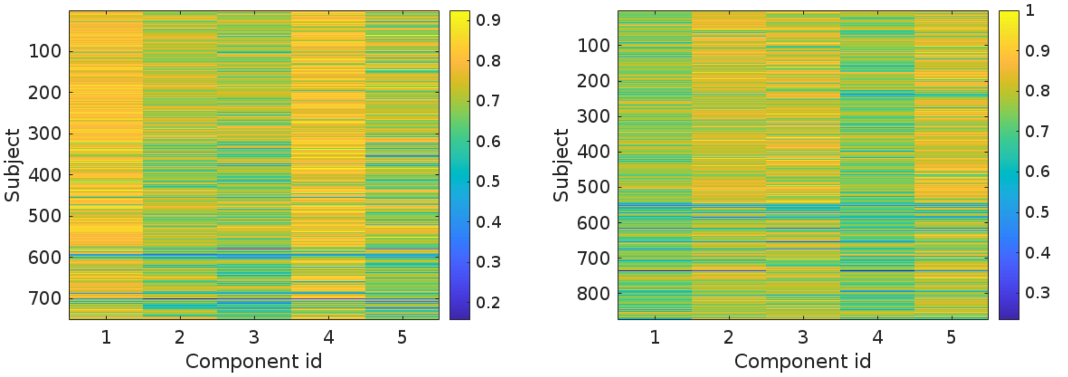


Figure 23: Hierarchical component having significant predictive power. It is comprised of component 5 and 2. Here the coarse-scale component is storing the relation between different fine-scale components comprising DMN, CEN, and VN.

out of 10.

Figure 22 displays components most predictive of MDD. Regions colored blue are anti-correlated with displayed areas in red. The extracted components store information between various parts of the human brain, i.e., whether the regions are correlated or anti-correlated with each other. These regions could be clustered to form one or multiple resting-state functional networks. Component 1 comprises of regions of Somatomotor Network (SMN), Default Mode Network (DMN) and Salience Network (SN), where regions of DMN and SMN are anti-correlated with SN. Component 2 comprises regions of DMN and Central Executive Network (CEN) anti-correlated with each other. Component 3 and 4 consists of regions of DMN and Visual Network (VN), and component 5 consists of regions of VN and CEN positively correlated with each other. In addition to the 5 fine scale significant components, we also recovered a significant hierarchical component comprising of 5 and 2 shown in Figure 23. The strength of each component in MDD and healthy individuals is shown in Figure 24. We observe decreased representation in components 2, 3 and 5 comprising the DMN, CEN and VN in MDD subjects compared to healthy. Increased representation is seen in the components comprising SMN, DMN and SN in MDD subjects compared to healthy.



(a) Heterogeneity captured by MDD subjects (b) Heterogeneity captured by healthy subjects

Figure 24: Heterogeneity captured by components $1-5$. The color represents the strength ($\mathbf{\Lambda}_1$) of each component in each individual. We normalized $\mathbf{\Lambda}_1$ for comparison purposes with 1 being the highest value and 0 lowest.

5.6 Discussion

5.6.1 Method

Our aim was to identify sparse hierarchical connectivity patterns discriminative of MDD and we achieved this with three coupled loss functions: 1) representation learning loss, 2) discriminative loss, and 3) adversarial loss. Our model cleverly exploits the rs-fMRI correlation matrix structure to extract sparse patterns through the representation learning loss. The model also serves as a dimensionality reduction technique helping to extract low-rank sparse decomposition. The classification loss in the model forces the decomposition to extract MDD specific group-level patterns. Notice that there is a slight tradeoff in classification performance at the expense of representation learning loss. We highlight this as it is essential for exploration.

This work provides proof-of-principle analysis focusing on age, sex, and site as diversity factors in the dataset. Adversarial loss helps reduce these factors, resulting in improved reproducibility, generalizability and prediction performance. Our results show that the largest improvement is achieved when accounting for all factors instead of focusing on just one. However, other factors could be considered for future work, such as comorbidity, ethnicity/race, open vs. closed eye during fMRI acquisition, etc. These factors could help further improve pattern stability and classification performance.

Here, we used split-sample reproducibility along with cross-validation accuracy to extract components with the aim to not just have high predictive power but also high reproducibility. We note here that after a certain number of components, the addition of more components does not increase the accuracy. However, the addition of components decreases the reproducibility, which can be attributed to the model capturing noisy components which are not predictive of MDD. Future investigations could look at various other reproducibility measures and classification metrics, and analyze optimality of components in multiple settings to assess the impact on downstream analysis.

Compared to existing seed-based approaches to investigate the disruptions in the functional connectivity, our method could extract patterns with high reproducibility and predictive power. One of the limitations of seed-based approaches is the dependency of results on the size of the seed and their inability to replicate even the consistent results, which might be due to high sensitivity to seed-region selected for analysis (Mulders et al., 2015). Our method is completely data-driven that could extract significant biomarkers by capturing the relation between different functional networks, within networks, and also at multiple scales, as shown in Figure 23. Compared to existing methods for analyzing MDD, our method can capture heterogeneity in the data, as shown in Figure 24, which could uncover promising subtypes of MDD (Drysdale et al., 2017; Grosenick et al., 2019; Dinga et al., 2019). This can help in optimizing the diagnosis and treatment of affected individuals. Although generating subtypes is not the aim of this chapter, future studies could focus on defining subtypes using the subject-specific information ($\boldsymbol{\Lambda}$) in a multi-site dataset.

5.6.2 Components

We identified 5 components that are highly predictive of MDD status. These components store information about the inter- and intra-connectivity within networks and suggest that MDD is characterized by disruptions in the following networks: intra-connectivity in the visual and default mode networks, inter-connectivity between the visual and central executive networks and inter-connectivity between the salience, default mode and central executive network. These discriminative patterns revealed by our framework are consistent with the recent literature on changes in functional connectivity patterns in MDD (Mulders et al., 2015; Luo et al., 2021).

We observed an increase in representation of component 1 storing anti-correlation between DMN and SN, which aligns with the previous finding of functional and structural connectivity disruption between DMN and SN (Mulders et al., 2015; Fang et al., 2012). This can be attributed to an increased response to negative stimuli, common in depressed patients. We also found a decrease in representation of component 2 storing anti-correlation between DMN and CEN. DMN has a role in awareness and directed attention is dominant during default state (Leech and Sharp, 2014), and CEN has a role in cognitive functioning (Corbetta and Shulman, 2002) and is dominant in the executive state. A change in the interaction between them could be a sign of difficulty switching from default to the executive state. (Hamilton et al., 2013). In addition, we found change in representation of SMN and SN in MDD patients aligning with previous research (Sacchet et al., 2016).

Impaired visual perception has been found in patients with MDD. It is considered an important aspect of the disease, whereby there is a positive correlation between the degree of visual disruption and the severity of symptoms (Song et al., 2021), reduced visual network connectivity (Zeng et al., 2012; Veer et al., 2010) and impaired connectivity between VN and DMN (Liu et al., 2020). Studies suggest that perceptual impairments are linked to abnormal cortical processing and disrupted neurotransmitter systems whilst retinal processing remains intact (Nikolaus et al., 2012; Salmela et al., 2021).

The DMN has frequently been implicated in MDD pathophysiology due to its role in producing negative, self-referential, ruminative thoughts (Hamilton et al., 2015). There are previous reports of both hypo- and hyper-connectivity within the DMN in depression (Kaiser et al., 2015; Tozzi et al., 2021) but Liang et al. (2020) suggest that depression is characterized by two subgroups of patients exhibiting opposing dysfunctional DMN connectivity. These inconsistent findings could be due to sample variations in symptom profiles as variability in connectivity within the DMN is positively correlated with levels of ruminative thoughts (Wise et al., 2017) while hypo-connectivity has been associated with symptom severity in recurrent MDD (Yan et al., 2019).

The inconsistent abnormalities of DMN connectivity in depression suggest that it could instead be the interplay between the DMN and other networks that leads to the variety of symptoms observed in depression. Indeed, our patient sample highly expressed abnormal connectivity across three networks, the CEN, DMN and SN, which have been put forward as being part of a triple network model of psychopathology (Menon, 2011). In this model, aberrant saliency attribution within the SN weakens the engagement of the CEN and disengagement of the DMN, leading to cognitive and emotional problems.

Therefore, identifying sparse connectivity patterns is crucial to understanding the interaction between networks that give rise to disease. These findings show that the proposed method can extract meaningful components with high reproducibility and clinical relevance without traditional seed-based methods, which rely too heavily on a priori regions of interest. In a nutshell, these findings could further our understanding of MDD from a functional network perspective.

5.6.3 Future work and limitations

There are several future directions from methodological and clinical perspectives. First, the model robustness can be improved by introducing masking of correlation matrices. It has been shown that masking of features (Devlin et al., 2019) while learning can improve the robustness of the model and its predictability power. Second, instead of reducing age and sex related heterogeneity, one could disentangle components and learn age, sex and MDD specific components. Our method is limited to finding effects of MDD on brain connectivity; in the future, we will use the proposed approach to study the effects of antidepressant medications on brain connectivity in MDD in resting-state or task-based fMRI (Gudayol-Ferré et al., 2015; Brakowski et al., 2017). Another important direction would be to combine structural connectivity information using Diffusion Tensor Imaging (DTI) in our optimization model. Unifying structural, functional and disease information would give a more comprehensive view of neurobiological abnormalities and altered brain functioning and improve MDD diagnosing ability.

Our proposed model has a few weaknesses, adding directions for future work. First, we only consider logistic regression as our classification model, the results of which might be sub-optimal. Our method can be modified to include various classifiers such as Support Vector Machines, Multi Layer Perceptron, etc., and a comparison study can be performed to find the optimal classifier with more emphasis on the classification performances. Here, using SVM like optimization models might be more straightforward than incorporating XGBoost like models. In this study, our focus was on the interpretation of the components; we did not evaluate the reproducibility and classification accuracy for broad values of k_1 and k_2. Future models could benefit from a thorough investigation of this shortcoming. However, this analysis is beyond the scope of this work. Second, the robust to covariate model is a multi-task learning model whose loss weights are manually selected. More advanced techniques such as the Pareto multi task learning model (Lin et al., 2019) and balanced multi task learning framework (Liang and Zhang, 2020) can be used to improve the results. Our study lumps all the MDD patients together and then analyzes the change in functional networks, but previous studies have signaled that it is a highly heterogeneous psychiatric disorder (Hyman, 2008; Miller, 2010). Clustering approaches can be used on $\mathbf{\Lambda}$ to find subtypes of MDD, the analysis of which is beyond the scope of the current work.

5.7 Conclusion

This work presents an effective matrix decomposition strategy to combine rs-fMRI data with clinical information. Our framework is completely data-dependent and makes minimal assumptions about the data. We extended the method of reducing site effects in hSCP by adding additional loss terms for reducing age and sex effects to estimate robust components discriminative of MDD. We added a discriminator to extract components that represent the MDD population. The problem is formulated as a minimax non-convex optimization problem and is solved using adaptive gradient descent. Experimentally, using a pooled dataset from five different sites, we showed that reducing heterogeneity introduced by age, sex, and site could improve the prediction capability of the components, which is validated using fivefold and leave one site out cross-validation. Our framework robustly identifies brain patterns characterizing MDD and provides an understanding of the manifestation of the disorder from a functional networks perspective. Our evaluation on a large multi-site dataset validates the reproducibility and generalizability of the framework. In addition, our model is not limited to MDD and can be easily adapted to other disorders such as ASD, ADHD, etc. Moreover, it can easily incorporate other models outside the medical domain, provided we have access to valid network measures as an input. This greatly broadens the method's applicability to numerous applications from varied fields.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

Extraction of site robust interpretable functional patterns using neuroimaging data is at an early stage. We hope this thesis has not only shed light on using adversarial learning and matrix factorization with functional neuroimaging data but will also lead to further understanding and progress in neuroscience communities.

In this thesis, we first propose a novel technique for the hierarchical extraction of sparse components from connectivity matrices, with application to rsfMRI data. The proposed hSCP method, an extension of SCPs, is a cascaded joint matrix factorization problem where a correlation matrix corresponding to each individual's data is considered an independent observation. This allowed us to model group-level hierarchical patterns and extract the 'strength' of these patterns in individual components, capturing heterogeneity across data. Experimentally we showed that our method is able to find sparse, low-rank hierarchical decomposition, which is highly reproducible across datasets. Importantly, our work provides a method to uncover hierarchical organization in the functioning of the human brain.

We extended the above method and used adversarial learning to enhance the hSCP method by increasing the reproducibility of hierarchical components. Our experimental results based on simulated data show that Adv hSCP can extract components accurately compared to other methods. Results using real-world rsfMRI data demonstrate that adversarial learning can improve the reproducibility of the components. This can help improve the confidence in using hSCP as biomarkers for downstream analysis.

In our third methodological contribution, we developed a method for estimating and reducing site effects in hSCP using adversarial learning. On the real dataset, we show that the proposed method can capture components with a better split sample and leave one site out reproducibility without losing biological interpretability and information. We also show that we can have a noisy estimate of the patterns without reducing site effects, resulting in a misleading downstream analysis. Using our purely data-driven analysis pipeline, we discovered significant patterns of functional connectivity changes associated with the aging brain in a large multi-site dataset.

The adversarial learning based formulation to extract robust patterns can be applied to any rsfMRI study to reduce heterogeneous changes in the functional connectivity. The method is not just limited to hSCP; it can be easily extended to various matrix factorization approaches such as Independent Component Analysis, Non-negative Matrix Factorization, Dictionary Learning, etc., to improve the reproducibility of functional networks/components. Our work not only has broader applicability in terms of methods used for estimation of components but also to different types of neuroscience data, which includes EEG, MEG, etc.

In our last contribution, we extended the above method and applied it to a large multi-site fMRI dataset containing healthy and MDD subjects. We combined the rshSCP with a discriminative optimization problem, extracting patterns predictive of MDD. The advantages of our method were evident from the results; the method provided robust, interpretable, and predictive sparse low dimensional patterns. This is the first study focusing on removing heterogeneity in the data to extract not only predictive but also interpretable and reproducible biomarkers. Our analysis can be easily adapted to other disorders such as ASD, ADHD, etc., and even other models outside the medical domain, provided we have access to valid network measures. This greatly broadens the method's applicability to numerous applications from varied fields.

6.2 Future Work

All the methods described in this thesis and the pre-processing of the data are completely data-driven and can be directly used for functional connectivity analysis in rs-fMRI. These methodological contributions lay the groundwork for many interesting methodological future directions and clinical applications for disease analysis. Our work has some limitations and improvements which are left for future work. Below we mention some of the future directions that can be pursued based on contributions so far:

1. **Uncertainty and consistency:** Recently, many matrix factorization methods have been developed for analyzing fMRI data that are being solved using non-convex optimization algorithms that might converge to local minima (Li et al., 2017; Wu et al., 2021). Due to reliability issues, either researcher starts their algorithm with good initialization to make the complete algorithm deterministic or compute decompositions several times with different initializations to verify the decomposition's reproducibility. Assessing reproducibility in the case of multiple initializations is non-trivial. Previously a clustering-based algorithm has been proposed which clusters decomposition based on their local minimums. They employ graph-based representation of the decompositions and then use clustering to get a low-rank approximation (Van Eyndhoven et al., 2019). Recently, some have used fixed initialization (Trigeorgis et al., 2017) including this work to make their algorithm deterministic. An exciting direction would be to find an efficient algorithm to estimate the most reproducible components. For this, one approach could be to use stochastic gradient descent as an approximate Bayesian inference scheme (Mandt et al., 2017). This would allow us to measure the reliability of the patterns at the convergence.

2. **Dynamic Functional Connectivity:** One major limitation of the current work is that we assumed that the coupling between two regions is static since we only computed a single correlation matrix for each individual using the complete time series. But, recent work has shown that the functional connectivity is not static, and multiple methods have introduced dynamics of the functional connectivity (Chang and Glover, 2010; Hutchison et al., 2013; Warnick et al., 2018; Zhang et al., 2018). The sliding time window technique is a popular method for studying dynamical functional connectivity and computing multiple correlation matrices for each individual, providing a time series of correlation matrices. Cai et al. (2017) proposed sliding time window based dynamic Sparse Connectivity Patterns where they find

multiple SCPs for an individual using the sliding time window technique. The proposed method by Cai et al. (2017) has limitations as they extract dynamics patterns at a single scale and do not address the problem of heterogeneity in multi-site datasets. Our method could be modified to include sliding time window based correlation matrices and extract dynamics patterns robust to site effects.

3. **Multimodal Integration of Functional and Structural Connectivity:** The human brain has been considered an interconnected network with structural pathways and functional signaling as its two key elements. These two elements are distinct but are related and provide complementary viewpoints. Structural connectivity can be extracted using DTI and functional using rs-fMRI studies providing a dual brain representation. Previous research has provided evidence of direct and indirect relationships between functional and structural pathways within the human brain (Skudlarski et al., 2008; Honey et al., 2009; Fukushima et al., 2018). Neuroimaging studies have revealed that there might be direct or indirect anatomical connections mediating the functional connectivity (Bowman et al., 2012; Atasoy et al., 2016). These interesting studies have pivoted clinical research towards multimodal integration to infer brain connectivity reliability and have provided key insights into brain dysfunction in neurological disorders such as Autism (Mueller et al., 2013; Cociu et al., 2017), Schizophrenia (Qureshi et al., 2017a; Li et al., 2020a), and ADHD (Qureshi et al., 2017b). However, hypothesis-driven discovery still faces challenges in this domain due to high data dimensionality, environmental confounds, and considerable inter-individual variability. Our work could be extended to develop joint generative frameworks to overcome the above limitations and extract a multimodal representation of brain connectivity.

4. **Disease classification:** Recent studies have shown that functional connectivity can be an important biomarker for predicting different brain disorders such as epilepsy (Rajpoot et al., 2015; Riaz et al., 2013), schizophrenia (Kumari et al., 2009; Koch et al., 2015), ADHD (Wang et al., 2013; Riaz et al., 2017), Alzheimer's disease (Wee et al., 2012) and Parkinson's disease (Díez-Cirarda et al., 2018; Wu et al., 2009). These disorders can alter the functional

connectivity of the human brain, and extracting these altered connections can help understand the underlying mechanisms of the disorder. Our framework robustly identifies brain patterns characterizing MDD and provides an understanding of the manifestation of the disorder from a functional networks perspective. Our model could be easily adapted to other disorders like ASD, ADHD, etc., and can robustly estimate robust patterns of dysfunction.

5. **Different adversarial learning approach:** There is one major limitation to adversarial learning in the model defined in Chapter 3. The model requires an instance-specific matrix $\mathbf{\Gamma}$ for the learning process, a hyperparameter, and a principled process is required to select the hyperparameter. The method is split into two parts: attack and defense. Instead of this two-step process, a simpler approach motivated by Cai et al. (2021) can be used by learning an adversary matrix $\mathbf{R}$ such that it maximizes the below reconstruction loss function:

$$\sum_{n=1}^{N} \|\mathbf{\Theta}^n + \mathbf{R} - \mathbf{W}_1 \mathbf{\Lambda}^n \mathbf{W}_1^\top\|_F^2,$$

where $\mathbf{R} \in \mathbb{R}^{P \times P}$, $\|\mathbf{R}\|_F^2 < \epsilon$ and $\mathbf{\Theta}^n + \mathbf{R} \succeq 0$. Then the joint optimization problem to learn $\mathbf{W}$ and $\mathbf{\Lambda}^n$ in a adversarial manner can be written as

$$\min_{\mathbf{W}_1, \mathbf{\Lambda}} \max_{\mathbf{R}} \sum_{n=1}^{N} \|\mathbf{\Theta}^n + \mathbf{R} - \mathbf{W}_1 \mathbf{\Lambda}^n \mathbf{W}_1^\top\|_F^2.$$

APPENDICES

Appendix A: Update rules

AMSgrad: Let g_i be the partial derivative of the objective function with respect to the parameter w_i at i^{th} iteration. Let m_i and v_i denote the decaying averages of past and past squared gradients, then the update rule for AMSgrad is defined as:

$$m_i = \beta_1 m_{i-1} + (1-\beta_1)g_i, \qquad v_i = \beta_2 v_{i-1} + (1-\beta_2)g_i^2$$
$$\hat{m}_i = \frac{m_i}{1-\beta_1^i} \qquad \hat{v}_i = \max(\hat{v}_{i-1}, v_i) \qquad w_{i+1} = w_i - \frac{\eta}{\sqrt{\hat{v}_i}+\epsilon}\hat{m}_i,$$

where $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$ and $\eta = 0.1$. β_1 and β_2 are the hyperparameters in the update rules described above. These are typical values for the practical applications Reddi et al. (2019).

NADAM: The update rule for NADAM is defined as:

$$\hat{m}_i = \frac{m_i}{1-\beta_1^i} \qquad \hat{v}_i = \frac{v_i}{1-\beta_2^i} \qquad w_{i+1} = w_i - \frac{\eta}{\sqrt{\hat{v}_i}+\epsilon}\left(\beta_1\hat{m}_i + \frac{(1-\beta_1)g_i}{1-\beta_1^i}\right),$$

where g_i, w_i, m_i and v_i are defined above. The typical values for the practical applications as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\eta = 0.1$.

ADAM: The update rule for ADAM is defined as:

$$\hat{m}_i = \frac{m_i}{1-\beta_1^i} \qquad \hat{v}_i = \frac{v_i}{1-\beta_2^i} \qquad w_{i+1} = w_i - \frac{\eta}{\sqrt{\hat{v}_i}+\epsilon}\hat{m}_i,$$

where g_i, w_i, m_i and v_i are defined above. The typical values for the practical applications as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\eta = 0.1$.

Appendix B: Quality Control

Quality control of the dataset is based on below metrics-

1. Mean Relative (frame-wise) Displacement (MRD): We used MRD calculated by MCFLIRT to quantify head motion Jenkinson (1999). We set a threshold 0.2mm.

2. Time course Signal to Noise Ratio (tSNR): tSNR is an important metric for evaluating the ability of the fMRI acquisition to detect neural signal changes in the time series. It is is defined as the ratio of mean intensity and standard deviation across time within the evaluated Region of Interest Triantafyllou et al. (2005). We excluded the subjects having temporal SNR less than 100.

3. Framewise Displacement (FD): It evaluates the head motion of each volume compared to the previous volume Power et al. (2012); Jenkinson et al. (2002). We set the threshold for FD to be 0.2 mm Power et al. (2012); Yan et al. (2013).

BIBLIOGRAPHY

Abraham, Alexandre; Milham, Michael P; Di Martino, Adriana; Craddock, R Cameron; Samaras, Dimitris; Thirion, Bertrand, and Varoquaux, Gael. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017.

Achard, Sophie and Bullmore, Ed. Efficiency and cost of economical brain functional networks. *PLoS computational biology*, 3(2):e17, 2007.

Adhikari, Bhim M; Jahanshad, Neda; Shukla, Dinesh; Turner, Jessica; Grotegerd, Dominik; Dannlowski, Udo; Kugel, Harald; Engelen, Jennifer; Dietsche, Bruno; Krug, Axel, and others, . A resting state fmri analysis pipeline for pooling inference across diverse cohorts: an enigma rs-fmri protocol. *Brain imaging and behavior*, 13(5):1453–1467, 2019.

Aharon, Michal; Elad, Michael, and Bruckstein, Alfred. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54 (11):4311–4322, 2006.

Akhavan Aghdam, Maryam; Sharifi, Arash, and Pedram, Mir Mohsen. Combination of rs-fmri and smri data to discriminate autism spectrum disorders in young children using deep belief network. *Journal of digital imaging*, 31(6):895–903, 2018.

Akiki, Teddy J and Abdallah, Chadi G. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific reports*, 9(1):1–15, 2019.

Akyildiz, Ian F.; Wang, Xudong, and Wang, Weilin. Wireless mesh networks: a survey. *Computer Networks*, 47(4):445–487, 2005. ISSN 1389-1286. doi: https://doi.org/10.1016/j.comnet.2004.12.001. URL https://www.sciencedirect.com/science/article/pii/S1389128604003457.

Al-Sharoa, Esraa; Al-Khassaweneh, Mahmood, and Aviyente, Selin. Tensor based temporal and multilayer community detection for studying brain dynamics during resting state fmri. *IEEE Transactions on Biomedical Engineering*, 66(3):695–709, 2018.

Alexander, Lindsay M; Escalera, Jasmine; Ai, Lei; Andreotti, Charissa; Febre, Karina; Mangone, Alexander; Vega-Potler, Natan; Langer, Nicolas; Alexander, Alexis; Kovacs, Meagan, and others, . An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1):1–26, 2017.

Alfaro-Almagro, Fidel; McCarthy, Paul; Afyouni, Soroosh; Andersson, Jesper LR; Bastiani, Matteo; Miller, Karla L; Nichols, Thomas E, and Smith, Stephen M. Confound modelling in uk biobank brain imaging. *NeuroImage*, 224:117002, 2021.

Allen, Elena A; Damaraju, Eswar; Plis, Sergey M; Erhardt, Erik B; Eichele, Tom, and Calhoun, Vince D. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, 24(3): 663–676, 2014.

Anderson, Ariana; Douglas, Pamela K; Kerr, Wesley T; Haynes, Virginia S; Yuille, Alan L; Xie, Jianwen; Wu, Ying Nian; Brown, Jesse A, and Cohen, Mark S. Non-negative matrix factorization of multimodal mri, fmri and phenotypic data reveals differential changes in default mode subnetworks in adhd. *NeuroImage*, 102:207–219, 2014.

Arbabshirani, Mohammad R; Plis, Sergey; Sui, Jing, and Calhoun, Vince D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, 145:137–165, 2017.

Armstrong, Nicole M; An, Yang; Beason-Held, Lori; Doshi, Jimit; Erus, Guray; Ferrucci, Luigi; Davatzikos, Christos, and Resnick, Susan M. Predictors of neurodegeneration differ between cognitively normal and subsequently impaired older adults. *Neurobiology of aging*, 75:178–186, 2019.

Ashourvan, Arian; Telesford, Qawi K; Verstynen, Timothy; Vettel, Jean M, and Bassett, Danielle S. Multi-scale detection of hierarchical community architecture in structural and functional brain networks. *PLoS One*, 14(5):e0215520, 2019a.

Ashourvan, Arian; Telesford, Qawi K; Verstynen, Timothy; Vettel, Jean M, and Bassett, Danielle S. Multi-scale detection of hierarchical community architecture in structural and functional brain networks. *PLoS One*, 14(5):e0215520, 2019b.

Atasoy, Selen; Donnelly, Isaac, and Pearson, Joel. Human brain networks function in connectome-specific harmonic waves. *Nature communications*, 7(1):1–10, 2016.

Bassett, Danielle S; Greenfield, Daniel L; Meyer-Lindenberg, Andreas; Weinberger, Daniel R; Moore, Simon W, and Bullmore, Edward T. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS comput biol*, 6 (4):e1000748, 2010.

Bayer, Johanna MM; Dinga, Richard; Kia, Seyed Mostafa; Kottaram, Akhil R; Wolfers, Thomas; Lv, Jinglei; Zalesky, Andrew; Schmaal, Lianne, and Marquand, Andre. Accommodating site variation in neuroimaging data using hierarchical and bayesian models. *bioRxiv*, 2021.

Beckmann, Christian F; DeLuca, Marilena; Devlin, Joseph T, and Smith, Stephen M. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1001–1013, 2005.

Beldzik, Ewa; Domagalik, Aleksandra; Daselaar, Sander; Fafrowicz, Magdalena; Froncisz, Wojciech; Oginska, Halszka, and Marek, Tadeusz. Contributive sources analysis: a measure of neural networks' contribution to brain activations. *Neuroimage*, 76:304–312, 2013.

Belmaker, Robert H and Agam, Galila. Major depressive disorder. *New England Journal of Medicine*, 358(1):55–68, 2008.

Benkarim, Oualid; Paquola, Casey; Park, Bo-yong; Kebets, Valeria; Hong, Seok-Jun; de Wael, Reinder Vos; Zhang, Shaoshi; Yeo, BT Thomas; Eickenberg, Michael; Ge, Tian, and others, . The cost of untracked diversity in brain-imaging prediction. *bioRxiv*, 2021.

Betzel, Richard F and Bassett, Danielle S. Multi-scale brain networks. *Neuroimage*, 160:73–83, 2017.

Betzel, Richard F; Mišić, Bratislav; He, Ye; Rumschlag, Jeffrey; Zuo, Xi-Nian, and Sporns, Olaf. Functional brain modules reconfigure at multiple scales across the human lifespan. *arXiv preprint arXiv:1510.08045*, 2015.

Bhaumik, Runa; Jenkins, Lisanne M; Gowins, Jennifer R; Jacobs, Rachel H; Barba, Alyssa; Bhaumik, Dulal K, and Langenecker, Scott A. Multivariate pattern analysis strategies in detection of remitted major depressive disorder using resting state functional connectivity. *NeuroImage: Clinical*, 16: 390–398, 2017.

Biswal, Bharat; Zerrin Yetkin, F; Haughton, Victor M, and Hyde, James S. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

Biswal, Bharat B; Mennes, Maarten; Zuo, Xi-Nian; Gohel, Suril; Kelly, Clare; Smith, Steve M; Beckmann, Christian F; Adelstein, Jonathan S; Buckner, Randy L; Colcombe, Stan, and others, . Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

Bowman, F DuBois; Zhang, Lijun; Derado, Gordana, and Chen, Shuo. Determining functional connectivity using fmri data with diffusion-based anatomical weighting. *NeuroImage*, 62(3): 1769–1779, 2012.

Brakowski, Janis; Spinelli, Simona; Dörig, Nadja; Bosch, Oliver Gero; Manoliu, Andrei; Holtforth, Martin Grosse, and Seifritz, Erich. Resting state brain network function in major depression–depression symptomatology, antidepressant treatment effects, future research. *Journal of Psychiatric Research*, 92:147–159, 2017.

Buckner, Randy L; Andrews-Hanna, Jessica R, and Schacter, Daniel L. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124 (1):1–38, 2008.

Buckner, Randy L; Krienen, Fenna M, and Yeo, BT Thomas. Opportunities and limitations of intrinsic functional connectivity mri. *Nature neuroscience*, 16(7):832–837, 2013.

Bush, George; Valera, Eve M, and Seidman, Larry J. Functional neuroimaging of attention-deficit/hyperactivity disorder: a review and suggested future directions. *Biological psychiatry*, 57 (11):1273–1284, 2005.

Button, Katherine S; Ioannidis, John; Mokrysz, Claire; Nosek, Brian A; Flint, Jonathan; Robinson, Emma SJ, and Munafò, Marcus R. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376, 2013.

Cai, Biao; Zille, Pascal; Stephen, Julia M; Wilson, Tony W; Calhoun, Vince D, and Wang, Yu Ping.

Estimation of dynamic sparse connectivity patterns from resting state fmri. *IEEE transactions on medical imaging*, 37(5):1224–1234, 2017.

Cai, Ting; Tan, Vincent YF, and Févotte, Cédric. Adversarially-trained nonnegative matrix factorization. *IEEE Signal Processing Letters*, 28:1415–1419, 2021.

Calhoun, Vince D; Adali, Tulay; Hansen, Lars Kai; Larsen, Jan, and Pekar, James J. Ica of functional mri data: an overview. In *in Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation*. Citeseer, 2003.

Calhoun, Vince D; Liu, Jingyu, and Adalı, Tülay. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1):S163–S172, 2009.

Caruana, Rich. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Casey, BJ; Cannonier, Tariq; Conley, May I; Cohen, Alexandra O; Barch, Deanna M; Heitzeg, Mary M; Soules, Mary E; Teslovich, Theresa; Dellarco, Danielle V; Garavan, Hugh, and others, . The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

Chang, Catie and Glover, Gary H. Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage*, 50(1):81–98, 2010.

Chaudhuri, Rishidev; Bernacchia, Alberto, and Wang, Xiao-Jing. A diversity of localized timescales in network activity. *elife*, 3:e01239, 2014.

Ciric, Rastko; Wolf, Daniel H; Power, Jonathan D; Roalf, David R; Baum, Graham L; Ruparel, Kosha; Shinohara, Russell T; Elliott, Mark A; Eickhoff, Simon B; Davatzikos, Christos, and others, . Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage*, 154:174–187, 2017.

Cociu, Bogdan Alexandru; Das, Saptarshi; Billeci, Lucia; Jamal, Wasifa; Maharatna, Koushik; Calderoni, Sara; Narzisi, Antonio, and Muratori, Filippo. Multimodal functional and structural brain connectivity analysis in autism: a preliminary integrated approach with eeg, fmri, and dti. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):213–226, 2017.

Cook, Ian A; Bookheimer, Susan Y; Mickes, Laura; Leuchter, Andrew F, and Kumar, Anand. Aging and brain activation with working memory tasks: an fmri study of connectivity. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 22(4): 332–342, 2007.

Corbetta, Maurizio and Shulman, Gordon L. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

Craddock, R Cameron; Holtzheimer III, Paul E; Hu, Xiaoping P, and Mayberg, Helen S. Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(6):1619–1628, 2009.

Crossley, Nicolas A; Mechelli, Andrea; Scott, Jessica; Carletti, Francesco; Fox, Peter T; McGuire, Philip, and Bullmore, Edward T. The hubs of the human connectome are generally implicated in the anatomy of brain disorders. *Brain*, 137(8):2382–2395, 2014.

Damoiseaux, Jessica S. Effects of aging on functional and structural brain connectivity. *Neuroimage*, 160:32–40, 2017.

Damoiseaux, Jessica S; Rombouts, SARB; Barkhof, Frederik; Scheltens, Philip; Stam, Cornelis J; Smith, Stephen M, and Beckmann, Christian F. Consistent resting-state networks across healthy subjects. *Proceedings of the national academy of sciences*, 103(37):13848–13853, 2006.

Dansereau, Christian; Benhajali, Yassine; Risterucci, Celine; Pich, Emilio Merlo; Orban, Pierre; Arnold, Douglas, and Bellec, Pierre. Statistical power and prediction accuracy in multisite resting-state fmri connectivity. *Neuroimage*, 149:220–232, 2017.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Di Martino, Adriana; Yan, Chao-Gan; Li, Qingyang; Denio, Erin; Castellanos, Francisco X; Alaerts, Kaat; Anderson, Jeffrey S; Assaf, Michal; Bookheimer, Susan Y; Dapretto, Mirella, and others, . The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Di Martino, Adriana; O'connor, David; Chen, Bosi; Alaerts, Kaat; Anderson, Jeffrey S; Assaf, Michal; Balsters, Joshua H; Baxter, Leslie; Beggiato, Anita; Bernaerts, Sylvie, and others, . Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, 4(1):1–15, 2017.

Díez-Cirarda, María; Strafella, Antonio P; Kim, Jinhee; Peña, Javier; Ojeda, Natalia; Cabrera-Zubizarreta, Alberto, and Ibarretxe-Bilbao, Naroa. Dynamic functional connectivity in parkinson's disease patients with mild cognitive impairment and normal cognition. *NeuroImage: Clinical*, 17: 847–855, 2018.

Dinga, Richard; Schmaal, Lianne; Penninx, Brenda WJH; van Tol, Marie Jose; Veltman, Dick J; van Velzen, Laura; Mennes, Maarten; van der Wee, Nic JA, and Marquand, Andre F. Evaluating the evidence for biotypes of depression: Methodological replication and extension of. *NeuroImage: Clinical*, 22:101796, 2019.

Dong, Qinglin; Ge, Fangfei; Ning, Qiang; Zhao, Yu; Lv, Jinglei; Huang, Heng; Yuan, Jing; Jiang, Xi; Shen, Dinggang, and Liu, Tianming. Modeling hierarchical brain networks via volumetric sparse deep belief network. *IEEE Transactions on Biomedical Engineering*, 67(6):1739–1748, 2019.

Doshi, Jimit; Erus, Guray; Ou, Yangming; Resnick, Susan M; Gur, Ruben C; Gur, Raquel E; Satterthwaite, Theodore D; Furth, Susan; Davatzikos, Christos; Initiative, Alzheimer's Neuroimaging, and others, . Muse: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *Neuroimage*, 127:186–195, 2016.

Doucet, Gaëlle; Naveau, Mikaël; Petit, Laurent; Delcroix, Nicolas; Zago, Laure; Crivello, Fabrice; Jobard, Gael; Tzourio-Mazoyer, Nathalie; Mazoyer, Bernard; Mellet, Emmanuel, and others, . Brain activity at rest: a multiscale hierarchical functional organization. *Journal of neurophysiology*, 105(6):2753–2763, 2011.

Dozat, Timothy. Incorporating nesterov momentum into adam. 2016.

Drevets, Wayne C and Raichle, Marcus E. Reciprocal suppression of regional cerebral blood flow during emotional versus higher cognitive processes: Implications for interactions between emotion and cognition. *Cognition and emotion*, 12(3):353–385, 1998.

Drysdale, Andrew T; Grosenick, Logan; Downar, Jonathan; Dunlop, Katharine; Mansouri, Farrokh; Meng, Yue; Fetcho, Robert N; Zebley, Benjamin; Oathes, Desmond J; Etkin, Amit, and others, . Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1):28–38, 2017.

Du, Yuhui and Fan, Yong. Group information guided ica for fmri data analysis. *Neuroimage*, 69: 157–197, 2013.

Duncan, Niall W and Northoff, Georg. Overview of potential procedural and participant-related confounds for neuroimaging of the resting state. *Journal of psychiatry & neuroscience: JPN*, 38 (2):84, 2013.

D'Souza, Niharika Shimona; Nebel, Mary Beth; Wymbs, Nicholas; Mostofsky, Stewart H, and Venkataraman, Archana. A joint network optimization framework to predict clinical severity from resting state functional mri data. *NeuroImage*, 206:116314, 2020.

Eavani, Harini; Satterthwaite, Theodore D; Gur, Raquel E; Gur, Ruben C, and Davatzikos, Christos. Discriminative sparse connectivity patterns for classification of fmri data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 193–200. Springer, 2014.

Eavani, Harini; Satterthwaite, Theodore D; Filipovych, Roman; Gur, Raquel E; Gur, Ruben C, and Davatzikos, Christos. Identifying sparse connectivity patterns in the brain using resting-state fmri. *Neuroimage*, 105:286–299, 2015a.

Eavani, Harini; Satterthwaite, Theodore D; Filipovych, Roman; Gur, Raquel E; Gur, Ruben C, and Davatzikos, Christos. Identifying sparse connectivity patterns in the brain using resting-state fmri. *Neuroimage*, 105:286–299, 2015b.

Fang, Peng; Zeng, Ling-Li; Shen, Hui; Wang, Lubin; Li, Baojuan; Liu, Li, and Hu, Dewen. Increased cortical-limbic anatomical network connectivity in major depression revealed by diffusion tensor imaging. 2012.

Farnia, Farzan; Zhang, Jesse, and Tse, David. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations*, 2018.

Ferrarini, Luca; Veer, Ilya M; Baerends, Evelinda; van Tol, Marie-José; Renken, Remco J; van der Wee, Nic JA; Veltman, Dirk J; Aleman, Andre; Zitman, Frans G; Penninx, Brenda WJH, and others, . Hierarchical functional modularity in the resting-state human brain. *Human brain mapping*, 30 (7):2220–2231, 2009.

Ferreira, Luiz Kobuti; Regina, Ana Carolina Brocanello; Kovacevic, Natasa; Martin, Maria da Graça Morais; Santos, Pedro Paim; Carneiro, Camila de Godoi; Kerr, Daniel Shikanai; Amaro Jr, Edson; McIntosh, Anthony Randal, and Busatto, Geraldo F. Aging effects on whole-brain functional connectivity in adults free of cognitive and psychiatric disorders. *Cerebral cortex*, 26 (9):3851–3865, 2016.

First, Michael B; France, Allen, and Pincus, Harold Alan. *DSM-IV-TR guidebook.* American Psychiatric Publishing, Inc., 2004.

Fitzpatrick, Annette L; Buchanan, Catherine K; Nahin, Richard L; DeKosky, Steven T; Atkinson, Hal H; Carlson, Michelle C, and Williamson, Jeff D. Associations of gait speed and other measures of physical function with cognition in a healthy cohort of elderly persons. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 62(11):1244–1251, 2007.

Fornito, Alex; Zalesky, Andrew, and Bullmore, Edward. *Fundamentals of brain network analysis*. Academic Press, 2016.

Fortin, Jean-Philippe; Parker, Drew; Tunç, Birkan; Watanabe, Takanori; Elliott, Mark A; Ruparel, Kosha; Roalf, David R; Satterthwaite, Theodore D; Gur, Ruben C; Gur, Raquel E, and others, . Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, 2017.

Fortin, Jean-Philippe; Cullen, Nicholas; Sheline, Yvette I; Taylor, Warren D; Aselcioglu, Irem; Cook, Philip A; Adams, Phil; Cooper, Crystal; Fava, Maurizio; McGrath, Patrick J, and others, . Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167: 104–120, 2018.

Fox, Michael D and Raichle, Marcus E. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*, 8(9):700–711, 2007.

Fox, Michael D; Snyder, Abraham Z; Vincent, Justin L; Corbetta, Maurizio; Van Essen, David C, and Raichle, Marcus E. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005.

Frances, Allen; First, Michael B, and Pincus, Harold Alan. *DSM-IV guidebook.* American Psychiatric Association, 1995.

Frick, Andreas; Gingnell, Malin; Marquand, Andre F; Howner, Katarina; Fischer, Håkan; Kristiansson, Marianne; Williams, Steven CR; Fredrikson, Mats, and Furmark, Tomas. Classifying social anxiety disorder using multivoxel pattern analyses of brain function and structure. *Behavioural brain research*, 259:330–335, 2014.

Friedman, Gary D; Cutter, Gary R; Donahue, Richard P; Hughes, Glenn H; Hulley, Stephen B; Jacobs Jr, David R; Liu, Kiang, and Savage, Peter J. Cardia: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology*, 41(11):1105–1116, 1988.

Fukushima, Makoto; Betzel, Richard F; He, Ye; van den Heuvel, Martijn P; Zuo, Xi-Nian, and Sporns, Olaf. Structure–function relationships during segregated and integrated network states of human brain functional connectivity. *Brain Structure and Function*, 223(3):1091–1106, 2018.

Geerligs, Linda; Maurits, Natasha M; Renken, Remco J, and Lorist, Monicque M. Reduced specificity of functional connectivity in the aging brain during task performance. *Human brain mapping*, 35 (1):319–330, 2014.

Geerligs, Linda; Renken, Remco J; Saliasi, Emi; Maurits, Natasha M, and Lorist, Monicque M. A brain-wide study of age-related changes in functional connectivity. *Cerebral cortex*, 25(7): 1987–1999, 2015.

Geranmayeh, Fatemeh; Wise, Richard JS; Mehta, Amrish, and Leech, Robert. Overlapping networks engaged during spoken language production and its cognitive control. *Journal of Neuroscience*, 34(26):8728–8740, 2014.

Gholami, Behnam; Sahu, Pritish; Rudovic, Ognjen; Bousmalis, Konstantinos, and Pavlovic, Vladimir. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29:3993–4002, 2020.

Gilbert, Charles D and Li, Wu. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013.

Glasser, Matthew F; Sotiropoulos, Stamatios N; Wilson, J Anthony; Coalson, Timothy S; Fischl, Bruce; Andersson, Jesper L; Xu, Junqian; Jbabdi, Saad; Webster, Matthew; Polimeni, Jonathan R, and others, . The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

Glasser, Matthew F; Coalson, Timothy S; Robinson, Emma C; Hacker, Carl D; Harwell, John; Yacoub, Essa; Ugurbil, Kamil; Andersson, Jesper; Beckmann, Christian F; Jenkinson, Mark, and others, . A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Godlewska, Beata R; Hasselmann, Helge WW; Igoumenou, Artemis; Norbury, Ray, and Cowen, Philip J. Short-term escitalopram treatment and hippocampal volume. *Psychopharmacology*, 231 (23):4579–4581, 2014.

Godlewska, Beata R; Browning, Michael; Norbury, Ray; Igoumenou, Artemis; Cowen, Philip J, and Harmer, Catherine J. Predicting treatment response in depression: the role of anterior cingulate cortex. *International Journal of Neuropsychopharmacology*, 21(11):988–996, 2018.

Gong, Qiyong and He, Yong. Depression, neuroimaging and connectomics: a selective overview. *Biological psychiatry*, 77(3):223–235, 2015.

Goodfellow, Ian J; Shlens, Jonathon, and Szegedy, Christian. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Goulas, Alexandros; Schaefer, Alexander, and Margulies, Daniel S. The strength of weak connections in the macaque cortico-cortical network. *Brain Structure and Function*, 220(5):2939–2951, 2015.

Grady, Cheryl; Sarraf, Saman; Saverino, Cristina, and Campbell, Karen. Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks. *Neurobiology of aging*, 41:159–172, 2016.

Greicius, Michael D; Flores, Benjamin H; Menon, Vinod; Glover, Gary H; Solvason, Hugh B; Kenna, Heather; Reiss, Allan L, and Schatzberg, Alan F. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological psychiatry*, 62(5):429–437, 2007.

Griffanti, Ludovica; Salimi-Khorshidi, Gholamreza; Beckmann, Christian F; Auerbach, Edward J; Douaud, Gwenaëlle; Sexton, Claire E; Zsoldos, Enikő; Ebmeier, Klaus P; Filippini, Nicola; Mackay, Clare E, and others, . Ica-based artefact removal and accelerated fmri acquisition for improved resting state network imaging. *Neuroimage*, 95:232–247, 2014.

Grosenick, Logan; Shi, Tracey C; Gunning, Faith M; Dubin, Marc J; Downar, Jonathan, and Liston, Conor. Functional and optogenetic approaches to discovering stable subtype-specific circuit mechanisms in depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4 (6):554–566, 2019.

Gudayol-Ferré, Esteve; Peró-Cebollero, Maribel; González-Garrido, Andrés A, and Guàrdia-Olmos, Joan. Changes in brain connectivity related to the treatment of depression measured through fmri: a systematic review. *Frontiers in human neuroscience*, 9:582, 2015.

Guye, Maxime; Bettus, Gaelle; Bartolomei, Fabrice, and Cozzone, Patrick J. Graph theoretical analysis of structural and functional connectivity mri in normal and pathological brain networks. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23(5):409–421, 2010.

Hamdi, Shah Muhammad; Wu, Yubao; Boubrahimi, Soukaina Filali; Angryk, Rafal; Krishnamurthy, Lisa Crystal, and Morris, Robin. Tensor decomposition for neurodevelopmental disorder prediction. In *International Conference on Brain Informatics*, pages 339–348. Springer, 2018.

Hamilton, J Paul; Chen, Michael C, and Gotlib, Ian H. Neural systems approaches to understanding major depressive disorder: an intrinsic functional organization perspective. *Neurobiology of disease*, 52:4–11, 2013.

Hamilton, J Paul; Farmer, Madison; Fogelman, Phoebe, and Gotlib, Ian H. Depressive rumination, the default-mode network, and the dark matter of clinical neuroscience. *Biological psychiatry*, 78 (4):224–230, 2015.

He, Xiangnan; He, Zhankui; Du, Xiaoyu, and Chua, Tat-Seng. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 355–364, 2018.

He, Ye; Byrge, Lisa, and Kennedy, Daniel P. Nonreplication of functional connectivity differences in autism spectrum disorder across multiple sites and denoising strategies. *Human Brain Mapping*, 41(5):1334–1350, 2020.

Heinsfeld, Anibal Sólon; Franco, Alexandre Rosa; Craddock, R Cameron; Buchweitz, Augusto, and Meneguzzi, Felipe. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23, 2018.

Hirabayashi, Toshiyuki; Takeuchi, Daigo; Tamura, Keita, and Miyashita, Yasushi. Functional microcircuit recruited during retrieval of object association memory in monkey perirhinal cortex. *Neuron*, 77(1):192–203, 2013.

Honey, Christopher J; Sporns, Olaf; Cammoun, Leila; Gigandet, Xavier; Thiran, Jean-Philippe; Meuli, Reto, and Hagmann, Patric. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009.

Hong, Seok-Jun; Valk, Sofie L; Di Martino, Adriana; Milham, Michael P, and Bernhardt, Boris C. Multidimensional neuroanatomical subtyping of autism spectrum disorder. *Cerebral Cortex*, 28 (10):3578–3588, 2018.

Horovitz, Silvina G; Fukunaga, Masaki; de Zwart, Jacco A; van Gelderen, Peter; Fulton, Susan C; Balkin, Thomas J, and Duyn, Jeff H. Low frequency bold fluctuations during resting wakefulness and light sleep: A simultaneous eeg-fmri study. *Human brain mapping*, 29(6):671–682, 2008.

Hu, Xintao; Huang, Heng; Peng, Bo; Han, Junwei; Liu, Nian; Lv, Jinglei; Guo, Lei; Guo, Christine, and Liu, Tianming. Latent source mining in fmri via restricted boltzmann machine. *Human brain mapping*, 39(6):2368–2380, 2018.

Huang, Heng; Hu, Xintao; Zhao, Yu; Makkie, Milad; Dong, Qinglin; Zhao, Shijie; Guo, Lei, and Liu, Tianming. Modeling task fmri data via deep convolutional autoencoder. *IEEE transactions on medical imaging*, 37(7):1551–1561, 2017.

Hutchison, R Matthew; Womelsdorf, Thilo; Allen, Elena A; Bandettini, Peter A; Calhoun, Vince D; Corbetta, Maurizio; Della Penna, Stefania; Duyn, Jeff H; Glover, Gary H; Gonzalez-Castillo, Javier, and others, . Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378, 2013.

Hyman, Steven E. A glimmer of light for neuropsychiatric disorders. *Nature*, 455(7215):890, 2008.

Iwabuchi, Sarina J; Krishnadas, Rajeev; Li, Chunbo; Auer, Dorothee P; Radua, Joaquim, and Palaniyappan, Lena. Localized connectivity in depression: a meta-analysis of resting state functional imaging studies. *Neuroscience & Biobehavioral Reviews*, 51:77–86, 2015.

Jenkinson, Mark. Measuring transformation error by rms deviation. *Studholme, C., Hill, DLG, Hawkes, DJ*, 1999.

Jenkinson, Mark; Bannister, Peter; Brady, Michael, and Smith, Stephen. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17 (2):825–841, 2002.

Jenkinson, Mark; Beckmann, Christian F; Behrens, Timothy EJ; Woolrich, Mark W, and Smith, Stephen M. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Johnson, W Evan; Li, Cheng, and Rabinovic, Ariel. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Jovicich, Jorge; Minati, Ludovico; Marizzoni, Moira; Marchitelli, Rocco; Sala-Llonch, Roser; Bartrés-Faz, David; Arnold, Jennifer; Benninghoff, Jens; Fiedler, Ute; Roccatagliata, Luca, and others, . Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: a multicentric resting-state fmri study. *Neuroimage*, 124:442–454, 2016.

Kaiser, Roselinde H; Andrews-Hanna, Jessica R; Wager, Tor D, and Pizzagalli, Diego A. Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity. *JAMA psychiatry*, 72(6):603–611, 2015.

Karahanoğlu, Fikret Işik and Van De Ville, Dimitri. Transient brain activity disentangles fmri resting-state dynamics in terms of spatially and temporally overlapping networks. *Nature communications*, 6(1):1–10, 2015.

Katsuki, Fumi and Constantinidis, Christos. Unique and shared roles of the posterior parietal and dorsolateral prefrontal cortex in cognitive functions. *Frontiers in integrative neuroscience*, 6:17, 2012.

Keller, Joseph B; Hedden, Trey; Thompson, Todd W; Anteraper, Sheeba A; Gabrieli, John DE, and Whitfield-Gabrieli, Susan. Resting-state anticorrelations between medial and lateral prefrontal cortex: association with working memory, aging, and individual differences. *Cortex*, 64:271–280, 2015.

Keshavan, Anisha; Paul, Friedemann; Beyer, Mona K; Zhu, Alyssa H; Papinutto, Nico; Shinohara, Russell T; Stern, William; Amann, Michael; Bakshi, Rohit; Bischof, Antje, and others, . Power estimation for non-standardized multisite studies. *NeuroImage*, 134:281–294, 2016.

Kessler, Ronald C. The costs of depression. *Psychiatric Clinics*, 35(1):1–14, 2012.

Khosla, Meenakshi; Jamison, Keith; Kuceyeski, Amy, and Sabuncu, Mert R. 3d convolutional neural networks for classification of functional connectomes. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 137–145. Springer, 2018.

Kia, Seyed Mostafa; Huijsdens, Hester; Dinga, Richard; Wolfers, Thomas; Mennes, Maarten; Andreassen, Ole A; Westlye, Lars T; Beckmann, Christian F, and Marquand, Andre F. Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data. In *International*

Conference on Medical Image Computing and Computer-Assisted Intervention, pages 699–709. Springer, 2020.

Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kinouchi, Osame and Copelli, Mauro. Optimal dynamical range of excitable networks at criticality. *Nature physics*, 2(5):348–351, 2006.

Koch, Stefan P; Hägele, Claudia; Haynes, John-Dylan; Heinz, Andreas; Schlagenhauf, Florian, and Sterzer, Philipp. Diagnostic classification of schizophrenia patients on the basis of regional reward-related fmri signal patterns. *PloS one*, 10(3):e0119089, 2015.

Kostro, Daniel; Abdulkadir, Ahmed; Durr, Alexandra; Roos, Raymund; Leavitt, Blair R; Johnson, Hans; Cash, David; Tabrizi, Sarah J; Scahill, Rachael I; Ronneberger, Olaf, and others, . Correction of inter-scanner and within-subject variance in structural mri based automated diagnosing. *NeuroImage*, 98:405–415, 2014.

Kumari, Veena; Peters, Emmanuelle R; Fannon, Dominic; Antonova, Elena; Premkumar, Preethi; Anilkumar, Anantha P; Williams, Steven CR, and Kuipers, Elizabeth. Dorsolateral prefrontal cortex activity predicts responsiveness to cognitive–behavioral therapy in schizophrenia. *Biological psychiatry*, 66(6):594–602, 2009.

Lancichinetti, Andrea; Radicchi, Filippo; Ramasco, José J, and Fortunato, Santo. Finding statistically significant communities in networks. *PloS one*, 6(4):e18961, 2011.

Lee, Kangjoo; Tak, Sungho, and Ye, Jong Chul. A data-driven sparse glm for fmri analysis using sparse dictionary learning with mdl criterion. *IEEE Transactions on Medical Imaging*, 30(5): 1076–1089, 2010.

Leech, Robert and Sharp, David J. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1):12–32, 2014.

Li, Hongming; Satterthwaite, Theodore D, and Fan, Yong. Large-scale sparse functional networks from resting state fmri. *Neuroimage*, 156:1–13, 2017.

Li, Hongming; Zhu, Xiaofeng, and Fan, Yong. Identification of multi-scale hierarchical brain functional networks using deep matrix factorization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–231. Springer, 2018a.

Li, Peng; Jing, Ri-Xing; Zhao, Rong-Jiang; Shi, Le; Sun, Hong-Qiang; Ding, Zengbo; Lin, Xiao; Lu, Lin, and Fan, Yong. Association between functional and structural connectivity of the corticostriatal network in people with schizophrenia and unaffected first-degree relatives. *Journal of Psychiatry and Neuroscience*, 45(6):395–405, 2020a.

Li, Xiaoxiao; Gu, Yufeng; Dvornek, Nicha; Staib, Lawrence H; Ventola, Pamela, and Duncan, James S. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 65:101765, 2020b.

Li, Yitong; Murias, Michael; Major, Samantha; Dawson, Geraldine, and Carlson, David E. Extracting relationships by multi-domain matching. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6799–6810, 2018b.

Liang, Sicong and Zhang, Yu. A simple general approach to balance task difficulty in multi-task learning. *arXiv preprint arXiv:2002.04792*, 2020.

Liang, Sugai; Deng, Wei; Li, Xiaojing; Greenshaw, Andrew J; Wang, Qiang; Li, Mingli; Ma, Xiaohong; Bai, Tong-Jian; Bo, Qi-Jing; Cao, Jun, and others, . Biotypes of major depressive disorder: Neuroimaging evidence from resting-state default mode network patterns. *NeuroImage: Clinical*, 28:102514, 2020.

Lin, Xi; Zhen, Hui-Ling; Li, Zhenhua; Zhang, Qing-Fu, and Kwong, Sam. Pareto multi-task learning. *Advances in neural information processing systems*, 32:12060–12070, 2019.

Liu, Feng; Guo, Wenbin; Liu, Ling; Long, Zhiliang; Ma, Chaoqiong; Xue, Zhimin; Wang, Yifeng; Li, Jun; Hu, Maorong; Zhang, Jianwei, and others, . Abnormal amplitude low-frequency oscillations in medication-naive, first-episode patients with major depressive disorder: a resting-state fmri study. *Journal of affective disorders*, 146(3):401–406, 2013.

Liu, Feng; Guo, Wenbin; Fouche, Jean-Paul; Wang, Yifeng; Wang, Wenqin; Ding, Jurong; Zeng, Ling; Qiu, Changjian; Gong, Qiyong; Zhang, Wei, and others, . Multivariate classification of social anxiety disorder using whole brain functional connectivity. *Brain Structure and Function*, 220(1):101–115, 2015.

Liu, Yujie; Chen, Yaoping; Liang, Xinyu; Li, Danian; Zheng, Yanting; Zhang, Hanyue; Cui, Ying; Chen, Jingxian; Liu, Jiarui, and Qiu, Shijun. Altered resting-state functional connectivity of multiple networks and disrupted correlation with executive function in major depressive disorder. *Frontiers in Neurology*, page 272, 2020.

Long, Mingsheng; Cao, Zhangjie; Wang, Jianmin, and Yu, Philip S. Learning multiple tasks with multilinear relationship networks. *Advances in neural information processing systems*, 30, 2017.

Lowd, Daniel and Meek, Christopher. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647, 2005.

Luo, Lei; Zhang, Yanfu, and Huang, Heng. Adversarial nonnegative matrix factorization. In *International Conference on Machine Learning*, pages 6479–6488. PMLR, 2020.

Luo, Liang; Wu, Huawang; Xu, Jinping; Chen, Fangfang; Wu, Fengchun; Wang, Chao, and Wang, Jiaojian. Abnormal large-scale resting-state functional networks in drug-free major depressive disorder. *Brain imaging and behavior*, 15(1):96–106, 2021.

Lv, Jinglei; Jiang, Xi; Li, Xiang; Zhu, Dajiang; Zhang, Shu; Zhao, Shijie; Chen, Hanbo; Zhang, Tuo; Hu, Xintao; Han, Junwei, and others, . Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *IEEE Transactions on Biomedical Engineering*, 62(4):1120–1131, 2014.

Lv, Jinglei; Lin, Binbin; Li, Qingyang; Zhang, Wei; Zhao, Yu; Jiang, Xi; Guo, Lei; Han, Junwei; Hu, Xintao; Guo, Christine, and others, . Task fmri data analysis based on supervised stochastic coordinate coding. *Medical image analysis*, 38:1–16, 2017.

Madry, Aleksander; Makelov, Aleksandar; Schmidt, Ludwig; Tsipras, Dimitris, and Vladu, Adrian. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Madsen, Kristoffer H; Churchill, Nathan W, and Mørup, Morten. Quantifying functional connectivity in multi-subject fmri data using component models. *Human brain mapping*, 38(2):882–899, 2017.

Mandt, Stephan; Hoffman, Matthew D, and Blei, David M. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18:1–35, 2017.

Marcus, Daniel S; Wang, Tracy H; Parker, Jamie; Csernansky, John G; Morris, John C, and Buckner, Randy L. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9): 1498–1507, 2007.

Menon, Vinod. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in cognitive sciences*, 15(10):483–506, 2011.

Meunier, David; Lambiotte, Renaud; Fornito, Alex; Ersche, Karen, and Bullmore, Edward T. Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3:37, 2009.

Miller, Greg. Beyond dsm: seeking a brain-based classification of mental illness, 2010.

Minshew, Nancy J and Keller, Timothy A. The nature of brain dysfunction in autism: functional brain imaging studies. *Current opinion in neurology*, 23(2):124, 2010.

Moosavi-Dezfooli, Seyed-Mohsen; Fawzi, Alhussein; Fawzi, Omar, and Frossard, Pascal. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

Mueller, Sophia; Keeser, Daniel; Samson, Andrea C; Kirsch, Valerie; Blautzik, Janusch; Grothe, Michel; Erat, Okan; Hegenloh, Michael; Coates, Ute; Reiser, Maximilian F, and others, . Convergent findings of altered functional and structural brain connectivity in individuals with high functioning autism: a multimodal mri study. *PloS one*, 8(6):e67329, 2013.

Mueller, Timothy I; Leon, Andrew C; Keller, Martin B; Solomon, David A; Endicott, Jean; Coryell, William; Warshaw, Meredith, and Maser, Jack D. Recurrence after recovery from major depressive disorder during 15 years of observational follow-up. *American Journal of Psychiatry*, 156(7): 1000–1006, 1999.

Mulders, Peter C; van Eijndhoven, Philip F; Schene, Aart H; Beckmann, Christian F, and Tendolkar, Indira. Resting-state functional connectivity in major depressive disorder: a review. *Neuroscience & Biobehavioral Reviews*, 56:330–344, 2015.

Munafò, Marcus R; Nosek, Brian A; Bishop, Dorothy VM; Button, Katherine S; Chambers, Christopher D; Percie du Sert, Nathalie; Simonsohn, Uri; Wagenmakers, Eric-Jan; Ware, Jennifer J, and Ioannidis, John. A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9, 2017.

Nakano, Takashi; Takamura, Masahiro; Ichikawa, Naho; Okada, Go; Okamoto, Yasumasa; Yamada, Makiko; Suhara, Tetsuya; Yamawaki, Shigeto, and Yoshimoto, Junichiro. Enhancing multi-center generalization of machine learning-based depression diagnosis from resting-state fmri. *Frontiers in Psychiatry*, 11:400, 2020.

Navrady, LB; Wolters, MK; MacIntyre, DJ; Clarke, Toni-Kim; Campbell, AI; Murray, AD; Evans, KL; Seckl, Jonathan; Haley, Christopher; Milburn, Keith, and others, . Cohort profile: stratifying resilience and depression longitudinally (stradl): a questionnaire follow-up of generation scotland: Scottish family health study (gs: Sfhs). *International journal of epidemiology*, 47(1):13–14g, 2018.

Nickel, Maximillian and Kiela, Douwe. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30:6338–6347, 2017.

Nielsen, Jared A; Zielinski, Brandon A; Fletcher, P Thomas; Alexander, Andrew L; Lange, Nicholas; Bigler, Erin D; Lainhart, Janet E, and Anderson, Jeffrey S. Multisite functional connectivity mri classification of autism: Abide results. *Frontiers in human neuroscience*, 7:599, 2013.

Nikolaus, Susanne; Hautzel, Hubertus; Heinzel, Alexander, and Müller, Hans-Wilhelm. Key players in major and bipolar depression—a retrospective analysis of in vivo imaging studies. *Behavioural brain research*, 232(2):358–390, 2012.

Niznikiewicz, Margaret A; Kubicki, Marek, and Shenton, Martha E. Recent structural and functional imaging findings in schizophrenia. *Current Opinion in Psychiatry*, 16(2):123–147, 2003.

Noble, Stephanie; Scheinost, Dustin; Finn, Emily S; Shen, Xilin; Papademetris, Xenophon; McEwen, Sarah C; Bearden, Carrie E; Addington, Jean; Goodyear, Bradley; Cadenhead, Kristin S, and others, . Multisite reliability of mr-based functional connectivity. *Neuroimage*, 146:959–970, 2017.

Noroozi, Ali and Rezghi, Mansoor. A tensor-based framework for rs-fmri classification and functional connectivity construction. *Frontiers in neuroinformatics*, page 46, 2020.

Nosek, Brian A and Errington, Timothy M. Reproducibility in cancer biology: Making sense of replications. *Elife*, 6:e23383, 2017.

Olivetti, Emanuele; Greiner, Susanne, and Avesani, Paolo. Adhd diagnosis from multiple data sources with batch effects. *Frontiers in systems neuroscience*, 6:70, 2012.

Otte, Christian; Gold, Stefan M; Penninx, Brenda W; Pariante, Carmine M; Etkin, Amit; Fava, Maurizio; Mohr, David C, and Schatzberg, Alan F. Major depressive disorder. *Nature reviews Disease primers*, 2(1):1–20, 2016.

Ozaktas, Haldun M. Paradigms of connectivity for computer circuits and networks. *Optical Engineering*, 31(7):1563–1567, 1992.

Park, Hae-Jeong and Friston, Karl. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158):1238411, 2013.

Peng, Xingchao; Bai, Qinxun; Xia, Xide; Huang, Zijun; Saenko, Kate, and Wang, Bo. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

Pluta, John; Yushkevich, Paul; Das, Sandhitsu, and Wolk, David. In vivo analysis of hippocampal subfield atrophy in mild cognitive impairment via semi-automatic segmentation of t2-weighted mri. *Journal of Alzheimer's disease*, 31(1):85–99, 2012.

Podosinnikova, Anastasia; Hein, Matthias, and Gemulla, Rainer. *Robust Principal Component Analysis as a Nonlinear Eigenproblem*. PhD thesis, Saarland University, 2013.

Pomponio, Raymond; Erus, Guray; Habes, Mohamad; Doshi, Jimit; Srinivasan, Dhivya; Mamourian, Elizabeth; Bashyam, Vishnu; Nasrallah, Ilya M; Satterthwaite, Theodore D; Fan, Yong, and others, . Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450, 2020.

Potluru, Vamsi K and Calhoun, Vince D. Group learning using contrast nmf: Application to functional and structural mri of schizophrenia. In *2008 IEEE International Symposium on Circuits and Systems*, pages 1336–1339. IEEE, 2008.

Power, Jonathan D; Barnes, Kelly A; Snyder, Abraham Z; Schlaggar, Bradley L, and Petersen, Steven E. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *Neuroimage*, 59(3):2142–2154, 2012.

Puxeddu, Maria Grazia; Faskowitz, Joshua; Betzel, Richard F; Petti, Manuela; Astolfi, Laura, and Sporns, Olaf. The modular organization of brain cortical connectivity across the human lifespan. *NeuroImage*, page 116974, 2020.

Qureshi, Muhammad Naveed Iqbal; Oh, Jooyoung; Cho, Dongrae; Jo, Hang Joon, and Lee, Boreom. Multimodal discrimination of schizophrenia using hybrid weighted feature concatenation of brain functional connectivity and anatomical features with an extreme learning machine. *Frontiers in neuroinformatics*, 11:59, 2017a.

Qureshi, Muhammad Naveed Iqbal; Oh, Jooyoung; Min, Beomjun; Jo, Hang Joon, and Lee, Boreom. Multi-modal, multi-measure, and multi-class discrimination of adhd with hierarchical feature extraction and extreme learning machine using structural and functional brain mri. *Frontiers in human neuroscience*, 11:157, 2017b.

Raichle, Marcus E. The restless brain. *Brain connectivity*, 1(1):3–12, 2011.

Raichle, Marcus E. The brain's default mode network. *Annual review of neuroscience*, 38:433–447, 2015.

Raichle, Marcus E; MacLeod, Ann Mary; Snyder, Abraham Z; Powers, William J; Gusnard, Debra A, and Shulman, Gordon L. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

Rajpoot, Kashif; Riaz, Atif; Majeed, Waqas, and Rajpoot, Nasir. Functional connectivity alterations in epilepsy from resting-state functional mri. *PloS one*, 10(8):e0134944, 2015.

Rauschecker, Josef P and Scott, Sophie K. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6):718–724, 2009.

Ravishankar, Hariharan; Madhavan, Radhika; Mullick, Rakesh; Shetty, Teena; Marinelli, Luca, and Joel, Suresh E. Recursive feature elimination for biomarker discovery in resting-state functional connectivity. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4071–4074. IEEE, 2016.

Reddi, Sashank J; Kale, Satyen, and Kumar, Sanjiv. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Resnick, Susan M; Pham, Dzung L; Kraut, Michael A; Zonderman, Alan B, and Davatzikos, Christos. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *Journal of Neuroscience*, 23(8):3295–3301, 2003.

Riaz, Atif; Rajpoot, Kashif, and Rajpoot, Nasir. A connectivity difference measure for identification of functional neuroimaging markers for epilepsy. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1517–1520. IEEE, 2013.

Riaz, Atif; Asad, Muhammad; Al-Arif, SM; Alonso, Eduardo; Dima, Danai; Corr, Philip, and Slabaugh, Greg. Fcnet: a convolutional neural network for calculating functional connectivity from functional mri. In *International Workshop on Connectomics in Neuroimaging*, pages 70–78. Springer, 2017.

Riaz, Atif; Asad, Muhammad; Alonso, Eduardo, and Slabaugh, Greg. Deepfmri: End-to-end deep learning for functional connectivity and classification of adhd using fmri. *Journal of neuroscience methods*, 335:108506, 2020.

Rosa, Maria J; Portugal, Liana; Hahn, Tim; Fallgatter, Andreas J; Garrido, Marta I; Shawe-Taylor, John, and Mourao-Miranda, Janaina. Sparse network-based models for patient classification using fmri. *Neuroimage*, 105:493–506, 2015.

Rozycki, Martin; Satterthwaite, Theodore D; Koutsouleris, Nikolaos; Erus, Guray; Doshi, Jimit; Wolf, Daniel H; Fan, Yong; Gur, Raquel E; Gur, Ruben C; Meisenzahl, Eva M, and others, . Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophrenia bulletin*, 44(5): 1035–1044, 2018.

Rubinov, Mikail and Sporns, Olaf. Weight-conserving characterization of complex functional brain networks. *Neuroimage*, 56(4):2068–2079, 2011.

Ruder, Sebastian. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Ruder, Sebastian; Bingel, Joachim; Augenstein, Isabelle, and Søgaard, Anders. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829, 2019.

Sacchet, Matthew D; Ho, Tiffany C; Connolly, Colm G; Tymofiyeva, Olga; Lewinn, Kaja Z; Han, Laura KM; Blom, Eva H; Tapert, Susan F; Max, Jeffrey E; Frank, Guido KW, and others, . Large-scale hypoconnectivity between resting-state functional networks in unmedicated adolescent major depressive disorder. *Neuropsychopharmacology*, 41(12):2951–2960, 2016.

Sahoo, Dushyant and Davatzikos, Christos. Learning robust hierarchical patterns of human brain across many fmri studies. *Advances in Neural Information Processing Systems*, 34, 2021.

Sahoo, Dushyant; Honnorat, Nicolas, and Davatzikos, Christos. Gpu accelerated extraction of sparse granger causality patterns. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 604–607. IEEE, 2018.

Salimi-Khorshidi, Gholamreza; Douaud, Gwenaëlle; Beckmann, Christian F; Glasser, Matthew F; Griffanti, Ludovica, and Smith, Stephen M. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.

Salmela, Viljami; Socada, Lumikukka; Söderholm, John; Heikkilä, Roope; Lahti, Jari; Ekelund, Jesper, and Isometsä, Erkki. Reduced visual contrast suppression during major depressive episodes. *Journal of Psychiatry and Neuroscience*, 46(2):E222–E231, 2021.

Santarnecchi, Emiliano; Galli, Giulia; Polizzotto, Nicola Riccardo; Rossi, Alessandro, and Rossi, Simone. Efficiency of weak brain connections support general cognitive functioning. *Human brain mapping*, 35(9):4566–4582, 2014.

Satterthwaite, Theodore D; Elliott, Mark A; Ruparel, Kosha; Loughead, James; Prabhakaran, Karthik; Calkins, Monica E; Hopson, Ryan; Jackson, Chad; Keefe, Jack; Riley, Marisa, and others, . Neuroimaging of the philadelphia neurodevelopmental cohort. *Neuroimage*, 86:544–553, 2014.

Schaie, K Warner and Willis, Sherry. *Handbook of the Psychology of Aging*. Academic Press, USA, 2021.

Sener, Ozan and Koltun, Vladlen. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Shen, Huawei; Cheng, Xueqi; Cai, Kai, and Hu, Mao-Bin. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8): 1706–1712, 2009.

Shinohara, Russell T; Oh, Jiwon; Nair, Govind; Calabresi, Peter A; Davatzikos, Christos; Doshi, Jimit; Henry, Roland G; Kim, Gloria; Linn, Kristin A; Papinutto, Nico, and others, . Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis. *American Journal of Neuroradiology*, 38(8):1501–1509, 2017.

Simon, Herbert A. The architecture of complexity. In *Facets of systems science*, pages 457–476. Springer, 1991.

Sinha, Aman; Namkoong, Hongseok, and Duchi, John. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Skudlarski, Pawel; Jagannathan, Kanchana; Calhoun, Vince D; Hampson, Michelle; Skudlarska, Beata A, and Pearlson, Godfrey. Measuring brain connectivity: diffusion tensor imaging validates resting state temporal correlations. *Neuroimage*, 43(3):554–561, 2008.

Smith, Stephen M; Fox, Peter T; Miller, Karla L; Glahn, David C; Fox, P Mickle; Mackay, Clare E; Filippini, Nicola; Watkins, Kate E; Toro, Roberto; Laird, Angela R, and others, . Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences*, 106(31):13040–13045, 2009.

Smith, Stephen M; Miller, Karla L; Salimi-Khorshidi, Gholamreza; Webster, Matthew; Beckmann, Christian F; Nichols, Thomas E; Ramsey, Joseph D, and Woolrich, Mark W. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Smith, Stephen M; Miller, Karla L; Moeller, Steen; Xu, Junqian; Auerbach, Edward J; Woolrich, Mark W; Beckmann, Christian F; Jenkinson, Mark; Andersson, Jesper; Glasser, Matthew F, and others, . Temporally-independent functional modes of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 109(8):3131–3136, 2012.

Smith, Stephen M; Hyvärinen, Aapo; Varoquaux, Gaël; Miller, Karla L, and Beckmann, Christian F. Group-pca for very large fmri datasets. *Neuroimage*, 101:738–749, 2014.

Song, Xue Mei; Hu, Xi-Wen; Li, Zhe; Gao, Yuan; Ju, Xuan; Liu, Dong-Yu; Wang, Qian-Nan; Xue, Chuang; Cai, Yong-Chun; Bai, Ruiliang, and others, . Reduction of higher-order occipital gaba and impaired visual perception in acute major depressive disorder. *Molecular Psychiatry*, pages 1–9, 2021.

Sporns, Olaf. *Networks of the Brain*. MIT press, 2010.

Sporns, Olaf and Betzel, Richard F. Modular brain networks. *Annual review of psychology*, 67: 613–640, 2016.

Spreng, R Nathan; Stevens, W Dale; Viviano, Joseph D, and Schacter, Daniel L. Attenuated anticorrelation between the default and dorsal attention networks with aging: evidence from task and rest. *Neurobiology of aging*, 45:149–160, 2016.

Stam, Cornelis J. Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10):683–695, 2014.

Stewart, Gilbert W. Perturbation theory for the singular value decomposition. Technical report, 1998.

Stolicyn, Aleks; Harris, Mathew A; Shen, Xueyi; Barbu, Miruna C; Adams, Mark J; Hawkins, Emma L; de Nooij, Laura; Yeung, Hon Wah; Murray, Alison D; Lawrie, Stephen M, and others, . Automated classification of depression from structural brain measures across two independent community-based cohorts. *Human brain mapping*, 41(14):3922–3937, 2020.

Sudlow, Cathie; Gallacher, John; Allen, Naomi; Beral, Valerie; Burton, Paul; Danesh, John; Downey, Paul; Elliott, Paul; Green, Jane; Landray, Martin, and others, . Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, 12(3):e1001779, 2015.

Tozzi, Leonardo; Zhang, Xue; Chesnut, Megan; Holt-Gosselin, Bailey; Ramirez, Carolina A, and Williams, Leanne M. Reduced functional connectivity of default mode network subsystems in depression: meta-analytic evidence and relationship with trait rumination. *NeuroImage: Clinical*, 30:102570, 2021.

Tramèr, Florian; Kurakin, Alexey; Papernot, Nicolas; Goodfellow, Ian; Boneh, Dan, and McDaniel, Patrick. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Triantafyllou, Christina; Hoge, Richard D; Krueger, Gunnar; Wiggins, Christopher J; Potthast, Andreas; Wiggins, Graham C, and Wald, Lawrence L. Comparison of physiological noise at 1.5 t, 3 t and 7 t and optimization of fmri acquisition parameters. *Neuroimage*, 26(1):243–250, 2005.

Trigeorgis, George; Bousmalis, Konstantinos; Zafeiriou, Stefanos, and Schuller, Björn W. A deep matrix factorization method for learning attribute representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):417–429, 2017.

Trivedi, Madhukar H; McGrath, Patrick J; Fava, Maurizio; Parsey, Ramin V; Kurian, Benji T; Phillips, Mary L; Oquendo, Maria A; Bruder, Gerard; Pizzagalli, Diego; Toups, Marisa, and others, . Establishing moderators and biosignatures of antidepressant response in clinical care (embarc): Rationale and design. *Journal of psychiatric research*, 78:11–23, 2016.

Uddin, Lucina Q; Clare Kelly, AM; Biswal, Bharat B; Xavier Castellanos, F, and Milham, Michael P. Functional connectivity of default mode network components: correlation, anticorrelation, and causality. *Human brain mapping*, 30(2):625–637, 2009.

Van Essen, David C; Ugurbil, Kamil; Auerbach, Edward; Barch, Deanna; Behrens, TEJ; Bucholz, Richard; Chang, Acer; Chen, Liyong; Corbetta, Maurizio; Curtiss, Sandra W, and others, . The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

Van Essen, David C; Smith, Stephen M; Barch, Deanna M; Behrens, Timothy EJ; Yacoub, Essa; Ugurbil, Kamil; Consortium, Wu-Minn HCP, and others, . The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Van Eyndhoven, Simon; Vervliet, Nico; De Lathauwer, Lieven, and Van Huffel, Sabine. Identifying stable components of matrix/tensor factorizations via low-rank approximation of inter-factorization similarity. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

Van Kerkoerle, Timo; Self, Matthew W; Dagnino, Bruno; Gariel-Mathis, Marie-Alice; Poort, Jasper; Van Der Togt, Chris, and Roelfsema, Pieter R. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences*, 111(40):14332–14341, 2014.

Veer, Ilya M; Beckmann, Christian; Van Tol, Marie-Jose; Ferrarini, Luca; Milles, Julien; Veltman, Dick; Aleman, André; Van Buchem, Mark A; Van Der Wee, Nic JA, and Rombouts, Serge AR. Whole brain resting-state analysis reveals decreased functional connectivity in major depression. *Frontiers in systems neuroscience*, 4:41, 2010.

Vega, Roberto and Greiner, Russ. Finding effective ways to (machine) learn fmri-based classifiers from multi-site data. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 32–39. Springer, 2018.

Vergani, Francesco; Lacerda, Luis; Martino, Juan; Attems, Johannes; Morris, Christopher; Mitchell, Patrick; de Schotten, Michel Thiebaut, and Dell'Acqua, Flavio. White matter connections of the supplementary motor area in humans. *Journal of Neurology, Neurosurgery & Psychiatry*, 85(12): 1377–1385, 2014.

Wagstaff, Kiri; Cardie, Claire; Rogers, Seth; Schrödl, Stefan, and others, . Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.

Wang, Huahua; Banerjee, Arindam, and Boley, Daniel. Common component analysis for multiple covariance matrices. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 956–964, 2011.

Wang, Lebo; Li, Kaiming, and Hu, Xiaoping P. Graph convolutional network for fmri analysis based on connectivity neighborhood. *Network Neuroscience*, 5(1):83–95, 2021.

Wang, Mingliang; Zhang, Daoqiang; Huang, Jiashuang; Yap, Pew-Thian; Shen, Dinggang, and Liu, Mingxia. Identifying autism spectrum disorder with multi-site fmri via low-rank domain adaptation. *IEEE transactions on medical imaging*, 39(3):644–655, 2019.

Wang, Xun-Heng; Jiao, Yun, and Li, Lihua. Identifying individuals with attention deficit hyperactivity disorder based on temporal variability of dynamic functional connectivity. *Scientific reports*, 8(1): 1–12, 2018.

Wang, Xunheng; Jiao, Yun; Tang, Tianyu; Wang, Hui, and Lu, Zuhong. Altered regional homogeneity patterns in adults with attention-deficit hyperactivity disorder. *European journal of radiology*, 82 (9):1552–1557, 2013.

Warnick, Ryan; Guindani, Michele; Erhardt, Erik; Allen, Elena; Calhoun, Vince, and Vannucci, Marina. A bayesian approach for estimating dynamic functional network connectivity in fmri data. *Journal of the American Statistical Association*, 113(521):134–151, 2018.

Wee, Chong-Yaw; Yap, Pew-Thian; Zhang, Daoqiang; Denny, Kevin; Browndyke, Jeffrey N; Potter, Guy G; Welsh-Bohmer, Kathleen A; Wang, Lihong, and Shen, Dinggang. Identification of mci individuals using structural and functional connectivity networks. *Neuroimage*, 59(3):2045–2056, 2012.

Wise, T; Marwood, L; Perkins, AM; Herane-Vives, A; Joules, R; Lythgoe, DJ; Luh, WM; Williams, SCR; Young, AH; Cleare, AJ, and others, . Instability of default mode network connectivity in major depression: a two-sample confirmation study. *Translational psychiatry*, 7(4):e1105–e1105, 2017.

Wolfers, Thomas; Floris, Dorothea L; Dinga, Richard; van Rooij, Daan; Isakoglou, Christina; Kia, Seyed Mostafa; Zabihi, Mariam; Llera, Alberto; Chowdanayaka, Rajanikanth; Kumar, Vinod J, and others, . From pattern classification to stratification: towards conceptualizing the heterogeneity of autism spectrum disorder. *Neuroscience & Biobehavioral Reviews*, 104:240–254, 2019.

Wong, Eric and Kolter, Zico. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

Wu, Fan; Cai, Jiahui; Wen, Canhong, and Tan, Haizhu. Co-sparse non-negative matrix factorization. *Frontiers in Neuroscience*, 15, 2021.

Wu, Qi-Zhu; Li, Dong-Ming; Kuang, Wei-Hong; Zhang, Ti-Jiang; Lui, Su; Huang, Xiao-Qi; Chan, Raymond CK; Kemp, Graham J, and Gong, Qi-Yong. Abnormal regional spontaneous neural activity in treatment-refractory depression revealed by resting-state fmri. *Human brain mapping*, 32(8):1290–1299, 2011.

Wu, Tao; Wang, Liang; Chen, Yi; Zhao, Cheng; Li, Kuncheng, and Chan, Piu. Changes of functional connectivity of the motor network in the resting state in parkinson's disease. *Neuroscience letters*, 460(1):6–10, 2009.

Xia, Mingrui; Si, Tianmei; Sun, Xiaoyi; Ma, Qing; Liu, Bangshan; Wang, Li; Meng, Jie; Chang, Miao; Huang, Xiaoqi; Chen, Ziqi, and others, . Reproducibility of functional brain alterations in major depressive disorder: Evidence from a multisite resting-state functional mri study with 1,434 individuals. *Neuroimage*, 189:700–714, 2019.

Xu, Jiansong; Potenza, Marc N; Calhoun, Vince D; Zhang, Rubin; Yip, Sarah W; Wall, John T; Pearlson, Godfrey D; Worhunsky, Patrick D; Garrison, Kathleen A, and Moran, Joseph M. Large-scale functional network overlap is a general property of brain functional organization: reconciling inconsistent fmri findings from general-linear-model-based analyses. *Neuroscience & Biobehavioral Reviews*, 71:83–100, 2016.

Yamashita, Ayumu; Sakai, Yuki; Yamada, Takashi; Yahata, Noriaki; Kunimatsu, Akira; Okada, Naohiro; Itahashi, Takashi; Hashimoto, Ryuichiro; Mizuta, Hiroto; Ichikawa, Naho, and others, .

Generalizable brain network markers of major depressive disorder across multiple imaging sites. *PLoS biology*, 18(12):e3000966, 2020.

Yamashita, Ayumu; Sakai, Yuki; Yamada, Takashi; Yahata, Noriaki; Kunimatsu, Akira; Okada, Naohiro; Itahashi, Takashi; Hashimoto, Ryuichiro; Mizuta, Hiroto; Ichikawa, Naho, and others, . Common brain networks between major depressive-disorder diagnosis and symptoms of depression that are validated for independent cohorts. *Frontiers in psychiatry*, 12:888, 2021.

Yan, Baoyu; Xu, Xiaopan; Liu, Mengwan; Zheng, Kaizhong; Liu, Jian; Li, Jianming; Wei, Lei; Zhang, Binjie; Lu, Hongbing, and Li, Baojuan. Quantitative identification of major depression based on resting-state dynamic functional connectivity: a machine learning approach. *Frontiers in neuroscience*, 14:191, 2020.

Yan, Chao-Gan; Craddock, R Cameron; He, Yong, and Milham, Michael P. Addressing head motion dependencies for small-world topologies in functional connectomics. *Frontiers in human neuroscience*, 7:910, 2013.

Yan, Chao-Gan; Chen, Xiao; Li, Le; Castellanos, Francisco Xavier; Bai, Tong-Jian; Bo, Qi-Jing; Cao, Jun; Chen, Guan-Mao; Chen, Ning-Xuan; Chen, Wei, and others, . Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019.

Yang, Jaewon and Leskovec, Jure. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM, 2013.

Yassin, Walid; Nakatani, Hironori; Zhu, Yinghan; Kojima, Masaki; Owada, Keiho; Kuwabara, Hitoshi; Gonoi, Wataru; Aoki, Yuta; Takao, Hidemasa; Natsubori, Tatsunobu, and others, . Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. *Translational psychiatry*, 10(1):1–11, 2020.

Ye, Ming; Yang, Tianliang; Qing, Peng; Lei, Xu; Qiu, Jiang, and Liu, Guangyuan. Changes of functional brain networks in major depressive disorder: a graph theoretical analysis of resting-state fmri. *PloS one*, 10(9):e0133775, 2015.

Yeo, BT Thomas; Krienen, Fenna M; Chee, Michael WL, and Buckner, Randy L. Estimates of segregation and overlap of functional connectivity networks in the human cerebral cortex. *Neuroimage*, 88:212–227, 2014.

Yu, Meichen; Linn, Kristin A; Cook, Philip A; Phillips, Mary L; McInnis, Melvin; Fava, Maurizio; Trivedi, Madhukar H; Weissman, Myrna M; Shinohara, Russell T, and Sheline, Yvette I. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human brain mapping*, 39(11):4213–4227, 2018.

Yu, Ming; Gupta, Varun, and Kolar, Mladen. Recovery of simultaneous low rank and two-way sparse coefficient matrices, a nonconvex approach. *Electronic Journal of Statistics*, 14(1):413–457, 2020.

Zeng, Ling-Li; Shen, Hui; Liu, Li; Wang, Lubin; Li, Baojuan; Fang, Peng; Zhou, Zongtan; Li, Yaming, and Hu, Dewen. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain*, 135(5):1498–1507, 2012.

Zhan, Liang; Jenkins, Lisanne M; Wolfson, Ouri E; GadElkarim, Johnson Jonaris; Nocito, Kevin; Thompson, Paul M; Ajilore, Olusola A; Chung, Moo K, and Leow, Alex D. The significance of negative correlations in brain connectivity. *Journal of Comparative Neurology*, 525(15): 3251–3265, 2017.

Zhang, Chao; Baum, Stefi A; Adduru, Viraj R; Biswal, Bharat B, and Michael, Andrew M. Test-retest reliability of dynamic functional connectivity in resting state fmri. *NeuroImage*, 183:907–918, 2018.

Zhang, Jie; Cheng, Wei; Wang, ZhengGe; Zhang, ZhiQiang; Lu, WenLian; Lu, GuangMing, and Feng, Jianfeng. Pattern classification of large-scale functional brain networks: identification of informative neuroimaging markers for epilepsy. *PloS one*, 7(5):e36733, 2012.

Zhang, Junyi; Wan, Peng, and Zhang, Daoqiang. Transport-based joint distribution alignment for multi-site autism spectrum disorder diagnosis using resting-state fmri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 444–453. Springer, 2020a.

Zhang, Wei; Zhao, Shijie; Hu, Xintao; Dong, Qinglin; Huang, Heng; Zhang, Shu; Zhao, Yu; Dai, Haixing; Ge, Fangfei; Guo, Lei, and others, . Hierarchical organization of functional brain networks revealed by hybrid spatiotemporal deep learning. *Brain connectivity*, 10(2):72–82, 2020b.

Zhang, Yipu; Xiao, Li; Zhang, Gemeng; Cai, Biao; Stephen, Julia M; Wilson, Tony W; Calhoun, Vince D, and Wang, Yu-Ping. Multi-paradigm fmri fusion via sparse tensor decomposition in brain functional connectivity study. *IEEE journal of biomedical and health informatics*, 25(5): 1712–1723, 2020c.

Zhang, Yu and Yang, Qiang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

Zhao, Han; Zhang, Shanghang; Wu, Guanhang; Moura, José MF; Costeira, Joao P, and Gordon, Geoffrey J. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31:8559–8570, 2018.

Zhao, Jianlong; Huang, Jinjie; Zhi, Dongmei; Yan, Weizheng; Ma, Xiaohong; Yang, Xiao; Li, Xianbin; Ke, Qing; Jiang, Tianzi; Calhoun, Vince D, and others, . Functional network connectivity (fnc)-based generative adversarial network (gan) and its applications in classification of mental disorders. *Journal of neuroscience methods*, 341:108756, 2020a.

Zhao, S; Li, B; Yue, X; Gu, Y; Xu, P; Hu, R; Chai, H, and Keutzer, K. Multi-source domain adaptation for semantic segmentation", advances in neural information processing systems. *Advances in neural information processing systems*, 2019.

Zhao, Sicheng; Li, Bo; Xu, Pengfei, and Keutzer, Kurt. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020b.

Zhao, Youjin; Niu, Running; Lei, Du; Shah, Chandan; Xiao, Yuan; Zhang, Wenjing; Chen, Ziqi; Lui, Su, and Gong, Qiyong. Aberrant gray matter networks in non-comorbid medication-naive patients with major depressive disorder and those with social anxiety disorder. *Frontiers in Human Neuroscience*, 14:172, 2020c.

Zhong, Xue; Shi, Huqing; Ming, Qingsen; Dong, Daifeng; Zhang, Xiaocui; Zeng, Ling-Li, and Yao, Shuqiao. Whole-brain resting-state functional connectivity identified major depressive disorder: a multivariate pattern analysis in two independent samples. *Journal of Affective Disorders*, 218: 346–352, 2017.

Zhou, Zhenyu; Ding, Mingzhou; Chen, Yonghong; Wright, Paul; Lu, Zuhong, and Liu, Yijun. Detecting directional influence in fmri connectivity analysis using pca based granger causality. *Brain research*, 1289:22–29, 2009.

Zhu, Xueling; Wang, Xiang; Xiao, Jin; Liao, Jian; Zhong, Mingtian; Wang, Wei, and Yao, Shuqiao. Evidence of a dissociation pattern in resting-state default mode network connectivity in first-episode, treatment-naive major depression patients. *Biological psychiatry*, 71(7):611–617, 2012.