Multi-Modal Data Fusion of Imaging Genetics for the Discovery of Alzheimer's Disease Pathology

by

Kristen Nalani Knight

(Under the Direction of Nicole Lazar and Liang Liu)

Abstract

Advances in technology and science have spurred an increasing accumulation of complex data sources, leading modern researchers into a diverse present-day data revolution. The new tenor of data collection has created a two-part problem: historical models are often not malleable enough to accommodate data diversity, all the while connections among multiple data sources are being identified as more promising expressions of scientific processes. These new challenges are requiring data scientists to abandon model-based algorithms for data-driven techniques which permit a multi-modal approach to data analysis. A brief discussion of the theoretical underpinnings and computational challenges will be presented to compare decomposition analyses and deep learning methods under this framework. Then, a novel, hybrid approach is extended to the algorithmic learning that permits the combination of multiple sources of information, while retaining interpretability, generalizability, and the capability of downstream analyses. This method has a direct application to imaging genetics, where the aim is to fuse various modalities of neuroimaging with genetic markers to elucidate the progression of neuropathological diseases. Results are showcased with the joint analysis of single nucleotide polymorphism (SNP) markers with anatomical and resting-state functional MRI signals in order to uncover the multi-source pathways underlying Alzheimer's Disease pathology. The data are derived from the most recent stage of the iconic Alzheimer's Disease Neuroimaging Initiative (ADNI).

Index words: multi-modal, data fusion, imaging genetics, decomposition, deep learning, open science

Multi-Modal Data Fusion of Imaging Genetics for the Discovery of Alzheimer's Disease Pathology

by

Kristen Nalani Knight

B.S., Austin Peay State University, 2014
M.S., University of Georgia, 2017

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Doctor of Philosophy

Athens, Georgia

2021

Multi-Modal Data Fusion of Imaging Genetics for the Discovery of Alzheimer's Disease Pathology

by

Kristen Nalani Knight

Major Professors: Liang Liu, Nicole Lazar

Committee: Changwei Li
Cheolwoo Park
Lynne Seymour

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2021

Dedication

The credit of this dissertation goes to my son, Daniel Brock Knight. His support, his kindness, his love, his joy and his humor provided the anchor I needed to achieve this PhD. Without him, I could not have completed this. He encouraged me through every hurdle, and he believed in me the whole way, even when I couldn't believe in myself. I am forever proud and grateful to have this tremendous, intelligent, caring, son who has been my best friend through the process. Daniel, you are forever my favorite person, and the one whom I love with all of my heart. With the acceptance of this degree, I can also fully commit to your successful future. As this degree opens the door for me to a lifetime with a solid and impactful career; the benefits will surely overflow to you. I continue to grow in awe with each passing year at the young man you are becoming, and I can't wait for the realization of who you will become as a grown man. I know I am the luckiest mom in the world. From the bottom of my heart, thank you for being you - my perfect son, Daniel.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern advances in technology and science have increased the *depth* and *breadth* of data collected across disciplines, leading the world into a new tenor of science: data science. The *depth* of data refers to the magnitude; data sets with a larger number of observations and/or measurements are becoming more common among studies. The issues spurred by this higher dimensionality are collectively known as the *big data* problem and is currently a "hot topic" in the fields of data science, machine learning and statistics. In conjunction with managing and analyzing big data, more and more researchers are also experiencing analysis challenges brought on by the *breadth* of data, when applications are realized by multiple sources of information. In fact, many new data analyses in scientific fields draw knowledge from multiple high-dimensional data sources. Thus, the issues incurred by a wider breadth of information must be generalized to varying degrees of depth. Many outcomes may be more accurately expressed by multiple processes, rather than a single *modality*, or data source. However, due to convenience and simplicity, analyses considering single modalities remain most common, which very well could ignore the true underlying expression from multiple sources. The purpose of this dissertation is to explore the development and methodological challenges of multi-modal data fusion. In order to showcase this work, the neurological and genetic underpinnings of Alzheimer's Disease will be explored in this context.

1.1 Motivation and Definition

The motivation for this work stems from recent advances in statistical learning in addition to a wide range of scientific and biomedical applications. With the big data problem comes the introduction of analysis techniques that allow researchers to learn from the data, rather than focusing on more constraining

models. This has opened up the door to answering one particular question, which is: how may we utilize the intertwining networks of structural and functional brain imaging jointly with genetic markers to depict the progression of the stepwise clinical stages of Alzheimer's Disease?

1.1.1 From Model-Restricted to Data-Driven

Much of historical methodology of data analysis is driven by preset models that are assumed to provide a close fit to the data at hand. These models are often parametric, meaning that specific quantities are put in place to express the relationship between the response and explanatory variables. In other words, a model is imposed on the shape of the data that may or may not line up with reality. Several examples of model-driven methods include linear regression, generalized linear mixed models and structural equation modeling. When enough is known about the underlying problem of the study and the typical pattern of the data, one may utilize these parametric models to make inferences. If the assumptions are valid, then the parametric approach will yield fairly accurate results. However, many methods are not robust to major deviations from the model assumptions. The outputs will quickly yield impractical results if this is the case. Not to mention, the model-driven framework lacks the flexibility necessary for the combination of three or more heterogeneous data sets (Calhoun & Sui, 2016; Lahat et al., 2015).

In this data revolution, the data that are collected are often larger and more complex, mandating more tractable analysis techniques. Unfortunately, many researchers continue to use model-based approaches and ignore the invalid assumptions. This may occur either because a certain model is the typical and accepted approach within a discipline, the researcher does not realize there are other options, or the concept of multiple processes yielding a better expression of the outcome is a novel idea. With diverse data at our fingertips, it is more acceptable to make fewer assumptions up front so that researchers are able to learn from the data. Statisticians and data scientists are now thinking "out of the box" of parametric models and are allowing the data itself to drive the algorithms. The paradigm shift of data analysis in this day and age is one from model-based to *data-driven*. Rather than restricting the data to a model, the data impels the underlying algorithm.

Prior to the shift of utilizing multiple data sources to study specific phenomena, there were already data-driven approaches in place. The short list includes principal component analysis (PCA), linear discriminant analysis (LDA) and various forms of regularization, such as partial least squares (PLS). Each of these modeling strategies are able to draw on the advantages of a data-driven

approach. For example, PCA and LDA both reduce dimensionality by finding linear combinations of the data in order to explain the maximal variance and to characterize two or more events (Ma & Dai, 2011). PLS, especially under further regularization assumptions, has proved to be critical in the analysis of sparse data. Nonetheless, these techniques work best under the assumption of unimodality and normality of the data. While the data drives the outcomes under these models, the restrictions are not malleable to the interworkings of multiple sources. This is due to the fact that the architecture for multi-modality is not built into the structure of these algorithms (Sui et al., 2010) .

Consequently, researchers must extend previous data-driven algorithms as well as develop new techniques in order to keep up with the influx of novel information. In the recent years, researchers have shifted their attention to multi-modal data fusion and several review articles have followed suit. With the focus of highlighting the presence of multi-modal integration across disciplines, Turk provided one of the first deeply comprehensive reviews of applications with multi-source interactions (Turk, 2014). Lahat et al. (2015) provide a comprehensive review of independent vector analysis and tensor matrix decomposition, beautifully expressing the benefits of modeling diversity through multi-modality. Calhoun and Sui (2016) introduce an overview of the literature that utilizes decomposition methods for multi-modal data fusion of neuroimaging data with emphasis on applications to psychopathology. Meng et al. (2020) give an in-depth comparison and contrast of deep fusion strategies through machine learning.

1.1.2 An Interdisciplinary Application

The data revolution is not constrained within one or even a few disciplines. The large influx of multi-modal data sources is an interdisciplinary problem, impacting the areas of biology, psychology, engineering, medicine, astrophysics, telecommunication, human-computer interactions (Lahat et al., 2015); even the liberal arts, in education, sociology, law and more (Turk, 2014). Multiple data sources may also come from multi-site data collection initiatives, which collect data representing various fields. Indeed, one study may require that professionals from several disciplines work together in an analysis. Thus, the data-driven technique must be adaptive in order to handle a variety of sources and yet parsimonious enough to be applied by analysts from varying backgrounds. Under the multi-source data framework, there is an interdisciplinary property woven from the manifold nature of the data and is referred to as *data diversity* (Lahat et al., 2015). The following examples are given to showcase the widespread influ-

ence of advances in data collection and to gain some familiarity in multi-source analysis.

Application 1: Signal Processing

Many studies exist in the area of signal processing, where the goal of the analysis is to detect the implicit source of a signal that is responsible for an observed process. The signal is hidden, or *latent*, to the observer and must be recovered. In fact, signal processing of an auditory system of multiple microphones led to the development of independent component analysis (ICA; see section 2.1.3 for more detail) (Hyvärinen & Oja, 1999). Now, suppose that multiple signals are crossed in an auditory speech study, and the aim is to parse the various signals so that one may understand the contribution of the different sources. Rivet et al. (2014) discuss the implications of multi-modality for an audiovisual (AV) study, where additional video signals are mixed among multiple audio signals. The data sets include position, direction, arrival and velocity in the presence of noise emitted from multiple microphones, sound reflected from the walls and background noise. In order to separate the signals from the noise, the possibilities of using independent vector analysis (IVA) or linear predictive coding (LPV) to derive separate linear combinations representative of each signal were explored, but they were not privy to identifying multiple noisy sources (Rivet et al., 2014). A new multi-modal approach must be employed for proper multi-signal and noise decomposition. Other applications in the AV arena include fusion for human-computer interfaces and intelligent environments (Shivappa et al., 2010), geometry calibration of distributed microphone arrays (Plinge & Fink, 2014), and human interactions based on gestures, touch, and facial expressions (Turk, 2014).

Application 2: Environmental Monitoring

Technological advances of the last decade have also enhanced earth monitoring through high-tech satellites and remote sensing. Hyperspectral, panchromatic and multispectral data are monitored across the electromagnetic spectrum, light wavelengths, and image reconstruction through spectral bands, respectively. A combination of these data produce the satellite projection one may see. In order to improve on the high-spatial resolution of the satellite, Yokoya et al (2011) use coupled non-negative matrix factorization and were able to achieve minimal spectral distortion. Geospatial researchers, Debes et al (2014, access multiple sources of Light Detection and Ranging (LiDAR) elevation information to extract the morphological features using a graph-based feature fusion method.

Other topics of remote sensing that require multi-modal analysis are wireless communication networks (Messer et al., 2006); monitoring of natural resources and damage detection for natural disasters (Dalla Mura et al., 2015); component substitution and multiresolution analysis for pansharpening of images (Vivone et al., 2015).

Application 3: Medical Detection

The detection of measurements from medical equipment are vital to make proper medical diagnoses. This procedure usually bases the outcome on the fusion of biological information from the patient with multiple readings from scanners, machines and/or probes. One application deals with the reading of an electrocardiogram (ECG) to exploit the intracardiac surface of patients so that physicians may detect if a patient is experiencing atrial fibrillation. In this application, periodic component analysis was utilized to separate the periodic pattern of the beating heart from the abnormal pounding experienced by the patients (Garibaldi & Zarzoso, 2013). Lei et al. (2012) explore both data-driven and model-based methods for integrating electroencephalography (EEG) and BOLD functional magnetic resonance imaging (fMRI) drawn on a different time scale and containing multiple measurements for each. The spatiotemporal resolution was balanced out by a hybrid fusion technique that extracted the components by applying ICA and used them as inputs in a general linear model (GLM). Other examples include multivariate brain fusion to understand the structural-functional brain association (Sui et al., 2010); spatial-anatomical regularization for disease classification of Alzheimer's Disease (Sun et al., 2018); various image classifications for medical diagnoses dependent on multiple organ functioning (Yadav & Yadav, 2020).

1.1.3 Imaging Genetics

Studying the classification and progression of disease is a prominent medical aspiration, requiring billions of healthcare dollars annually and the scientific manpower to continually test and analyze new data sources. In this booming age of technology and information, medical data sets are increasing in dimensionality and complexity. The influx of this data necessitates the extension of current statistical methodology and analytical computing power to aid scientists in new medical discoveries. Advanced imaging techniques allow for further understanding of psychiatric illnesses that are known to cause changes at the brain level. Many genetic disorders are now able to be monitored before subjects experience symptoms due to a calculated risk from a genetic evaluation. One area

experiencing substantial growth in the development of ideas and applications is a joint analysis of neuroimaging and genetic data, called *imaging genetics*. In this section, I define the new field of imaging genetics and provide the scientific motivation for pursuing this research area as the focus of a multi-modal data analysis.

Patients diagnosed with psychiatric illnesses are often plagued by crippling mental disabilities which can have devastating effects on their daily lives, diminishing their ability to maintain healthy relationships and achieve career aspirations. For decades, the source and development of these illnesses have been under speculation. The majority of diagnoses are based on the individual's account of fluctuating behavioral outcomes and lifestyle interruptions. This ambiguity of mental illness has consequently left scientists with the task of developing preventative tests and discovering treatments. In order to treat psychiatric illness as a medical disease, the classification of the illnesses must move beyond obscure behavioral measures to uncover the underlying biological mechanisms.

Biological components of psychiatric illnesses are being unmasked through recent neuroimaging studies. Technological advances in magnetic resonance imaging (MRI) have revealed that neurology may in part explain the weakening of mental faculties. Disorders such as schizophrenia, bipolar disorder, Alzheimer's Disease, psychosis and others have shown differences in both structural (sMRI) and functional MRI (fMRI) outcomes in comparisons between healthy and affected individuals (Hariri & Weinberger, 2003; Meyer-Lindenberg, 2012; Meyer-Lindenberg & Weinberger, 2006). Areas of brain activation, observed in fMRI studies, are often fewer in those with a psychiatric illness, suggesting that the neurological connections formed for healthy responses are lacking (Chung et al., 2018; Hibar et al., 2015; J. Liu et al., 2009). Neurodegeneration of brain regions of interest (ROI) is being linked to specific disorders. For instance, structural atrophies of the hippocampus in the medial temporal lobe have shown to be correlated with the progression of Alzheimer's Disease (Gatz et al., 2006; Grasby et al., 2018; M. Weiner & ADNI, 2013). While there is no doubt that changes occur at the brain level for those diagnosed with psychiatric illnesses, it remains a challenge for neuroscientists to quantify the differences while controlling for heterogeneity of brain anatomy and function.

One unchangeable physical trait of humans is our genetic make-up. Geneticists have found that this biological underpinning has a link to mental illness. Upon the completion of the human genome in 2004 , follow-up genetic studies have tested the association of genetic markers with psychiatric diagnoses (et al. F.S. Collins, E.S. Lander, J. Rogers, 2004). Indeed, many neuropsychiatric illnesses have shown high genetic familiality (Meyer-Lindenberg &

Weinberger, 2006). Genome-wide association studies (GWAS) search for an association between a phenotype, often the presence/absence of a disease, and a genotype, some measurement of the human genetic constitution, such as single-nucleotide polymorphisms (SNPs). Each individual has approximately 4-5 million SNPs in their genetic make-up. However, a smaller portion of variable SNPs have allowed geneticists to narrow down their search for marked associations (Laurie et al., 2010; Lehne et al., 2011). The standard genome-wide study may analyze 100,000-500,000 SNPs per individual. Databases have been formed that track meta-analyses of these studies to keep lists of significant genes and the corresponding SNPs implicated with diseases (Bertram et al., 2007; Lambert et al., 2013).

Similar to previous GWAS studies, imaging genetics incorporates an association analysis involving a genotype-phenotype relationship. However, the phenotype under scrutiny is now a physiological response of the brain collected from an MRI (Bigos & Weinberger, 2010; Hariri & Weinberger, 2003). That is, the mapping of neural structure and functioning is analyzed as a function of risk genes in order to bridge the gap between pathological behavior and psychiatric diagnosis. This technique necessitates preliminary research into the neural mechanisms and genetic underpinnings that elucidate the disease. One common phenotype is a volumetric measure of grey matter in the brain, measured by sMRI; another structural method, diffusion tensor imaging (DTI), monitors the microstructural changes characterizing abnormal tissues. The functional impacts of the disease can be monitored by electrical changes in the brain via electroencephalography (EEG), regional metabolic changes in positron emission tomography (PET), or hemodynamic BOLD contrast fMRI (Crosson et al., 2010; Lindquist, 2008).

If brain changes and genetic markers have both shown promising results for understanding psychiatric illnesses, would the combination of these data types strengthen the evidence for a diagnosis? Can we assess an individual's risk of developing an illness based on their genetics and simultaneously track the progression of the illness through brain scans? If scientists now have access to biological outcome measures, will preventative or treatment plans be developed? Imaging genetics is the study of the interaction of neuroimaging and genetics data and their combined role in understanding the basis and progression of mental illness. The goal of these studies is to provide a foundation of knowledge from which scientists, doctors and biopharmaceutical researchers can create prevention and treatment therapies in hopes of improving the lives of those affected by a neuropsychiatric disease.

Although imaging genetics remains in the early stages of methodological development, the literature highlights promising modes for detecting the additive and interactive effects of brain changes and genetic risk contributing to the abnormal developmental trajectories underlying neuropsychiatric disorders. The incorporation of this research utilizes concepts and analytical tools from a multitude of disciplines: psychiatrists, neuroscientists, geneticists, biostatisticians, radiologists, computer scientists, cognitive psychologists and human physiologists. In support of this large-scale innovative research, the National Institute of Mental Health (NIMH) has called for the establishment of imaging genetics databases from which every scientist interested in contributing to this research problem may have easy access to real-life data (Van Essen et al., 2013).

1.2 Current Statistical Tools

The act of intertwining genetic and neuroimaging information into one analytical enterprise employs a statistical tool called *data fusion*. Modern scientists seek to learn from these combined efforts in order to make sense of complicated biological issues and to expand their knowledge basis of diseases. Imaging genetics necessitates extended methodology in data fusion that enables researchers to integrate multiple complex data sources in order to make novel discoveries and support groundbreaking treatments for psychiatric illnesses. Fusing human pathological behavior, anatomical and functional brain abnormalities and genetic underpinnings provides statistical power to such studies. Research in this area has progressed from mass association tests, to detecting multivariate patterns, to following the course of disease pathogenesis longitudinally, and, finally, to robust multi-source feature networks through large-scale association studies, discriminative analysis and deep learning.

1.2.1 Mathematical Formulation of Imaging Genetics

In this section, the mathematical formulation is introduced by considering each data set separately, then data fusion will be introduced through past models applied in the literature. The goal of this dissertation is to aggregate and jointly analyze three types of data: sMRI, referring to regional brain volume and/or thickness, to gauge anatomical atrophy; FDG uptake measured by PET imaging, which monitors synaptic dysfunction; and genetic expressions at the SNP level.

The genetic constitution of an individual in its largest form is the chromosome, and each human has 23 chromosome pairs. These chromosome pairs contain two tightly coiled strands of deoxyribonucleic acid (DNA) that are

wrapped around proteins. The DNA contains the genes, which are clusters of SNPs that provide the genetic "coding" determined by bases adenine (A), cytosine (C), guanine (G), and thymine (T). Each SNP has two alleles, which are each coded by one of these letters. The coding that is present for a specific SNP in the majority of the population is considered the major allele. A minor allele is the recessive coding for that particular SNP. An allele in its raw form will usually be coded as either an A,C, G or T, but this is translated into the number of minor allele frequencies 0,1, or 2 for each SNP, or pair of alleles. This recoding is done in PLINK, an open-source whole genome association analysis toolset (Purcell et al., 2007).

Consider a sample of $j = 1, ..., m$ SNPs on $\ell = 1, ..., N$ individuals, then the recoded allele frequencies would yield the matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)$, where $\mathbf{x}_{\ell j} \in \{0, 1, 2\}$ so that each column in the matrix is the jth SNP expression for all n subjects, and each row contains all m SNP expressions for the ℓth subject. In a GWAS study, typically this is modeled as the explanatory variable on a phenotypic response, a binary coding representative of the presence or absence of a disease. Under this framework, a hypothesis test may be carried out testing the association of the disease with the $\{0, 1, 2\}$ coding of the jth SNP as a χ^2 statistic. Since many tests are run, the p-value threshold of significance is set very low; the adjustment is based on the correction made for multiple testing (Lehne et al., 2011).

Now, take the same ℓ individuals and consider sMRI data, which yields voxel-by-voxel grey matter brain volumes and cortical thickness (mm^3). Let the number of voxels be $v = 1, ..., p$ at baseline for subjects, then we can signify the matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_p)$, where $\mathbf{y}_{\ell v} \in (0, \infty)$, a continuous measure on the positive real number line. Each column represents the volume of the pth voxel for all N subjects, and each row represents all v voxels for the ℓth subject. Similarly, take the FDG uptake collected by PET imaging. Let the matrix of uptake be denoted $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_q)$, constrained to the continuous values $\mathbf{z}_{\ell k} \in (0, 2)$. Keep in mind that both sMRI and FDG-PET uptake are taken at one time-point; therefore, both are modeled using a "structural approach". Additionally, both data sources may instead be summarized over $k = 1, ..., q$ regions of interest (ROIs), rather than at the voxel level, from the same set of subjects at baseline.

Complications arise due to the matrices being on different scales of magnitude with unbalanced data quantities. Note this data description is merely at the baseline level. While the genetic data are collected only at one office visit, the sMRI and FDG-PET imaging data are collected longitudinally throughout the study so that changes may be tracked. Thus, $\mathbf{Y}$ and $\mathbf{Z}$ may be expressed as

arrays with $t = 1, ..., r$ time points of the data collection. In addition, the data are collected in different frequencies for the various diagnoses. Denote the vector of baseline diagnosis $\mathbf{D} = (d_1, d_2, ..., d_N)^T$. This may be extended at each time point as well because diagnosis is evaluated at each point of data collection, so that $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_r)$ where $\mathbf{d}_{\ell t} \in \{CN, MCI, AD\} = \{0, 1, 2\}$ for the case of Alzheimer's Disease. The columns of this matrix represent the diagnosis for all N subjects at time point t.

The mathematical formulation of imaging genetics displays the major problem of high-dimensionality, that is, $N <<< m \times p \times q$. In the next section, statistical methodology of imaging genetics is introduced, beginning with genome-wide, brain-wide analyses. Then, I introduce multivariate models and dimension reduction techniques for a multi-modal analysis that allow for more than two sources of data to be fused for a deeper understanding of AD pathology.

1.2.2 Early Methodology

Early methodology for modeling stemmed from mass univariate hypothesis tests that sought to discover genome-wide, brain-wide associations. The most granular form tested each biological relationship unit-by-unit. For instance, each SNP was matched with each voxel one-by-one to explore the relationships between the two data sources. Later, studies focused on specific candidate genes shown to have associations in the past with suggestive ROIs. While some important contributions were made as a result, these simplified studies may not have adequately uncovered the complex pathology that is present when linking genetic and neuroimaging material. Multivariate approaches soon followed as researchers incorporated reduction of high-dimensional data sets with faster computational frameworks for joint analysis. Reflecting both natural and disease-oriented heterogeneity remained a challenge, engendering novel Bayesian methods to issue uncertainty quantification. As studies collected data on patients over the course of one, five, and even ten years, the rate of neurological changes and how genetics play into this risk of mental decline became the focus of study (Nathoo et al., 2019).

Mass Univariate Methods

The massive univariate analyses formed the basic building blocks of imaging genetics studies. The primary advantage of this approach is the ability to search the genome and the brain in their entirety. However, performing this many tests raises several issues when operating under this simplification: multiple

testing, speed of computation and restricting assumptions. Multiple testing occurs when one runs many hypothesis tests independently. For each test, there is a significance threshold, say α for simplicity. When running all h hypothesis tests, the probability of rejecting a true null for all of the tests combined is $1-(1-\alpha)^h$. That is, the probability of making a Type I Error is inflated exponentially. Correcting for multiple testing means setting an adjustment to the significance level based on the number of tests run. However, when such stringent thresholds are imposed, it can be difficult to gain even one significant finding among the vast number of tests (Meyer-Lindenberg et al., 2008). The millions, or even billions, of analyses further requires higher computational power, and these computations can take up to weeks even with parallel processing in place. In addition, when running many individual SNP-by-voxel test, this ignores the interactions among the SNPs in the genome and the voxels in the brain by not imposing a multivariate covariance structure that would explain the variability among the voxels or SNPs. Early researchers tackled these issues head-on by advancing current methodology and discovering new ways to model these complex hindrances.

Stein et al. (2010) paved the way for a voxel-wise GWAS under the univariate approach. Each voxel-by-SNP test was set up as a linear regression problem with the response of the brain volume per voxel, $y_\ell(v)$. They examined 448,293 SNPs, modeled as the minor-allele frequency in $x_{\ell j}$, in each of 31,622 voxels, v, of the entire brain for 740 subjects, including covariates for age and gender so that the model is given by

$$y_\ell(v) = \beta_0 + \beta_1 Age_\ell + \beta_2 Gender_\ell + \alpha_{vj} x_{\ell j} + e_\ell(v,j). \tag{1.1}$$

This resulted in $\sim 1.4 \times 10^{10}$ tests with the null hypothesis $\alpha_{vj} = 0$. The top 20 associated SNPs were analyzed in correspondence with voxels in the brain, which were found to be located primarily in the temporal lobe (Stein, Hua, Morra, et al., 2010). For simplicity, the minimum p-value is saved at each voxel, and an adjustment was made to the p-values. The adjustments are based on the CDF of the minimum p-values, followed by an FDR correction. In part, this method was applied instead of permutation, the "gold-standard" for deriving the null distribution, in order to lighten the load of the computational burden. With 300 cluster nodes in parallel, each analysis took 27 hours to complete. The clinical subgroups were compared by analyzing the trend level difference of the allele frequencies among the diagnostic groups and the related odds ratios for the top significant SNPs (Stein et al., 2010).

As a follow-up study, Stein et al (2010a) reduced the number of tests by applying his analysis as a SNP-by-voxel test with voxels located only in the temporal lobe and then again with voxels only in the hippocampus. Two SNPs met the still very stringent measures and other SNPs were identified as associations with voxels “of interest”.

Clinical groups were compared using the Chi-Square Test for Independence (Stein, Hua, Morra, et al., 2010). Hibar et al (2012) extended Stein's work by applying this methodology to genetic markers at the gene level, so that regression was performed on each voxel-by-gene pair. In order to uncorrelate the clusters of SNPs, PCA was used to summarize the SNP data for each gene and the resulting "eigenSNPs" which contained 95% of the explained variation in the SNPs were used as the predictors (Hibar et al., 2012). Let $g = 1, ..., w$ denote the smaller subset of genes. Then the fourth term on the RHS of equation (2.1) becomes $x_{\ell g}$, the g uncorrelated eigenvectors of the PCA analysis, yielding

$$y_\ell(v) = \beta_0 + \beta_1 Age_\ell + \beta_2 Gender_\ell + \alpha x_{\ell g} + e_\ell(v, g). \tag{1.2}$$

The multiple partial-F statistic was calculated for each gene at each voxel for 18,044 genes. The same procedure was carried out as in Stein's work for the multiple testing correction and the determination of significance.

The latter studies were able to achieve a large-scale analysis, but their models failed to capture the embedded spatial representation of the images or the genetic pathways of the SNPs. With this goal, Ge et al. (2012) used random field theory (RFT) on a multi-locus least squares kernel machines model to study the interaction of multiple genetic variants. In 2015, this model was extended to allow for potential interactions between non-genetic variables such as disease-risk factors, environmental exposures, and epigenetic markers (Ge et al., 2015). Both were applied to 18,043 genes by 448,294 SNPs with Bonferroni-corrected p-values to control the family-wise error (FWE). Findings revealed significant interactions between genes and risk factors on hippocampal volume. In the paper, the authors did not specify how the model was computed. An alternative fast voxel-wise GWAS (FVG-WAS) approach was developed by Huang et al. (2015) where the explicit aim was to reduce the computational load. This objective was carried out in three components: a heteroscedastic linear model, a global sure independence screening procedure and detection based on wild bootstrap, with FWE correction. The phenotype responses were brain volumes of 193,275 voxels containing 93 ROIs against the top five PCs of 501,584 SNPs. The fast multivariate computation took approximately 56 hours with one CPU (Huang et al., 2015).

Multivariate Methods

While vast univariate approaches utilize all of the available data, the phenotype-genotype dependence relationships are almost certainly ignored, and significance thresholds tend to mask significant findings. In addition, comparing disease groups is more challenging when running separate tests for each SNP-by-voxel or gene-by-voxel relationship. Since unrealistic assumptions are imposed on the complex temporal and spatial structures, "potential efficiency gains arising from borrowing information across brain regions are not realized" (pg4) (Nathoo et al., 2019). Some discoveries have found that only a few SNPs are highly associated when there are likely groups of SNPs whose

combined effects drive the phenotypic relationships. One approach to study the connections between neuroimaging and genetic data is a multivariate high-dimensional regression model. In this multi-dimensional framework, the scale of the data is reduced by summarizing brain measures across key ROIs and grouping SNPs based on linkage disequilibrium and gene locations (Nathoo et al., 2019).

The first comparison of univariate and multivariate approaches by power analysis was carried out by Vounou et al. (2010). In their study, a sparse reduced-rank regression (sRRR) model was proposed regressing $q = 111$ anatomical ROIs on $m = 437,577$ SNPs from simulated data. Sparsity was imposed on the matrix of regression coefficients so that each redundant (k, j) pair shrunk to 0, and the rank was reduced so that it would be less than $min(k, j)$. The simulation results showed that basing the approximation of the high-dimensional regularized regression coefficient matrix on a low rank, sparse matrix achieved higher power gains to detect meaningful genetic markers. Wang et al. (2012) proposed a group-sparse multitask model where estimates of the regression coefficients were based on penalized least squares with weight matrix $\mathbf{W}$. Regularization was of the form

$$\min_{\mathbf{W}} \sum_{\ell=1}^{N} ||\mathbf{W}^T \mathbf{x}_\ell - \mathbf{y}_\ell||_2^2 + \lambda_1 \sum_{i=1}^{h} ||\mathbf{W}^i||_{G_{2,1}} + \lambda_2 \sum_{j=1}^{m} ||\mathbf{w}^j||_{2,1} \qquad (1.3)$$

for $i = 1, ..., h$ genetically-linked SNP groups and $j = 1, ..., m$ SNPs. Sparsity was assigned by a group $\ell_{2,1}$ norm, denoted $G_{2,1}$, on the weight matrix in the following form:

$$||\mathbf{W}||_{G_{2,1}} = \sum_{i=1}^{h} \sqrt{\sum_{i} \sum_{j=1}^{m} w_{ij}^2} \qquad (1.4)$$

The penalty selected allowed for a nested sparsity structure, grouping the coefficients of a given SNP across all phenotypes. At the gene level, all SNPs were grouped within a given gene using the top 37 genes associated with the disease (the remaining number of genes after applying quality control). This nesting of regularization leveraged the interrelationships among the ROIs. When compared to standard univariate methods, the prediction power of the proposed method showed improved detection (Wang et al., 2012).

One drawback to Wang's approach was that only a point estimate was provided for the association between a SNP that was summarized across all ROIs. This circumvented inferential applications to the otherwise informative model because variability estimates could not be derived under this approach. Using a Bayesian hierarchical model, Greenlaw et al. (2017) extended Wang's work, now allowing uncertainty quantification to obtain interval estimates for the regression coefficients of the selected genetic markers. The enhanced procedure came at the cost of a lengthier computation of the posterior distribution via the Gibbs sampling algorithm (Greenlaw et al., 2017). Zhu et al (2014) used a low rank regression model with inference applied in the Bayesian

framework involving 93 ROIs and 1,071 SNPs. Robust standard errors extracted from the latent structure allowed for some inferential procedures. However, none of these methods compared the phenotype-genotype relationships between subject groups. In addition, the phenotypes mainly focused on structural brain measures, avoiding alterations in the underlying functional connectivity.

The association between fMRI measures and SNP data were analyzed along with a comparison of these measures for disease statuses by Chen et al (2012) and Stingo et al (2013). The multivariate approach implemented by Chen et al combined PCA with parallel ICA to identify groups of SNPs that contributed to four neurotransmitter pathways in the brain. Disease subgroups were incorporated by reducing the test to include the top 300 SNPs that showed significant differences among healthy and infected individuals before applying the model. In addition, a dichotomous predictor representing the diagnosis was included in the final model (Chen et al., 2012). Stingo et al. (2013) exploited a hierarchical Bayesian mixture model in order to study the association of spatial patterns in brain connectivity to genetic factors. The mixture components of the proposed model were then used to classify the study subjects into subgroups based on disease status, and the allocation of subjects to these mixture components were linked to genetic markers with structural dependencies on ROIs (Stingo et al., 2013).

Longitudinal Analysis

The call of the NIH for collaborative, large-scale studies initiated longitudinal data collection, where patients are examined at baseline and then are repeatedly tested in subsequent time intervals. The purpose of longitudinal data collection is to track the progression of neuropsychiatric diseases to be able to catch the disease early and to aid in treatment discovery. At what point does the weakening of mental structure and function contribute to noticeable behavioral and psychological detriments to the extent that a diagnosis is warranted? Once a patient is diagnosed, can we develop a timeline based on the rate of brain deterioration in concordance with genetic risk factors?

Szefer et al. (2017) developed a bootstrap-enhanced sparse canonical correlation analysis to identify the linear association among the phenotype-genotype relationships, thereby reducing their analysis to a list of $m = 1694$ SNPs and $k = 22$ ROIs. The phenotype of the linear mixed model (LMM) was the rate of change of cortical thickness and volume. Fixed effects (βs) of the main and interactive effects of time and disease status and a random subject-specific intercept and slope (γs) for time t were included:

$$Y_{\ell kt} = \beta_{0k}+\beta_{1k}MCI+\beta_{2k}AD+\beta_{3k}t+\beta_{4k}MCI*t+\beta_{5k}AD*t+\gamma_{1\ell k}+\gamma_{2\ell k}t+\epsilon_{\ell kt}. \tag{1.5}$$

Inverse probability weights were included in the model to adjust for the biased sampling design. The predicted rates of change by subject for each ROI were taken as the sum of the disease-specific and subject-specific estimated rates of change: $\hat{\beta}_{3k} + \hat{\beta}_{4k}MCI + \hat{\beta}_{5k}AD + \hat{\gamma}_{2\ell k}$ (Szefer et al., 2017). However, the assumed covariance structure placed on the model was not discussed in the paper.

One of the earliest studies to incorporate baseline values and subsequent longitudinal data into the prediction of mild cognitive impairment (MCI) conversion to Alzheimer's Disease (AD) was by Ye et al (2012). The research employed sparse logistic regression and stability selection for predicting MCI to AD conversion using baseline data. This study combined MRI, cerebral spinal fluid (CSF), demographic, genetic and cognitive measures into their model for robust feature selection in order to perform classification based on four years of data collection. To build the classifier, support vector machine (SVM) was applied and the performance was evaluated by the leave-one-out area under the curve (AUC). Stability selection of the regularization parameter was computed by bootstrapping. Fifteen various features from MRI scans, APOE genotyping and cognitive measures achieved an AUC of 0.8587 (Ye et al., 2012).

A Bayesian approach was proposed to perform longitudinal analysis of multivariate neuroimaging phenotypes and candidate genetic markers. A low rank longitudinal regression model was specified where penalized splines were incorporated to characterize an overall time effect, and a sparse factor analysis model coupled with random effects was proposed to account for spatio-temporal correlations of longitudinal phenotypes (Lu et al., 2017). A useful feature of this framework was the allowance for interactions between genetic main effects and time. The data were reduced to the top 10 and top 40 candidate genes and SNPs, based on statistical significance, respectively, for 93 ROIs; thus, keeping the dimensions of the analysis manageable. Nonetheless, the model became intractable when monitoring thousands of potential phenotype-genotype relationships (Lu et al., 2017).

1.2.3 Emerging Multi-Modal Inference

As research progressed and imaging genetics analyses continued, researchers realized that modeling genotype-neuroimaging relationships with disease requires more advanced statistical methodology. In recent studies, many of these methods have instead included multiple sources of information. A multi-source analysis utilizes two or more resources of data to understand neuropsychiatric illnesses. The approach of joining the study of two or more biological measures for the discovery of disease pathology is in parallel with modern approaches that have coined the term *multi-modal* to describe the combination of multiple big data sources for analysis. In this dissertation, the phrase *multi-modal neuroimaging genetics* (MMNG) encompasses this school of research techniques for specific application to neuropathological diseases with a genetic basis. Calhoun & Sui have contributed tremendous knowledge on the topic of multi-modal data analysis through several review papers of multi-modal techniques and analysis papers for these flexible, data-driven models. These discussions and results showcase the importance of leveraging cross-information among the modalities. Calhoun warns that the learning curve for such integrative analyses is steep but is worthwhile for the advantages to be gained (Calhoun & Sui, 2016). Modern methods involve extensions to principal component analysis (PCA), canonical correlation analysis (CCA),

independent component analysis (ICA), linear discriminant analysis (LDA), support vector machine (SVM), deep learning, and other machine learning methods.

Independent Component Analysis

The rich information gleaned from these studies requires dimension reduction to retain only the necessary information and minimize lengthy computational procedures. PCA is a popular method that finds orthogonal vectors such that the data are uncorrelated and the number of vectors is chosen by the desired percentage of variation explained by the data. Nonetheless, the pervasive correlation structures among MMNG data may be better expressed by higher-order statistics to identify independent vectors. Therefore, multiple forms of ICA have been created and applied to MMNG studies. Initially, ICA assumes a statistically independent, non-Gaussian and unknown linear mixing process (Calhoun & Sui, 2016) on the data $\mathbf{X}$ such that latent sources, $\mathbf{s}$ may be uncovered through a mixing matrix $\mathbf{A}$, or $\mathbf{X} = \mathbf{A}\mathbf{s}$ (Hyvärinen & Oja, 1999). When $\mathbf{s}$ are assumed to be independent, recovering the "sources" follows closely with the idea of blind source separation (BSS), where an unknown signal must be detected through an unmixing matrix $\mathbf{W} = (\mathbf{A})^{-1}$ for $\mathbf{s} = \mathbf{W}\mathbf{X}$. That is, the columns of $\mathbf{W}$ are an orthonormal basis that maps the raw data $\mathbf{X}$ to a new set of features, $\mathbf{s}$ (Guillén, 2017). The optimization of this approach is discussed in detail in section 3.1.3.

ICA is a multivariate approach that intrinsically summarizes the data without parametric restrictions to uncover the natural patterns in the data. This method can be applied to brain signals in order to derive spatially and temporally independent components, while also exploiting the genetic origins (Calhoun & Sui, 2016). Studies of fMRI data have implemented ICA for at least twenty years and, initially, ICA was used extensively for a single fMRI data source (Calhoun & Adali, 2006). One of the earliest reviews of ICA for MMNG analyses was provided by Calhoun et al (Calhoun et al., 2009). In this review, two extensions of ICA data fusion were discussed, joint ICA (jICA) and parallel ICA (paraICA), where the components provide groupings of brain activity into regions that share the same response revealing a natural measure of functional or structural connectivity (depending on the modality used). While jICA combines multiple modalities with a strong regularization assumption to derive a shared contribution matrix to link the modalities, paraICA updates separate ICA processes using the correlation between the subject profiles for the two modalities, enabling one to identify interactions between brain imaging and genetic information (Calhoun et al., 2009).

The paraICA approach is applied by Liu et al (2009) to investigate an application to schizophrenia data (n=63), targeting SNP arrays with relationships to abnormal brain functionality. When operating under two modalities, paraICA seeks to maximize the independence between components for the two modalities separately, and then determines the relation between them. The relationship of two modalities, $\mathbf{Y}$ and $\mathbf{Z}$, is examined by using paraICA to extract information from both modalities by simulta-

neously solving:

$$\max\{H(Y_k) + H(Z_j) + Corr(A_k, A_j)^2\} =$$

$$\{-E[\ln f(Y)] - E[\ln f(Z)] + \frac{Cov(A_k, A_j)^2}{V(A_k)V(A_j)}\};$$

$$Y = (1 + e^{-U_k})^{-1},\ U_k = W_k X_k + W_{k0},\ A_k = W_k^{-1};$$

$$Z = (1 + e^{-U_j})^{-1},\ U_j = W_j X_j + W_{j0},\ A_j = W_j^{-1}, \tag{1.6}$$

where $H(.)$ is the entropy (defined in Section 3.1.3, Equation 3.16) and mixing matrices (A_k,A_j) and unmixing matrices (W_k, W_j) of ($\mathbf{Y}$, $\mathbf{Z}$), respectively. After identifying relationships between the independent components of each modality, differences in the pair of loading parameters with the highest correlation (.38) were examined with a t-test between schizophrenia and healthy groups. Meda et al (Meda et al., 2012) examined this method on 757 subjects enrolled in an AD study. The loading parameters represented the weights of the overall component for each subject. Four of the pairwise correlations derived between the ideal number of components for both modalities were found to be statistically significant, and key ROIs were identified. The association of the top ten genes relative to disease status were analyzed, confirming past findings and including the discovery of new significant genes. A study by Wolf et al (Wolf et al., 2020) applied paraICA for the comparison of patients with and without Parkinson's Disease for two modalities, sMRI and resting state fMRI (rsfMRI).

Some researchers have critiqued whether paraICA adequately models the heterogeneous correlation structures. With the goal of extracting linked components from multiple modalities where each modality differs in units, signal-to-noise ratios, voxel counts, spatial smoothnesses and intensity distributions, Groves et al (Groves et al., 2011) developed a novel ICA model called Linked ICA. The model is an extension of Bayesian ICA where variational Bayes allows for the configuration of tensor ICA or spatially-concatenated ICA decompositions, or both at the same time. Simulation studies revealed that Linked ICA has more accurate recovery of the modalities (Groves et al., 2011). Miler et al (2020) developed a hybrid of dictionary deep learning and group ICA to understand multi-scale brain connectivity patterns in a schizophrenia study. Multi-channel 1D sparse convolutional dictionary learning (SCDL) assisted ICA in detracting time-varying representations of functional networks. Group-wise correlations were derived for the significant pairs of components (Miler & Calhoun, 2020).

Deep Neural Networks

In the past decade, novel analysis techniques have been engineered to accommodate diverse data that deviate from historical models and to effectively manage the sheer

volume of information collected. Data scientists are impelled by the ingenuity of a modern approach called *machine learning*, a branch of artificial intelligence that is responsible for the development of algorithms that operate much as a machine would in the processing, or *learning*, of data. Under this framework, the idea is to allow the computer system to manipulate the data and build algorithms in an instinctive fashion that is unique to the data at hand. In section 1.1.2, data with a greater depth were introduced as large data sets, or big data, that mandate non-traditional methods for analysis. Machine learning (ML) was devised initially for similar applications, and since has been employed for multi-modal analyses containing at least one big data source (Xue et al., 2019). A purely data-driven approach to ML which executes an automatic end-to-end learning procedure is called *deep learning*. Deep learning (DL) makes data-driven fusion possible for multiple modalities through the extraction and representation of the most relevant features among the data. The term "deep" describes the convoluted hidden layers built into the architecture of the data, instigated by a *deep neural network* (DNN) (Meng et al., 2020).

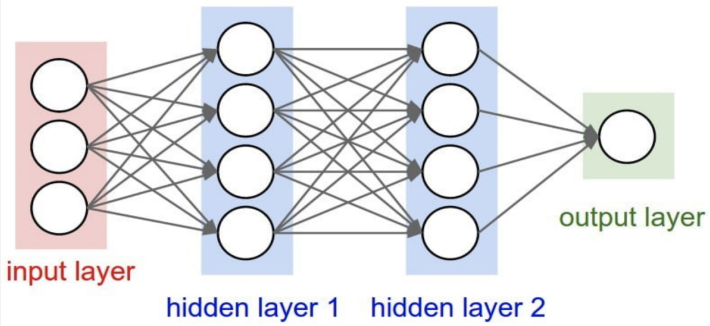


Figure 1.1: Utilizing data as the inputs, the DNN contrives hidden layers to process the end-to-end automatic learning algorithm for the extraction and fusion of the relevant features, or outputs (https://www.bmc.com/blogs/deep-neural-network/).

Figure 1.1 displays the most basic parts of the DNN algorithm: three data sets (left) are the *inputs* of the model; two *hidden layers* (represented by the two verticle columns of four circles) consider multiple combinations of the data subject to a feature algorithm learned by the data; the *output layer* is collectively the fused data features that are reduced from the initial state. Under this framework, algorithms may be broken up into three categories based on the element of the procedure that is emphasized, that is input-, output-, or layer-based (Meng et al., 2020). For the fusion of multiple heterogeneous data sets, there are double layers, or at least two points, of fusion that take place: within the hidden layers as the multiple combinations are considered and then as the final outputs are united to one fused outcome. That is, both input-based and output-based layers are combined to carry out the learning process. For more information on the various methods available see (J. Liu et al., 2020; Meng et al., 2020).

Based on the success of DNN in performing data-driven feature extraction for a unimodal analysis, researchers devised a method to extend this concept to multiple data sets. Zhou et al (2017) developed a stage-wise DNN for multiple modalities that performed fusion at three different stages. First, each modality was used as an input into its own output-based DNN. This stage utilized all available experimental units, though they may have been unbalanced across modalities. That is, under the DL framework (DNN in particular), the observation indices allow modality-specific samples, $i = 1, ..., n_m$. Furthermore, the first stage of the algorithm reduced the data to a more tractable dimension that highlighted only the relevant features specific to that modality. In the next stage, each possible pairwise consideration of the features extracted from the first phase were used as inputs and outputted as a two-way data fusion (double-layered). In the final stage, all outputs from the second stage were concatenated and further fused based on the relevant features from the $\binom{M}{2}$ combinations. That is, the final output was the complete data fusion of all M modalities. Figure 1.2 displays this process.

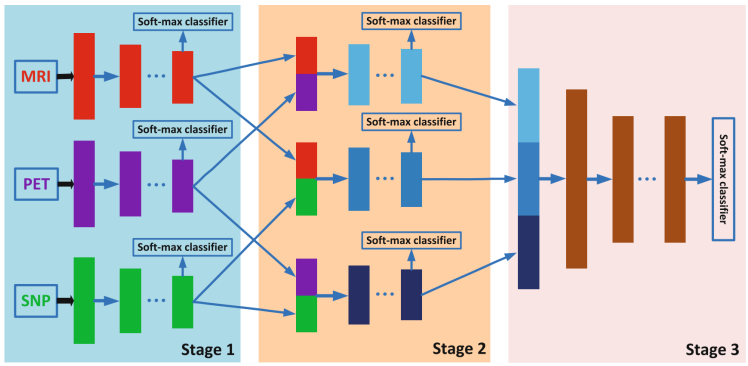


Figure 1.2: A three-stage DNN for a 3-way multi-modal data fusion using a soft-max classifier (Zhou et al., 2017)

The algorithm underlying the hidden layers is another changing factor representative of various DNN approaches. Each DNN in the three-stage network described is filtered through a soft-max classifier as the output. The classifier helps to keep track of the representative modalities by attaching labels to each feature. This is the automatic procedure within the hidden layers that permits the algorithm to learn from the data with each developing stage. A computer program is able to carry out this learning procedure so that the researcher merely inputs the data and remains "hands-off" until the data fusion is completed. For more details on the soft-max procedure, see (S. Liu et al., 2015). The same authors of the three-stage procedure later enhanced the deep learning algorithm by building a latent representation layer into the soft-max classifier. The latent sources of the features were learned through a multi-class group sparse feature selection formulated as

$$\min_{\mathbf{G}^{(m)}} \Big\{ \sum_{m=1}^{M} ||\mathbf{G}^{(m)T}\mathbf{X}^{(m)} - \mathbf{Y}||_F^2 + \beta||\mathbf{G}^{(m)}||_{2,1} \Big\}, \tag{1.7}$$

where $\mathbf{G}(m)$ denotes the feature matrix of the mth modality, $\mathbf{Y}$ is the corresponding label matrix in the soft-max classifier, $||\cdot||_F$ is the Frobenis norm, or the square root of the sum of the absolute squares of its elements, and $||\cdot||_{2,1}$ is the square root of the sum of squares. This negates the need for a stage-wise implementation through the augmentation of a Lagrange multipler (Zhou et al., 2019). Other DL algorithms that have been applied to multi-modal applications may be perused: sparse regularization and support vector machines (SVM) (Hao et al., 2020a); stacked autoencoders (S. Liu et al., 2015); translation-based ML through a dynamic functional network connectivity matrix (dFNC) (Plis et al., 2018), and others (Vielzeuf et al., 2019; Xue et al., 2019; W. Zhang et al., 2018)

Classification by Feature Extraction

One major challenge for MMNG studies is to parse the interrelationships of neuroimaging and genetic information while studying the disease statuses for the psychiatric illness under review. Past research highlights multiple sources of biological data which could potentially replace the standard criteria of cognitive exams for classification of patients into diagnosis groups. Thus, researchers of MMNG must have the ingenuity to utilize multiple modalities for the discrimination of disease groups while simultaneously modeling the associations among the modalities. Various strategies have been implemented with this goal in mind, including supervised learning algorithms, such as LDA, and deep learning neural networks with latent representation for classification.

A methodological framework for discriminant analysis was implemented by Dai et al (Dai et al., 2012) for a multi-modal imaging technique, studying multi-level characters with a multi-classifier (M3). Four imaging modalities, one of sMRI and three measures of resting-state fMRI (r-fMRI), were combined consisting of information on 90 ROIs on AD patients and healthy controls. Feature selection was done using a nonparametric rank-sum test followed by a maximum uncertainty LDA (mLDA) classification procedure. LDA is a common classification method that seeks to distinguish groups by simultaneously maximizing their between-class separability S_b while minimizing their within-class variability S_w. Let the number of diagnosis groups equal $d = 1, ..., g$, then LDA seeks to find a projection matrix P_L such that

$$S_b = \sum_{d=1}^{g} N_d(\bar{\mathbf{x}}_d - \bar{\mathbf{x}})(\bar{\mathbf{x}}_d - \bar{\mathbf{x}})^T; \quad S_w = \sum_{\ell=1}^{N_g} N_d(\bar{\mathbf{x}}_{d\ell} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{d\ell} - \bar{\mathbf{x}})^T;$$

$$P_L = \arg\max_P \frac{|P^T S_b P|}{|P^T S_w P|} \tag{1.8}$$

This is done by deriving eigenvectors and eigenvalues, P and Γ respectively, such that $(S_w^{-1}S_b) = P\Gamma$. For high-dimensional data this maximization is a challenge because with $n << p$ the within scatter is singular (the rank is less than the number of features). Thus, S_w was replaced with $S_w^* = (P\Gamma^* P)^T(N-g)$, where Γ^* is a new matrix of eigenvalues based on the largest dispersion values (Thomaz et al., 2006). Next, leave-one-out cross-validation (CV) estimated the performance of the classifier and identified the maximally distinguishable features between groups.The binary classification accuracy was 90% with sensitivity of 88% and specificity of 91% (Dai et al., 2012).

Often there are more than two disease group statuses, such as the prodromal stage of MCI within the ADNI data (Dai et al., 2012). For multi-class identification, Zhu et al (2016) generated a regularized sparse multitask learning procedure by combining two subspace learning methods, LDA and locality preserving projection (LPP). LDA condenses the high-dimensional data into a low-rank space, while LPP reduces overall noise. Through the incorporation of two imaging modalities, sMRI and FDG-PET, the authors achieve 69.49% and 61.86% accuracy for AD/MCI/NC and AD/LMCI/EMCI/CN, respectively. Still, the proposed framework outperformed standard methods including PCA, LDA, M3T, LDA and LPP alone. It's also important to note that when considering only two classes at a time, the highest accuracy achieved was 96% for AD vs NC, 80% for MCI vs NC, and 71% for EMCI vs LMCI (Zhu, 2016).

One supervised approach allowed for the inclusion of three biomarker modalities, sMRI, FDG-PET and CSF, among three classes, control normals (CN), MCI, and AD. Under this framework, Zhang et al (2011) embeds a kernel combination into the standard SVM (k-SVM) for data fusion. Classification accuracy was checked using a 10-fold cross validation and was found to be 93.2% for AD vs NC and 76.4% for MCI vs NC. Higher accuracy was achieved for MCI classification as the number of ROIs increases. Furthermore, it was shown that the combined modalities offered more powerful discriminatory properties than any one modality individually (D. Zhang et al., 2011). A much more recent advancement of MMNG methodology involved k-SVM but with additional steps prior to its implementation. Hao et al (2020) studied two AD samples with the joint analysis of sMRI and FDG-PET data. By implementing random forest, the authors derived a similarity matrix for each modality, such as the Manhattan distances, in order to uncover the differences within. Then, the authors performed joint selection of features through multi-task learning (MTL) with regularization imposed for sparsity and an added similarity constraint for modality leverage. The following data fusion was carried out via multi k-SVM (mk-SVM) for classification into multiple subgroups. This multi-modal neuroimaging feature selection with consistent metric constraint (MFCC) using mk-SVM attained the highest accuracy among competing procedures for all four classes: AD vs NC is 97.6%, MCI vs NC is 84.47% and EMCI vs LMCI is 77.76%. Furthermore, their method was validated on a second AD sample (Hao et al., 2020a).

Unsupervised methods of feature learning have also been proposed in recent studies of imaging genetics data that used deep learning algorithms to extract the layers of

hidden features. Liu et al (S. Liu et al., 2015) constructed a framework that utilized a stacked auto-encoder (SAE) with a softmax logistic regressor that was then followed by mk-SVM. The method randomly hid one modality at a time to allow the neural network to recognize the missing modality by inferred correlations between the features. Another study implemented a stacked deep polynomial network (s-DPN) to obtain more robust feature representation (Zheng et al., 2016). Both of these methods employed information from only two modalities, PET and sMRI. In addition, only samples with complete data across all modalities were applied. This is not a realistic assumption for most multi-modal data collections and could greatly reduce the sample size available.

Zhou et al (2017) used a three-stage deep neural network for feature extraction of hidden layers using the maximum number of available samples per modality. Each stage implemented feature extraction from sMRI, FDG-PET and SNP data separately, pairwise, and concatenated jointly by assigned labels, respectively. While the advantage of this framework was the volume of information used in the study, the classification accuracy was lower than past studies. In addition, the authors only compared their methods to base procedures, such as PCA and Lasso. Later, the model was improved by incorporating a latent representation learning method for classification (Zhou et al., 2019). This method reduced noise, made use of all the available subjects to train the model, and coherently achieved feature learning and classifier training to capture the inherent association among the modalities. The highest classification quantity this model achieves is an AUC of .755 (Zhou et al., 2019).

N-way Component Extensions

Although the deep learning methods are currently popular for multi-modal classification, the methodology may not align with the goals of MMNG. So far, the classification accuracy is not as high as other multivariate methods. Other than examining the accuracy, the interpretability of the models falls short. It's difficult to analyze the relationships among the modalities in terms of identifying which SNP-ROI pairs subsume the overall association results. In addition, as the number of modalities increases, uncovering the hidden multi-layers in deep learning quickly proliferates in complexity. Data fusion based on component analysis has a direct subject-specific interpretation of the loading parameters. The weights indicate which of the multivariate phenotype or genotype markers have the strongest associations. Obtaining quantifiable group comparisons is reachable. Extensions of component analysis models can support n-way MMNG models, so that three or more modalities are fused to unveil the pathogenesis of neuropsychiatric disease.

An analysis of the phenotype-genotype relationships among modalities may be carried out with mCCA+jICA. Under this framework, each modality is reduced to a feature by maximizing the correlated components among the modalities in order to enhance tractability (Y. O. Li et al., 2009). The data sets are then jointly decomposed

into matrices (A_k) of mixing coefficients and components (C_k), or spatial maps, by which pairwise modality correlations are maximized, where $k = 1, ..., K$ equals the number of components extracted. The columns of A_k, canonical variants, are linked among the modalities and can be used to examine the intra-relationships and detect group differences. The first tri-modal analysis of schizophrenia was carried out with mCCA+jICA for the joint analysis of r-fMRI, sMRI and DTI data. The influence of cognitive tests were examined among the modalities and significant differences were observed between healthy and affected individuals. The model was extended to support the addition of jICA with mCCAR with reference to concurrently maximize inter-modality covariation and correlations with cognitive tests. Using the joint loadings, a linear regression model was built to predict cognitive scores and further examine a difference between the disease subgroups (Sui et al., 2015).

Another form of dementia, called vascular cognitive impairment (VCI), is often difficult to distinguish from AD. In a breakthrough study, Raja et al (2020) combined four measurements from DTI and another four measurements from diffusion kurtosis imaging (DKI) to use a total of eight modalities for the multi-class comparison of AD, VCI and CN. The analysis utilized mCCA+jICA in order to examine in-depth differences among the white-matter tracts in the brain connectome. The data fusion method obtained a higher AUC value (AUC=.913) than any one of the unimodal features (AUC=.77). Three modal components were extracted using mCCA+jICA by Gao et al (2020). The fused modalities included grey matter brain volumes from sMRI, white matter from ROI by DTI, and rsfMRI. Then SVM with recursive feature elimination and nested 10-fold CV extracted the most significant features. With this reduced information, k-SVM was implemented to achieve a classification task into two groups. The binary classification task attained 96.56% and 87.63% accuracy on training and testing data, respectively (Gao et al., 2020).

1.3 Remaining Challenges

Although MMNG data fusion methodology has advanced tremendously in the past few years, these approaches have some drawbacks that future researchers must take into consideration. In addition, there is a lot of work that remains to be done in order to formalize the statistical methodology under this framework. The issues of MMNG analyses that I will discuss in this section are challenges that I plan to tackle in this dissertation: exploring the ICA algorithms in a simulation setting and multi-modal framework with the inclusion of genetic data with various imaging modalities; comparing the two different mainstream fusion frameworks under ICA; assessing the statistical reliability of traditional ICA algorithms and comparing this with novel optimization techniques; discussing the management and pre-processing of imaging genetics data; creating new computational tools for MMNG analyses. All of this will be done using multiple sources of data collected from a longitudinal, multi-site study of AD. Em-

phasis will be placed on the interpretability and generalizability of the decomposition toolbox for imaging genetics studies with the additional creation of an Imaging Genetics Pipeline. Finally, in this work, I highlight the need for improved open science tools in order to foster replicability of this high-impact multi-disciplinary research area.

While these methods jointly analyze multiple imaging modalities, very few studies incorporate genetic data in their models. Genetic data, whether at the SNP-level or gene-level, are quite different from neuroimaging data in their distributional form. While some studies have shown what may be gained from including this modality, past studies have not performed a simulation study of SNP data in a multi-modal setting. Past researchers have not explored the impact of various ICA algorithms on the data-driven models using genetic data alone or jointly with varying imaging modalities.

Many of the past decomposition analyses use jICA for data fusion, despite the fact that this approach may be restrictive. One reason for this is that jICA concatenates the data and estimates a joint unmixing matrix for all modalities. However, this makes the strong assumption that the data types modulate the same way across all subjects, having the same linear covariation for the modalities. By including multi-way CCA (mCCA) prior to the joint concatenation, it is assumed that this problem is mitigated. On the other hand, paraICA separately performs ICA on each modality, then performs a correlation analysis on the unmixing matrices for the data types. Some argue that this approach should retain the natural correlation while also allowing a multi-modal approach on the modal-specific component vectors. Another difference between these two approaches is that jICA requires the same number of components to be extracted from the data sources, while paraICA permits the number of ICs extracted per modality to be selected individually. Despite these differences, a comparison between these two frameworks have not been performed on the same data set.

Even beyond the selection of a multi-modal framework, most MMNG studies utilize traditional algorithms that are limited to linear or constrained optimization techniques. While new techniques have been developed, research has not been devoted to the exploration of algorithms in a multi-modal framework. In addition, with the introduction of neural networks or other deep learning approaches, more efficient algorithms may be written into ICA architecture. Each of the optimization approaches require varying levels of processing to be applied before ICA is run. In some cases the data must be whitened and PCA must be run before being fed into the multi-modal ICA model. Others require little processing and are reduced by sparsity penalization. How does the level of processing impact the outcomes of the MMNG application?

Furthermore, the impact that the optimization strategy has on the statistical reliability of the ICA estimates has not been explored for the MMNG structure. While ICA is desirable in the sense of interpretability and generalizability, the lack of global minima of the algorithms means that different ICA runs obtain different results even when the same data set is used. If the components may be considered as extracted features for the inputs of downstream analyses, what implication does this have on the statistical reliability of the estimates? Reliability may be measured in different ways,

but the focus in this dissertation is on the stability of the IC estimates with multiple runs.

Prior to the creation of imaging genetics studies, the two data sources were processed and modeled separately. These data sources still require separate procedures from data collection to quantifying the biological material and then to the pre-processing and transformation of the data. In addition, both data sets fall under the umbrella of "big data", which can not be analyzed with traditional methods. The large dimensions of both data sets further complicate the analysis in terms of data storage. This leads to computational issues due to the volume of the data at hand coupled with the advanced algorithms. Even after dimension reduction, the data still require storage, manipulation, cleaning or imputation prior to the analysis. Therefore, an imaging genetics pipeline will be presented that lays out the process of MMNG studies from the point of data collection to visualization. Within the pipeline, I discuss roadblocks that one may face in achieving replicability from an MMNG study. Researchers will necessarily gain and share access to open source information in the form of scientific knowledge, data collection and processing, including shared code for the computing programs used in carrying out the analyses. This idea of sharing and engaging in open access to scientific and computational resources is referred to as *open science* and is a vital contribution of this dissertation.

Last, but not least, understanding the neuropsychiatric application is vital to an MMNG analysis and should be the motivation of the study. Therefore, I will highlight the focus of this study in an application to AD using data derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Data will be pulled using different structural and functional neuroimaging modalities and combined with genetic information from patients in different stages of the disease. Decomposition analyses for MMNG studies permits group comparisons of disease categories by the features extracted, and interpretations will be provided in relation to the disease status of the subjects. An introduction and description of the ADNI data collection, followed by preliminary statistical analyses, will be presented in Chapter 2. The purpose of the preliminary analyses is to explore and confirm the relationship of biomarkers with clinical data. The clinical data includes diagnosis status, sex, age, education, race, and the results of cognitive exams. Biomarker data includes APOEϵ_4 expression, the top ten significant SNPs in Alzgene, tau accumulation measured by CSF, whole brain FDG-PET uptake, sMRI ROIs, and amyloid β (Aβ).

In Chapter 3.1, the methodology of MMNG will be presented. This begins with a description of processing that occurs prior to ICA, including correlation analyses, PCA and whitening. Next, the ICA framework is extended to the multi-modal scenario for joint and parallel analyses. Parallel tIVA with mCCA following the analyses is proposed for a three-way analysis of ADNI data, which, to date, has not been considered in the past. Four different algorithms are introduced, two of which are novel techniques used in this dissertation. Methods for comparison are introduced. A simulation setting is present in Chapter 3.2, describing how to simulate imaging genetics

data and how to use ICASSO to understand statistical reliability. The last section of Chapter 3 presents imaging genetics analyses as a step-by-step pipeline, from data collection and download to computation.

The results are presented in Chapter 4. I devote the first section to the simulation study of SNP, sMRI and rsfMRI data using a comparison of the four algorithms defined in Chapter 3. Section 4.2 uses ICA on sMRI, FDG-PET and rsfMRI data, providing an explanation of component analysis for both structural and functional neuroimaging data. The last section performs two-way and three-way multi-modal data fusion. All sections include a comparison of two traditional algorithms with two novel algorithms, and the MMNG analyses are performed on mCCA+jICA with a comparison to parallel+tIVA, a new approach to multi-way fusion of imaging genetics. Group comparisons are provided for three stages of an AD diagnosis: CN, MCI and AD.

Chapter 2

Alzheimer's Disease and Preliminary Results

The motivation of MMNG problems is to gain a greater understanding of the underlying biological basis and expression of neuropsychiatric illnesses. In this chapter, I will describe the focus of this research on a debilitating memory disorder, dementia. After describing the data collection process, I will perform an exploratory data analysis (EDA) on baseline data from over 2000 subjects of the possible disease statuses. The relationships and associations will be explored with supporting visualizations and interpretations.

2.1 Application to Alzheimer's Disease

Due to the steady stream of new data sources in imaging genetics, there is a wide spectrum of mental disorders for application: schizophrenia, bipolar disorder, post-traumatic stress disorder (PTSD), Parkinson's Disease, Alzheimer's Disease, and others. In this research, I will focus on Alzheimer's Disease (AD). The goal is to utilize multiple modalities, including structural and functional neuroimaging data with genetic data in order to elucidate the clinical expression or diagnosis. In this section, I will introduce the longest and most comprehensive study of AD, which is the Alzheimer's Disease Neuroimaging Initiatve (ADNI). Then I will discuss the data collection process through ADNI.

2.1.1 Alzheimer's Disease Neuroimaging Initiative

The most comprehensive study thus far of the crippling memory disorder, AD, is called the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2008). Dr. Michael Weiner, Professor of Radiology at the School of

Medicine at the University of California in San Francisco gathered a team of scientists to propose a large-scale initiative of research for AD. The aim of the Alzheimer's Disease Neuroimaging Inititiative (ADNI) is to understand and compare neurological and biomarkers changes that present themselves in AD pathology. What followed this scientific coalition was approval for a total of $218 million of combined funding provided by public and private sectors, including the National Institute of Aging, pharmaceutical companies, and the National Institute of Health, to collect neuroimaging, genetic and other biomarker data believed to be important in the classification of AD. Beginning in 2004, data collection is ongoing and will be completed in 2021. To date, ADNI is the only multicenter study, collecting data from over 60 different sites, with the largest and longest longitudinal study, keeping track of over 800 patients for 17 years, for the study of AD (Mueller et al., 2008; M. W. Weiner et al., 2015; M. W. Weiner, Veitch, Aisen, Beckett, Nigel, et al., 2017; M. Weiner & ADNI, 2013).

The biomarkers involved in the onset of AD follow a series of stages as the disease progresses from one clinical status to the next - normal controls (NC) to mild cognitive impairment (MCI) to an AD diagnosis. A longitudinal study of patients in each of the clinical stages with a collection of various biomarkers will increase knowledge of the biological path AD takes (Petersen & Jack, 2009). The five biomarkers under consideration are: 1) b-amyloid ($A\beta$) measured in CSF or by amyloid PET imaging (Shaw et al., 2009); 2) tau protein measured by synaptic dysfunction in FDG-PET (Butler et al., 2019); 3) brain atrophy measured by sMRI (Jack et al., 2008); 4) memory loss as measured by cognitive tests (Petersen et al., 2014); 5) clinical function, also measured by cognitive tests, and a rating of the severity of life interruptions caused by memory malfunctioning. The first three biomarkers are observable before the memory concerns get out of hand. Unfortunately, by the time the patient is experiencing elevated levels of the last two biomarkers, there is nothing that can be done to stop the steep decline in memory since treatment has yet to be discovered (M. Weiner & ADNI, 2013).

While most of the early studies focused on the joint analysis of genetic markers and structural brain abnormalities, functional changes may allow even earlier detection. Recent research has revealed that tau protein levels are more closely linked to cognitive decline than $A\beta$. The basis of this fact is that AD is characterized by disruptions in the brain connectome at the synaptic level. When functioning diminishes, then the anatomical atrophy begins to occur (M. W. Weiner, Veitch, Aisen, Beckett, Nigel, et al., 2017). Neurodegeneration of the tau proteins are measured by first administering a small amount of the radioactive tracer fluorodeoxyglucose (FDG). PET imaging reveals how well the sub-

stance flows within the brain and displays whether an adequate number of connections are being made in essential ROIs. Deficiencies in these connections are a direct reflection on connections made from diseased tissue (Kunkle et al., 2019). Thus, a paired analysis of FDG-PET imaging and sMRI will potentially identify an earlier stage of cognitive impairment than has previously been identified. Instead of analyzing the pathology of AD as a $(\text{big data})^2$ problem with one imaging approach and genetic data, a three prong approach joining structural brain atrophies, functional brain disconnections and genetic risk will be considered; that is, big-data-cubed $(\text{big data})^3$.

2.1.2 Data Collection

Inference involving multiple modalities of big data requires large sample sizes and consistent techniques for the data collection process. The ADNI researchers processed and uploaded the data online from four protocols, ADNI1, ADNI-GO, ADNI2 and ADNI3. The repository is constantly updated as new information is received and processed from the 63 collection sites. The online ADNI database can be found at https://ida.loni.usc.edu/login.jsp. This massive database includes: clinical data, such as medical history, results of neuropsychological exams, education, marital status, and other subject characteristics; biospeciman results and procedures, like blood and CSF; imaging data, both the actual images of the brain slices and multiple levels of pre-processed quantitative data of multiple neuroimaging forms, sMRI, fMRI, various PET imaging, DTI, etc.; genetic data, for a whole genome approach, individual SNP data, DNA methylation profiling and protein genotyping.

Other documents describing the procedures for data collection, how the data is pre-processed, variable descriptions, the study methods for collection of the biomarkers, the goals of the ADNI protocol, news on recent publications and findings, suggested methods and tools, and even a place to sign up as a participant may be found on the ADNI website (http://adni.loni.usc.edu/. Any further questions a researcher may have can be posted to the ADNI Data forum on Google (https://groups.google.com/forum/#!forum/adni-data). Questions for any of the ADNI Cores (Neuropathology, Genetics, MRI, PET, Clinical, Biostatistics, Informatics, etc) may also be posted at this site: http://adni.loni.usc.edu/support/experts-knowledge-base/ask-experts/.

After ADNI1 was implemented, geneticists and neuroscientists began their analyses by pre-processing the data. This is a key step in the analysis because any alterations across collection sites in the formation of data sets may result in observed differences at the analysis level that are in fact meaningless. The genetic data are processed at baseline for each subject using consistent proce-

dures (Saykin et al., 2015a). However, standardizing this process for the noisy neuroimaging data is not so easy. ADNI researchers have recognized these challenges and thereby formalized the pre-processing technique, then publicized collections of this data along with instructions on standard procedures for researchers to adhere to (Jack et al., 2008; M. Weiner & ADNI, 2013; Wyman et al., 2013). The sMRI and FDG-PET imaging used in this study, and any future analysis will be taken from the standardized collection of imaging data to ensure reliability and allow for replication.

Genetic material is obtained through blood samples drawn at baseline by the ADNI Biomarker core which are then sent to and processed by the National Institute on Aging-sponsored National Cell Repository for AD (NCRAD). DNA is extracted from buffy coats using whole blood collected in EDTA tubes, and the Qiagen PAXgene Blood RNA Kit purifies and extracts total RNA from a 2.5 mL tube. The genome-wide genotyping is processed by the Illumina Human 610-Quad BeadChip by TGen (for ADNI1) with bead intensity data used to call genotypes by BeadStudio 3.2 (et al Saykin, 2010). APOE genotyping is carried out separately by polymerase chain reaction (PCR) amplification, Hhal restriction enzyme digestion and subsequent standard gel resolution and visualization processes. Assay kits were then added to DNA samples. Genotypes were called using LGC Genomics' Kraken software and then returned to the Genetics core for QC, the same as the genome-wide data, including sex and identity checks (Saykin et al., 2015a).

The MRI Core is responsible for overseeing the standard processing procedure to maintain coherence across sites and over time. Image quality characteristics such as contrast-to-noise, spatial resolution, resistance to artifact, reliability, speed, etc. are managed and aligned from time point to time point (Jack et al., 2008). Even differences in scanners due to technical malfunctions or upgrading to newer versions were also considered, and the data were organized accordingly. QC on the 3D 1.5 & 3 T_1-weighted sequences is conducted by the Aging and Dementia Imaging Research Laboratory at the Mayo Clinic. Although data is available at each stage of this process, it is recommended that researchers utilize those with the maximal level of correction, the MPRAGE files processed by Freesurfer. Gradient nonlinearity (geometric distortion) is corrected for by 3D gradwarp, then is followed by intensity inhomogeneity correction from nonuniform receiver coils by employing B1 calibration scans. Next, N3 reduces the residual intensity due to a wave or dieletric effect. Spatial scaling is improved by phantom-based distortion correction (Wyman et al., 2013).

The ADNIMERGE R package is recommended for researchers interested in studying the joint association of genetic and neuroimaging data because it is

easy to access, well organized and has already been processed according to the standard procedures. The package consists of a total of 234 data sets containing detailed data collection information, from a diagnosis and symptoms checklist to blood drawing procedures for each subject, from lab data to neuropathology status. Dr. Mike Donahue from UC San Diego collected the ADNIMERGE information initially for R and provided an ADNIMERGE R package that can be uploaded in conjuction with the Hmisc package. For information on the former package see https://adni.bitbucket.io/, which provides variable descriptions and some basic methods for studying the data. Data sets may be uploaded and merged easily by RID.

2.2 Exploratory Data Analysis

The extensive longitudinal data collection of ADNI and the comprehensive preliminary research in response to this data has set the stage for future researchers to make groundbreaking contributions to AD pathology. Past research has focused on understanding how each biomarker changes throughout the disease trajectory, assessing the risk of developing AD based on genetic make-up, and identifying structural and functional changes that occur at the brain level. I discussed a multitude of methods, including brain-wide genome-wide analyses, dimension reduction techniques, and various data fusion strategies. After perusing the background research, it is clear that there is still a lot of work that remains to be done. Many methods have been proposed for combining genetic and neuroimaging data, but most do not incorporate a multi-source three-prong approach. The pathology of AD and the diligent work of past researchers highlights the need for statisticians to develop statistical methodology for the multi-modal data fusion of sMRI, PET imaging and genetic markers while controlling for diagnosis.

At the time of this preliminary analysis, the last date of collection for ADNIMERGE was September 16, 2019. Each subject has multiple lines of observations, one for each time point. Note that $N = 2,217$ baseline values are collected across all four protocols. The 116 variables contain site and patient IDs, diagnoses at each time point, subject characteristics, multiple memory exam results, biomarker outcomes, ROI brain volumes and APOEϵ_4 allele frequency. Table 2.1 lists the variables that will be considered. Natural patient heterogeneity exists from basic subject characteristics, such as age, sex, education level or race. The average age at baseline is 73.2 with a range of (54.4, 91.4). Slightly more men than women are represented in the study, making up 53.2% of the total sample. 91.8% of the sample is of a Caucasian race. The vast majority is

well-educated, since 73% have at least a four-year college degree. In the data set, education is given by the number of years. The assumption is that those who completed high school have been to school for 12 years; a four-year college is 16; a masters degree is 18; and an advanced degree is longer.

Table 2.1: Variable names and descriptions pulled from ADNIMERGE.

Variable	Description	Possible values
RID	participant roster ID	2-6799 (2217 distinctive)
EXAMDATE	date of data collection	yyyy-mm-dd
VISCODE	the timepoint of data collection	bl=baseline, m03, every 6m
ORIGPROT	original ADNI protocol at baseline	ADNI1, Go, 2, 3
COLPROT	ADNI protocol at time of collection	ADNI1, Go, 2, 3
AGE	age at the time of screening	54.4-91.4
DX.bl	diagnosis at baseline	CN, SMC, EMCI, LMCI, AD
DX	diagnosis at each collection time point	CN, MCI, Dementia
PTGENDER	gender	Female, Male
PTEDUCAT	education level by years in school	16= bachelors, >18 is advanced degree
RACE	white, Am Ind/Alas, Hawaii, Black, ...	categorical based on racial orientation
CDRSB	clinical dementia rating, diagnostic	0-18, >1 non-CN
RAVLT.immediate	Rey Auditory Verbal Learning Test recognition	0 to 75, $<$ 30 memory impaired
RAVLT.perc.forget	RAVLT percentage of forgetting	0-100%, >50% concerning
MMSE	Mini-Mental State Examination, diagnostic	0-30, $\leq$26 MCI/AD
ADAS13	Alzheimer's Disease Assessment Scale, version 13	0 to 85, >20 for AD
*WLMS	Wechsler Logical Memory Scale, education-adj	0 to 25, $\leq$8 MCI/AD
APOE4	APOEϵ_4 allele frequency	0=homo maj, 1=hetero, 2=homo min
ABETA	Aβ, or b-amyloid measured by CSF	0 to 1700, lower for MCI/AD
TAU	tau protein measured by CSF	8 to 103, higher for MCI/AD
FDG	synapse by fluorodeoxyglucose-PET	.694 to 1.7, lower for MCI/AD
Hippocampus	hippocampal grey matter volume	2990-10800mm^3
MidTemp	medial temporal grey matter volume	9380-32300mm^3
WholeBrain	whole brain grey matter volume	669k-1490kmm^3

2.2.1 Cognitive Tests in ADNI

Clinical data pertaining to the diagnosis and neuropsychological exam results will also be extracted from ADNIMERGE and considered in the preliminary analysis. The cognitive exams included in the data test a range of abilities, from episodic memory to actual recall. CDR-SB is a sum of box method that rates the severity of dementia on a scale from 0 to 18, where 0 shows no sign of dementia and 18 shows extreme severity (B. Li et al., 2017). RAVLT provides various measures of episodic memory such as retention or how forgetful one may be. Past research has shown that overall medial temporal grey matter volume is positively associated with retention, but forgetfulness is negatively associated with hippocampal volume (Kueper et al., 2018). Both measures will be included,

RAVLT-recall and RAVLT-forgetting, respectively. A 30-question test assessing attention, memory and recall is MMSE, and those who are CN will score higher on a scale from 0 to 30 than those with MCI or AD. The pharmacometric ADASCog-13 (ADAS-13), the "gold standard" for past AD research, assesses written and verbal responses of subjects that are related to fundamental cognitive functions and is used in the qualification of target anti-dementia drugs (Kueper et al., 2018; B. Li et al., 2017). WLMS provides a comprehensive mental rating of recall, memory and recognition (M. Weiner & ADNI, 2013).

Each of the ADNI protocols has a Procedures Manual discussing all details involving patient screening and technical procedures (https://adni.loni.usc.edu/wp-content/uploads/2010/09/ADNI_GeneralProceduresManual.pdf). The diagnostic categories at baseline are CN, SMC (significant memory concern), EMCI, LMCI, and AD. SMC was included beginning with ADNI 2 to cover those patients with CN memory test results but who have self-reported a notable concern with their memory. Three of the neuropsychological tests used as diagnosis criteria are WLMS, MMSE and CDR. The WLMS gives a score based on whether a patient operates at the average recall ability, given the level of schooling they obtained in their lifetime. For subjects who attended a four-year college or greater: a score of 9 or more were classified as CN/SMC, a score of 8 or less were classified as MCI/AD. An MMSE score of 24-30 was expected for CN/SMC/MCI or between 20-26 for AD; a score of 0 for CDR-SB for CN/SMC, at least 0.5 for MCI and at least 1 for AD. Adjustments for EMCI, LMCI and AD were made based on a combination of the scores and patient/physician reporting.

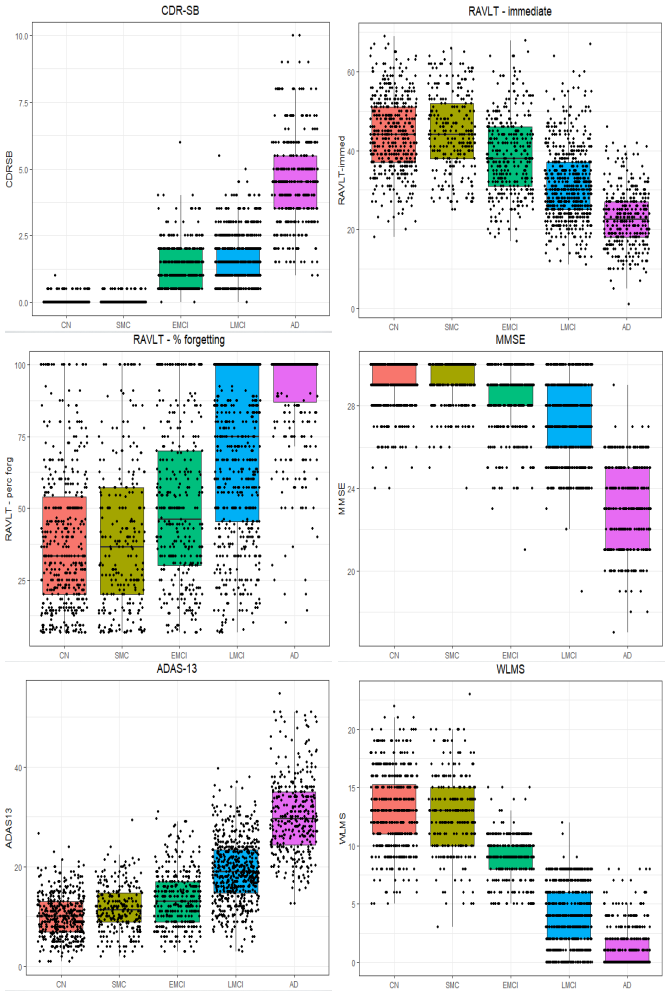
Table 2.2: Mean/standard deviation of diagnosis categories at baseline.

Quantitative	CN	SMC	EMCI	LMCI	AD
N=2188	513	289	372	636	378
Age	73.8 (5.56)	71.2 (6.45)	71.3 (7.29)	73.8 (7.43)	74.9 (7.73)
Education	16.4 (2.68)	16.8 (2.44)	16.1 (2.62)	15.9 (2.93)	15.2 (2.95)
CDR-SB	.03 (.116)	.05 (.157)	1.29 (.737)	1.64 (.89)	4.45 (1.61)
RAVLT rec	45.3 (9.65)	45.9 (9.99)	39.6 (10.4)	31.5 (9.81)	22.7 (7.2)
RAVLT for	33.8 (2.81)	35.6 (3.1)	46 (2.64)	66.8 (2.36)	89.8 (1.79)
MMSE	29.1 (1.1)	29.1 (1.15)	28.3 (1.56)	27.2 (1.78)	23.1 (2.01)
ADAS-13	9.8 (4.25)	11.3 (4.65)	13.4 (5.28)	18.9 (6.53)	30.4 (7.74)
WLMS	13.3 (3.31)	13 (3.21)	9 (1.75)	3.9 (2.71)	1.4 (1.88)
Aβ	1015 (382)	1083 (374)	943 (355)	767 (328)	632 (258)
CSF tau	238 (86.6)	239 (89.9)	256 (119)	309 (124)	368 (141)
FDG-PET	1.3 (.11)	1.32 (.11)	1.28 (.12)	1.21 (.13)	1.07 (.14)
Hippocam	7355 (911)	7536 (900)	7272/ (1015)	6475 (1135)	5773 (1007)
Med temp	20.3k (2.69k)	20.9k (2.52k)	20.8k (2.61k)	19k (2.82k)	17.2k (3.11k)
Whole br	1027k (101k)	1060k (93k)	1067k (106k)	1011k (108k)	977k (115k)

Table 2.2 provides the mean and standard deviation by diagnosis category for age, education, the cognitive exams and biomarkers. Age appears to be higher for more advanced stages of AD, while the standard deviation of the age increases for each successive category. This shows that there is more variability in the age group of later stages of AD. he opposite is true for, the number of years of education appears to be higher for healthier individuals and lower for LMCI and AD. The variability of education also increases for later stages. Recall that low values of RAVLT-recall, MMSE, and WLMS are indicative of advanced stages of AD, while this is true for high values of CDR-SB, RAVLT-forgetting, and ADAS-13. These expectations are reflected for the mean values of each diagnosis status. Consequently, the standard deviations are higher for LMCI and AD categories than the lower stages for CDR-SB, MMSE and ADAS-13. However, we see more variability in healthier individuals for the RAVLT tests. This is interesting, because RAVLT-forgetting expects higher values for AD patients but still has a lower variability than the lower disease statuses. The biomarkers will be discussed in Section 2.2.2.

Since CDR-SB, MMSE AND WLMS are directly used as diagnostic criteria, three additional cognitive tests (both RAVLT tests and ADAS-13) are added as consideration to the diagnosis. Past research supports the use of these exams as regular criteria for the diagnosis of AD (Davis et al., 2013; Kueper et al., 2018; B. Li et al., 2017). Figure 2.1 gives boxplots of these neuropsychological tests for the different diagnoses at baseline for all subjects from ADNI1, 2, GO and 3 combined. There is little to no difference between CN and SMC from a testing standpoint. Differences in clinical outcomes do not begin to emerge until comparing SMC to EMCI (SMC-EMCI). The largest observable difference of SMC-EMCI is with WLMS, where the variation is much larger for SMC (IQR=5) and the median is much smaller for EMCI (9). WLMS drops another five points for EMCI-LMCI and then another 3 for LMCI-AD. The disease differences among EMCI-LMCI-AD show stage-by-stage decreases for RAVLT-immediate and similar increases for ADAS-13. The percent of forgetting measured by RAVLT is the neurological exam with the most variable responses for all categories but AD, where the median is 22% (of forgetting) higher than LMCI. Considerable increases for CDR-SB and decreases for MMSE are observed in LMCI-AD. Table 2.2 also provides the mean and standard deviation for each of the clinical stages. Percent of forgetting shows a difference in averages of 56% for CN-AD. As expected based on the diagnostic criteria, the averages are consistently different for CN/SMC-LMCI/AD categories. In Table 2.3, the proportions for the categorical variables are given. There is a greater representation of males, at 53%, with 92% white participants. Over half of the

Figure 2.1: Baseline results for cognitive memory tests used as diagnostic tools in the ADNI study: CDR-SB, RAVLT (immediate recognition & percent forgetting), MMSE, ADAS-13 and WLMS. Five disease categories are compared in each plot: CN, SMC, EMCI, LMCI, AD (left to right).



subjects have the homozygote major expression of $APOE\epsilon_4$. The biomarkers will be discussed in Section 2.2.2.

Table 2.3 provides summary statistics on the same variables discussed in Table 2.2, now for all of the disease categories combined. The overall means and standard deviations may be compared to the means and standard deviations for the individual stages of the disease. The average age of all subjects is 73.2, and the mean is higher than average for LMCI and AD categories. Similarly, years of education are lower than average for LMCI and AD. Scores on the cognitive

Table 2.3: Summaries of clinical and biomarker variables at baseline.

Quantitative	min	Q1	Q2	Q3	max	mean	sd
Age	54.4	68.3	73.2	78.3	91.4	73.2	7.38
Education	4	14	16	18	20	15.8	2.79
CDR-SB	0	0	1	2	10	1.48	1.54
RAVLT rec	0	27	36	45.5	71	36.5	12.8
RAVLT for	0	28.6	57.1	100	100	57.4	33.4
MMSE	17	26	28	29	30	27.4	2.65
ADAS-13	0	10	15	22	54.7	16.8	9.28
WLMS	0	3	8	12	23	7.74	5.45
Aβ	203	562	752	1090	1700	846	365
CSF tau	81.5	194	258	350	852	287	129
FDG-PET	.694	1.14	1.24	1.33	1.7	1.23	.153
Hippocam	2990	5950	6860	7650	10800	6790	1190
Med temp	9.4k	17.5k	19.4k	21.5k	32.2k	19.5k	31k
Whole br	669k	941k	1020k	1100k	1490k	1020k	112k

Categorical	Categories	Proportions
Gender	female/male	.47/.53
Race	White/Am Ind-Alas/Haw-PI	.92/.002/.02
	Black/Mixed/Unkown	.05/.01/.001
APOEϵ_4	homo maj/hetero/homo min	.54/.36/.10

measures for CN, SMC and, in some cases, EMCI from Table 2.2 tends to be close to the mean and median values for these measurements in Table 2.3.

2.2.2 Biomarker Differences by Disease Category

The field of imaging genetics is in place to induce more powerful measurements than memory scores as diagnostic criteria for each clinical disease stage. Corresponding to the biomarkers introduced in the section on AD pathology, the biomarkers included in the preliminary analysis are: APOEϵ_4 values 0, 1, 2 signifying the allele frequencies for homozygote minor, heterozygote, and homozygote major respectively; Aβ measured by CSF; CSF tau protein levels; the average FDG-PET of angular, temporal, and posterior cingulate; hippocampal, medial temporal and whole brain grey matter volumes. Descriptions and statistical summaries of these variables are laid out in Tables 2.1-2.3. Table 2.2 provides the means and standard deviations of the biomarkers for each diagnosis status, while Table (2.3) provides summary statistics of the biomarkers for the subjects overall.

Disease-specific means decrease progressively for each successive stage of AD diagnosis for Aβ, FDG-PET uptake and all instances of brain volume. The

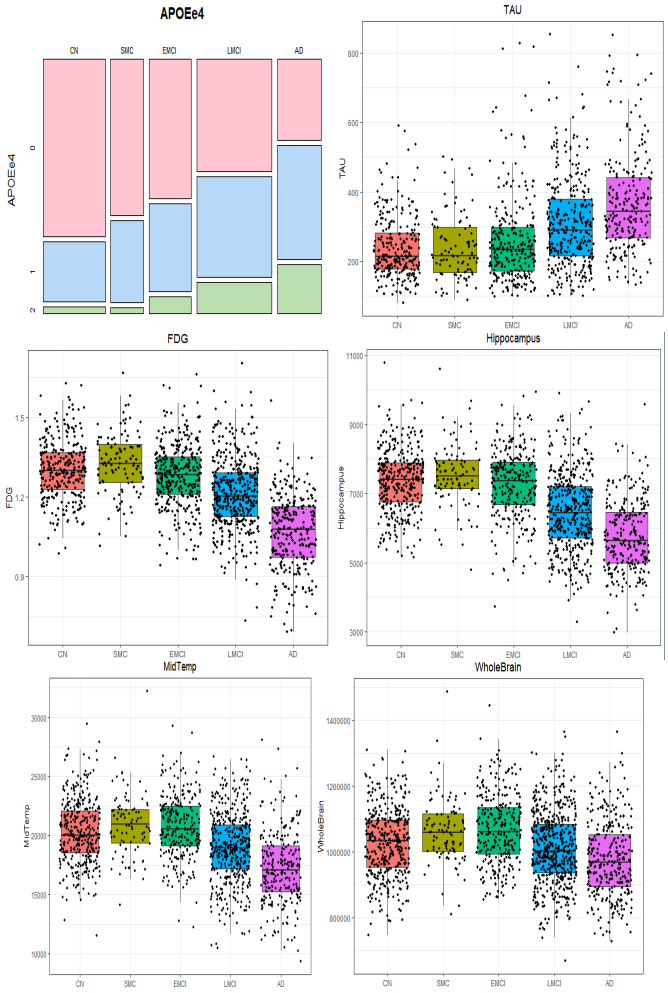
means of the hippocampus, medial temporal, and whole brain volumes decrease with each advanced stage of AD. On the other hand, tau accumulation measured by CSF increases for each advanced stage. Recall that the diagnoses were primarily based on the cognitive measures presented in section 2.2.2. The consistent increase/decrease of the biomarkers reveals that these biological measures can supply specific disease-level information as well. With the exception of $A\beta$ and CSF tau, the standard deviations increase for advanced stages of AD even though the means decrease for all. This means there is more variability among the later stages of the disease for the biomarkers when compared to the cognitive measures. In studying the entire patient sample in Table 2.3, one can see that the averages of the biomarkers tends to fall between the EMCI and LMCI measures in Table 2.2. This highlights the fact that there is a steady decline in the volume throughout the stages. In other words, all of the biological measures show substantial differences for all 5 categories. Given the increases in disease-specific variation for the biomarkers in (Table 2.2), it is not surprising to see that the overall standard deviations are skewed right, which reveals the impact of the later stages of the disease on subject variability in the atrophy of the brain.

Figure 2.2 shows boxplots of the CSF and brain-related biomarkers, and a mosaic plot is provided to showcase the APOEϵ_4 expression across disease status. Thisreveals that those with the heterozygotic or homozygotic minor alleles make up a greater proportion of the advanced clinical stages LMCI/AD than those with a 0. An increase in heterozygotic representation is observed even for CN-SMC. Only one-third of those in the AD category have the major homozygotic gene expression. CSF measurements of Aβ/tau decreases/increases, for each clinical category from CN/SMC/EMCI-LMCI-AD. This coincides with the understanding that a drop in soluble levels of Aβ occurs at the onset of AD and an increase in tau protein reveals neuronal damage (M. Weiner & ADNI, 2013). Anatomically, the hippocampus displays the most noticeable decreases in brain volume, followed closely by the medial temporal lobe. Grey matter volume as a whole also decreases, but not as obviously. The averages of these biomarkers in

2.2.3 Relationships with the Diagnoses

The overall goal of the EDA is to test the combined association of genetic and imaging data with the diagnosis. For this reason, I tested each variable individually with the diagnosis status. After validating certain conditions, I conducted tests for association for each categorical variable with the baseline diagnosis using the Chi-Square Test for Independence. For the continuous variables, an ANOVA or Kruskal-Wallis test was run to assess whether the mean responses differed for each clinical group. Checking the assumptions was required to de-

Figure 2.2: Baseline results for biomarkers APOE, CSF tau, FDG-PET, Hippocampus volume, Medial Temporal volume, and whole brain volume from ADNI. Five disease categories are compared in each plot: CN, SMC, EMCI, LMCI, AD (left to right).



termine which test would be used. If the more stringent ANOVA assumptions did not qualify, then the non-parametric version, the Kruskal-Wallis, would test the hypothesis of whether both variables come from unrelated populations, meaning the samples do not affect each other.

The collection of ADNI data took place with volunteers, so the simple random sample condition should be relaxed. As a result, one must keep in mind that this study may include volunteer bias. The categorical variables included are gender, race and APOEϵ_4. Each of these variables has two or more categories,

including diagnosis which has five. Finally, even though there is a large sample size, the cell counts requirement is not met for some of the racial minorities, so some of the categories were combined to meet this assumption (see Table 2.3 again). χ_1 refers to the test statistic of this Chi-square test. The association results between overall diagnosis and gender, race and APOEϵ_4 may be viewed in the first column of Table 2.4; the test statistics, χ_1, degrees of freedom and associated p-values are presented. The right-hand side of Table 2.4 gives the standardized residuals of these tests.

The Chi-square Test for Independence tests the null hypothesis of whether two categorical variables are independent, or have no relationship. Although gender, race and APOE are all significant, with small p-values, the degrees of freedom (df) vary for each. The impact of race category is weak when compared to the result given by gender and APOE expression. Relative to the degrees of freedom, APOEϵ_4 has the largest test statistic with a relatively small df when compared to Gender. This means there is a moderate but significant impact of gender on the level of diagnosis status. However, the genetic influence is even greater. That is, diagnosis status, determined by cognitive measures, as an association with gender, race and APOEϵ_4.

Table 2.4: Standardized residuals for baseline diagnosis vs. categorical variables.

Variable	level	CN	SMC	EMCI	LMCI	AD
Gender, $df = 4$	female	1.854	3.462	-.723	-2.691	-1.050
$\chi_1 = 46, p < .001$	male	-1.733	-3.237	.676	2.516	.982
Race, $df = 12$	Asian	-.34	1.39	-1.22	-.14	.58
$\chi_1 = 38, p = .035$	Black	1.9	.33	-1.84	-.04	-.67
	Other	-1.59	2.77	2.68	-1.75	-.95
	White	-.16	-.65	.22	.27	.19
APOEϵ_4, $df = 8$	0	5.425	2.158	0.736	-2.743	-5.569
$\chi_1 = 196, p < .001$	1	-4.171	-.805	-.164	1.888	3.422
	2	-4.792	-3.576	-1.435	2.855	6.600

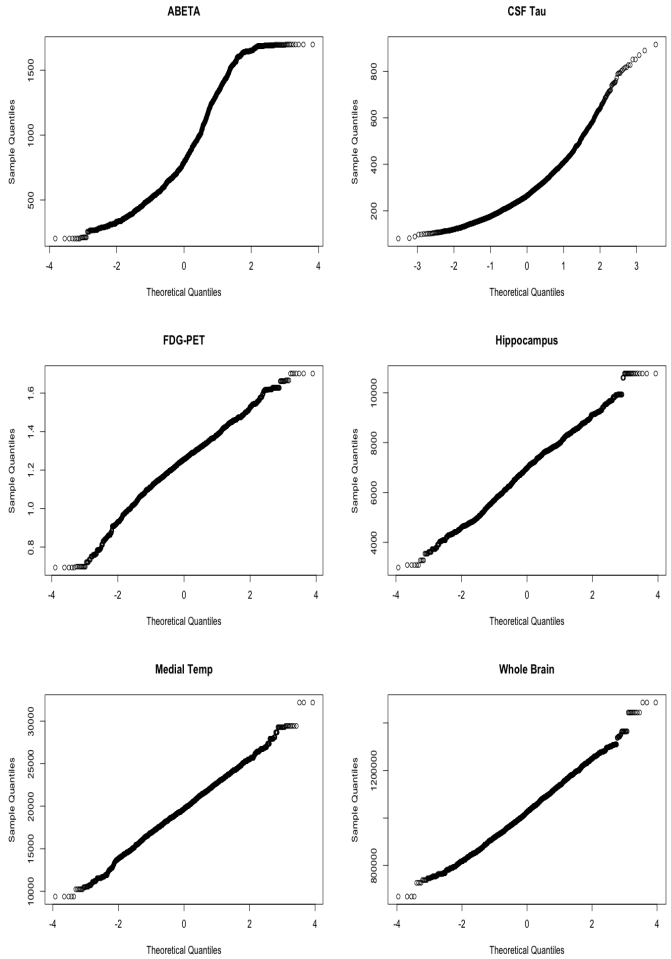
Table 2.4 also provides the standardized residuals for gender, race and APOEϵ_4. The magnitude of the residuals reveal the contribution of that category to the outcome, and the sign (+/-) gives the direction of the influence. The test statistics of gender and race were moderate, and for these standardized residuals, two or three categories carry the largest positive or negative weights, more than 2 or less than -2. There is a gender difference observed with a large negative outcome of -2.691 for females and large positive outcome of 2.516 for males in the LCMI category. This confirms past research which suggests that men are more likely to experience AD symptoms with old age. For the race analysis, the "Other" category is a combination of American Indian-Alaskan, Hawaiian-PI islanders, mixed and unknown. The SMC and EMCI categories have large positive residu-

als of 2.77 and 2.68, respectively, suggesting that race-specific differences may be present at earlier stages of AD progression. Lastly, the presence/absence of two minor alleles of APOEϵ_4 results showed prominent confirmation of AD/CN statuses. Those with two major alleles show a residual of 5.425 in the CN category and those with two minor alleles show a residual of 6.6 in the AD category, which is consist with the mosaic plot observed earlier that showed the greater proportion of the CN category has 0 while the greater proportion of the AD category has 2. The flip side of this is that significant negative residuals are observed in those with 1 and 2 minor alleles for the CN category and for those with 0 minor alleles in the AD category. Notice that even those with 1 minor allele show a positive residual of 3.422 in the AD status.

Next, an association analysis was conducted for each of the continuous variables against diagnosis status. It doesn't make sense to calculate an association, per se, between a categorical response and a continuous predictor. However, the average values for each subgroup, or clinical stage, may be compared. The normality assumption was checked by looking at qqplots of each continuous variable, and departures from normality were observed especially in the tails of the distributions. See Figure 2.3 to view the qqplots. The variances were compared in each disease stage, and several biomarkers showed greater variability in specific categories (see Table 2.2 again). Thus, none of the continuous variables met the assumptions for an ANOVA test. In place, the Kruskal-Wallis rank sum test is run, denoted by the corresponding test statistic χ_2. Relaxing the assumption of normality, the Kruskal-Wallis tests the null hypothesis of whether the variables come from unrelated populations.

Column one of Table 2.5 displays the Kruskal-Wallis results testing the relationship of diagnosis with the continuous variables. The degrees of freedom are 4 for all of these tests, and each of the tests show a p-value< 0.001. This means the overall diagnosis, DX, has a significant association with age, the clinical exams, and all of the biomarkers considered. Due to the same degrees of freedom for each, we may more easily compare the results. The largest test statistics are observed for CDR-SB, MMSE, ADAS-13 and WLMS. These results are expected based on the diagnosis definition which is based on these cognitive measures. We observe identical results for age as we do for the biological measure of Aβ, revealing the lowest relationship in the table, although is still moderately strong. The largest association among the biomarkers can be shown with the FDG-PET and hippocampal volume, with 444 and 414, respectively. It is interesting to note that the value of the test statistic of medial temporal lobe is much lower than the hippocampus, even though the hippocampus is located iwthin this lobe. This confims previous results that show brain atrophy

Figure 2.3: Baseline results for biomarkers APOE, CSF tau, FDG-PET, Hippocampus volume, Medial Temporal volume, and whole brain volume from ADNI. Five disease categories are compared in each plot: CN, SMC, EMCI, LMCI, AD (left to right).



in the hippocampus more than any other part of the medial temporal lobe. The remaining values displayed in this table will be discussed in Section 2.2.4.

2.2.4 Collinearity among the biomarkers

A preliminary analysis is also desirable to discover the relationship between APOEϵ_4 and the continuous biomarkers. Therefore, I examined the relationship of APOEϵ_4 expression with age, cognitive scores, and biomarkers are presented in the second column of Table 2.5. After assumptions were verified, the

Table 2.5: Tests of association among DX, mental exams and biomarkers.

Variable	DX.bl	APOEϵ_4	Aβ	TAU	FDG	Hippo	MedT	WhBr
Age	$*\chi_2 = 81$	$*F = 18$	-.08	.10	-.12	-.37	-.21	-.29
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2, 2047$						
CDR-SB	$*\chi_2 = 1847$	$*\chi_2 = 169$	-.36	.34	-.56	-.47	-.38	-.
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2, 2047$						
RAVLT rec	$*\chi_2 = 901$	$*F = 56$	.42	-.32	.5	.44	.31	.15
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2, 2044$						
RAVLT for	$*\chi_2 = 99$	$*\chi_2 = 133$	-.32	.33	-.36	-.46	-.25	
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2$						
MMSE	$*\chi_2 = 1134$	$*\chi_2 = 129$	.38	-.35	.54	.48	.38	.22
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2$						
ADAS-13	$*\chi_2 = 1182$	$*\chi_2 = 166$	-.44	.38	-.62	-.53	-.42	-.
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2$						
WLMS	$*\chi_2 = 1471$	$*\chi_2 = 182$	.42	-.36	.49	.51	.37	.22
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2$						
Aβ	$*\chi_2 = 81$	$*\chi_2 = 94$	1					
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2$						
CSF tau	$*\chi_2 = 167$	$*\chi_2 = 121$	-.27	1				
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2$						
FDG-PET	$*\chi_2 = 444$	$*F = 46$	.38	-.29	1			
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2$						
Hippocam	$*\chi_2 = 414$	$*F = 43$	.32	-.32	.43	1		
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2, 1479$						
Med temp	$*\chi_2 = 234$	$*F = 11.9$	.22	-.29	.41	.60	1	
	$p < .001$	$p < .001$						
	$df = 4$	$df = 2, 1454$						
Whole br	$*\chi_2 = 119$	$F = 1.96$	.09	-.20	.21	.58	.75	1
	$p < .001$	$p = .14$						
	$df = 4$	$df = 2, 1677$						

ANOVA test was applied to age, RAVLT percent forgetting, FDG-PET, and grey matter volume of the hippocampus, medial temporal lobe and the whole brain. The F tests were run where the assumptions held, otherwise the Kruskal-Wallis test was used, with test statistics denoted by F and χ_2, respectively. The

degrees of freedom vary for the F test, but they are 2 for the Kruskal-Wallis test. Almost all results showed p-values less than 0.05. However, the test of association between whole brain and APOE were not significant. RAVLT-recall has the highest test statistic for the F test, with FDG-PET having the second highest. This is interesting to show that the cognitive measures are still strongly associated with the genetic expression of APOE. Similar to diagnosis results, FDG-PET has the strongest relationship with APOE with hippocampal volume next. These associations showcase the fact that genetic, neuroimaging and clinical material are not only representative of AD pathology, but also interact with the cognitive decline experienced.

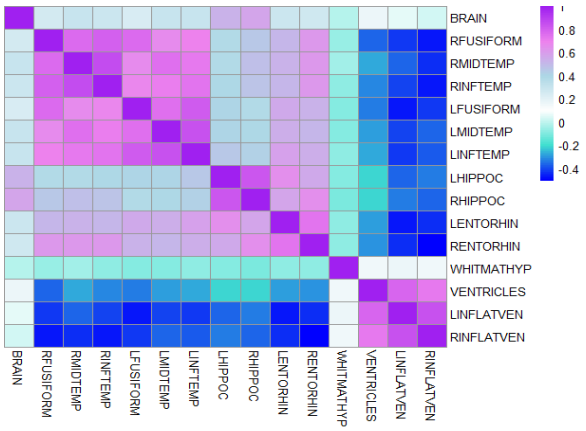
Working with MMNG data presents many analysis options. Often, the disease status or the imaging phenotypes are thought of as the response. However, the results discussed raise the question of how APOEϵ_4 is involved in AD progression and whether it should be incorporated into the diagnosis. One can recode this variable so that it is an indicator variable equal to 1 for subjects who have two minor alleles and a 0 for those with at at most one minor allele, and a multiple logistic regression may be computed. Considering the mental exam data, diagnosis as a factor, three ROI volumes and FDG, one can view the odds ratios, odds ratio 95% CIs, estimates and p-values in Table 2.6. The highest odds ratios obtained are given by the MCI and Dementia (AD) factors at 2.27 and 2.45 respectively. This means that the MCI group are 2.27 times more likely to have 2 minor alleles and the AD group are 2.45 times more likely to have 2 minor alleles. Of the brain volumes, we see that the hippocampus has the highest negative value for the test statistic with a correspondingly low p-value. Holidng the other variables constant, this means that those with a reduction in hippocampal volume have a greater chance of having the two APOEϵ_4 alleles. Of the cognitive exams, only WLMS showed a strong association with this outcome.

Table 2.6: Logistic regression results on two minor alleles of APOEϵ_4.

Variable	Hippoc	MedTemp	WhlBrn	Age	Male	FDG-PET
Odds Ratio	1	1	1	0.92	1.20	0.21
95% CI	(.999,1)	(1,1)	(1,1)	(.89,.94)	(.79,1.84)	(.05,.81)
z stat	-3.84	2.69	-0.62	-6.24	0.86	-2.25
p-val	0.003	.007	.534	0.000	0.388	0.025
Variable	WLMS	MMSE	CDRSB	MCI	Dementia	
Odds Ratio	0.93	1.03	1.03	2.27	2.45	
95% CI	(.88,.98)	(.95,1.11)	(.91,1.16)	(1.13,4.85)	(.941,6.57)	
z stat	-2.56	0.68	0.44	2.22	1.81	
p-val	0.011	0.499	0.661	0.026	0.070	

Neuroimaging data are expected to have high multicollinearity among the various regions, because the functional and structural regions in the brain are interconnected through neural pathways. The correlation was derived for each pair of continuous biomarkers using the correlation coefficient r, provided in the right-hand side of Table 2.5. The results highlight potential collinearity issues, with moderate correlations observed among $A\beta$, CSF tau and FDG-PET. FDG-PET imaging averages show a moderate correlation of .43 and .41 for the hippocampal and medial temporal lobe volumes, respectively. As expected, the brain ROIs show high correlation. This is especially true because the hippocampus is contained within the medial temporal lobe, which is - of course - contained within the brain. Future research should study mutually exclusive ROIs. What is most alarming to me is the strong negative correlations we see between FDG-PET and the mental exams (as high as -.62), and, similarly, the hippocampal volume and the mental exams (as high as -.53). Such results point towards a possible relationship among brain atrophy and poor performance on neuropsychological testing. A correlation plot of 15 different grey matter ROIs in AD patients is given in Figure 2.4. The regions with the highest correlations are concentrated in the left and right temporal lobes as well as the left and right angular gyri. This is consistent with past research that implicates the temporal lobe and hippocampus have potential spatial correlation maps within these regions (M. W. Weiner et al., 2015; M. W. Weiner, Veitch, Aisen, Beckett, Nigel, et al., 2017).

Figure 2.4: Heatmap of correlation among 15 ROIs.



2.2.5 Principal Component Analysis

Past work has shown that inherent spatial and temporal correlation exists among neuroimaging data, and highly correlated genetic pathways may express the dis-

ease under study. Component analysis (CA) can be seen as a first step to handle some of these issues of complexity as well as dimension reduction prior to analyses. Under this framework, a linear combination of the variables in the data are formed without the presence of a response variable and without parametric restrictions, permitting a completely data-driven method to uncover the natural heterogeneity. Principal component analysis is the most common method for dimension reduction. It affords analysts with a computationally simple approach backed by mathematical theory that is relatively understandable to those from many backgrounds. The aim is to visualize the features, reduce the dimensionality and mitigate the effects of multicollinearity before applying it to an inferential analysis. Since PCA does not tie the variables to any outcome variable, these inputs are called *features*. Take the matrix of d features, X_d. Traditionally, after normalizing the features, the eigenvalues, called loadings, and the eigenvectors, called the principal components (PC), of the covariance matrix $Cov(X)$ are derived using singular value decomposition (SVD). The number of PCs is equal to the rank of $Cov(X)$, k, and they are derived such that the PCs are orthogonal to each other. So this method effectively uncorrelates the data while reducing the dimension of the feature matrix from $d-$dimensional to $k-$dimensional. The weights of the PCs are subject-specific influences on the component. A linear combination of the PCs with the data may later be used as covariates in a downstream analysis. Similarly, the loadings are the eigenvalues of the eigenvectors, and the values provide the magnitude and direction of the contribution of each feature on that component. Thus, PCA captures the essence of the variability of many features while retaining interpretability.

PCA was derived from ADNI1 baseline data (n=790) for the following data sets: (i) mental exams (WLMS, CDR-SB, MMSE, ADAS-Cog13, RAVLT recall and RAVLT forgetfulness); (ii) ADNI1 biomarkers (APOEϵ_4, TAU, ABETA, FDG-PET, and grey matter volumes from the hippocampus, medial temporal lobe and whole brain volume); (iii) i and ii combined; (iv) grey matter volume of 15 ROIs and whole brain white matter hyperintensity (merged by two additional data sets in ADNIMERGE, *ucsdvol* and *ucd_ADNI1_wmh*, respectively); (v) the top 10 Alzgene SNPs, merged from the ADNI1 GWAS using PLINK, plus APOEϵ_4 and the poly-T variant TOMM40 allele one and two frequencies (merged by *adnimerge* and *tomm40* data sets); (vi) i-v combined ; (vii) ADNI2 biomarkers; (viii) ADNI2 biomarkers with the mental exam results.

Table 2.7 yields the cumulative proportion of variance for the first 3 PCs for each combination. The data set with the highest cumulative proportion explained by the third PC is the ADNI1 mental exam data, at 86%. The "next best" is model (vii), the ADNI2 biomarker data. A close third is (ii), which

uses the ADNI1 biomarker data, followed by the 15 ROIs and then the ADNI2 combination of mental and biomarker data. Looking at just (i) alone, one may think that perhaps the mental exam data is able to explain more of the variance overall. However, ADNI1 biomarker data, included in (ii) & (iii), have many NAs. PCA does not allow NAs in the data, so essentially I had to throw out all of the observations that yielded NAs across at least one of the features. This reduced models (ii),(iii) & (vi) down to n=290! Recall that model (vi) combines data from (i)-(v), which includes a total of 34 features. However, with the exception of the SNP data, each of the separate PCA models were able to account for a great proportion of the variance than (vi). This suggests that when many NAs enter the model, the increase in features and simultaneous decrease in n results in poorer results. ADNI2 data were included because there are fewer missing observations so that n=548. This explains the improvement from model (ii) to model (iii). One can see that the SNP model performed very poorly at accounting for variance among the data. In order to proceed with a PCA approach for SNP data, one would need to utilize a categorical version of PCA that may be fused with the neuroimaging technique.

Table 2.7: PCA: Cumulative proportion of variance for models i-viii

model	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
PC1	.698	.427	.472	.497	.156	.298	.447	.496
PC2	.805	.625	.610	.630	.285	.381	.656	.633
PC3	.866	.760	.692	.731	.393	.449	.776	.714

In addition to the relevant interpretation of a PCA model, PCA also allows one to visualize the highly variable and high-dimensional multivariate data through unique plots. Figures 2.5 and 2.6 provides the PCA biplots and 3d plots with the data points separated by baseline diagnosis, respectively. The diagnosis provided in the data for ADNI 1 is only 1 of 3 categories: CN, MCI and AD. The biplots give three sources of information: the eigenvectors of the variables, or the loadings (the arrows), the subject-specific PC weights plotted as PC1 on the x-axis and PC2 on the y-axis (the data points), and the diagnosis clusters (ellipses) divided by color (red=dementia, green=MCI, blue=CN). The direction of the arrow determines the positive/negative effect of the features on the PCs. The length of the arrow provides the overall contribution of that variable to the PC. When mental and biomarker data are combined, we observe longer eigenvectors representative of the biomarker data in comparison to the mental exam vectors. The plots with the most NAs and the genetic data displays the largest overlap of the diagnosis ellipses.The clustering in the plots is most clear for the mental exams, of course. This is to be expected because the diagnosis comes mostly from this quantity. An easier way to observe this clustering is in

3D, which incorporates an additional component, PC3, on the z-axis so that the variation explained matches the bottom row of Table 2.7. Once again, the most distinctive clustering is observed with the mental exam data. However, we do see considerable improvement for data (vii) and (viii).

The PCA models show that much of the variance in the ADNI data may also be explained by the neuroimaging and genetics data in addition to the neuropsychological outcomes. However, several problems remain when using PCA as a dimension reduction tool of MMNG data. For one, PCA doesn't handle NAs very well, so an adequate imputation method would be needed to fix this. Second, the incomplete data problem seems to exacerbate as the number of features increases. This is a severe issue for MMNG data because this example represents only a teeny tiny fraction of the dimensionality that will be used in the final analysis. Next, standard PCA doesn't allow room for a valuable integration of genetic data due to the different data type. Finally, the PCA method is useful in that it uncorrelates the data, but it may not fully parse out the complex connectivity among and within the phenotype-genotype relationships.

For example, the connectivity among the brain ROIs suggest that many regions of the brain work together for memory and recall tasks. Yet, PCA will not really tell us the combined effect of, say, hippocampus and ventricle data compared to the rest. It simply parses the loadings all at once per PC. In addition, linkage disequilibrium is a known source of confounding within DNA material, which could result in masking the interactive effects by modeling them as separate entities. For instance, for 15 years past research have confirmed that APOEϵ_4 is highly associated with AD. However, TOMM40, also located on chromosome 11, has shown significant association with AD as well. These two genes are known to be in linkage disequilibrium. A closer look at the SNP results shows the other individual SNPs from the Alzgene database as carrying more of the weight in the PC derivations. Could an additional problem here be that this model doesn't include the interactive effects of TOMM40 and APOEϵ_4? Past research in bioinformatics suggests that specialized versions of PCA such as sparse PCA or supervised PCA will reflect more realistic results for genetic data. (Lever et al., 2017; Ma & Dai, 2011; Price et al., 2006).

2.2.6 Classification

The EDA and component analysis give strong evidence that these biomarkers have the potential to be classifiers or predictors for assessing AD risk or progression. Past research confirms these results, as imaging genetics has applied advanced techniques for the joint analysis of two identifiers at a time (Hao et al.,

Figure 2.5: Biplots for PCA from models i-viii

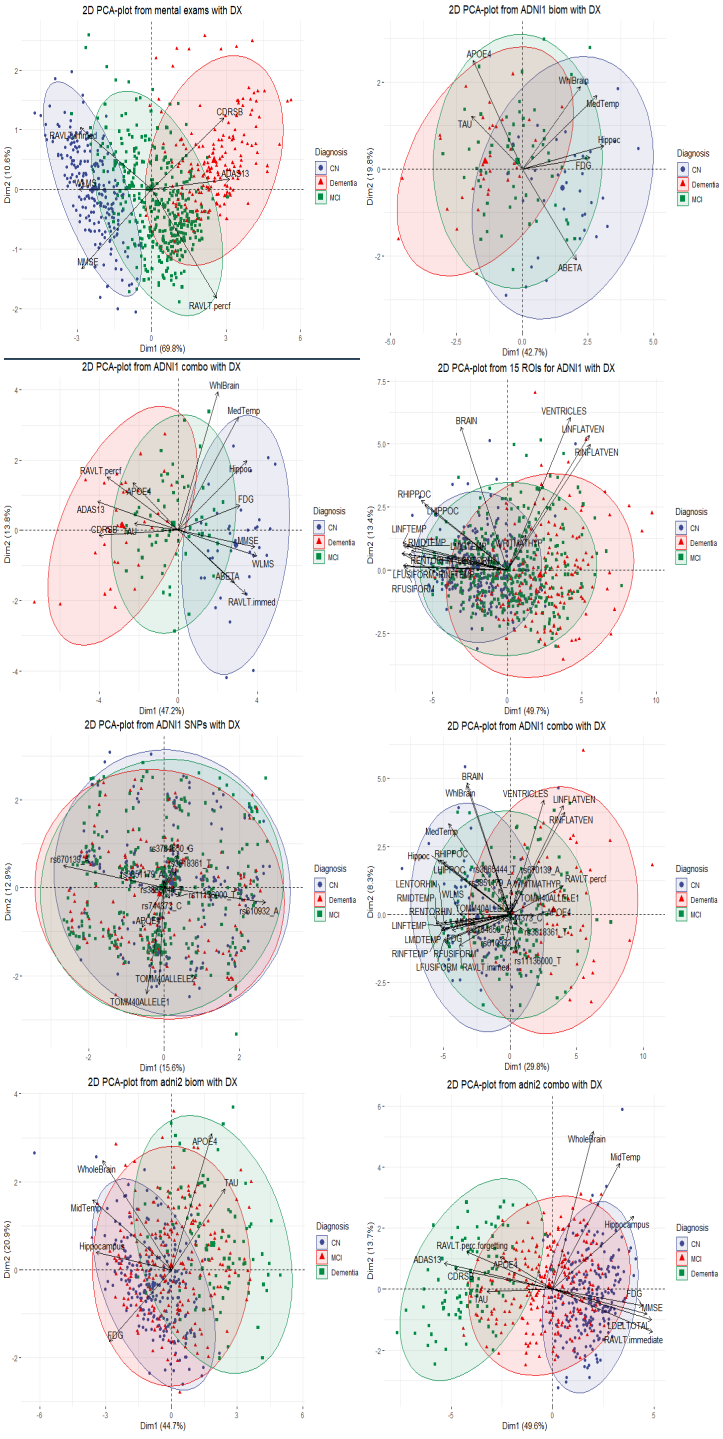
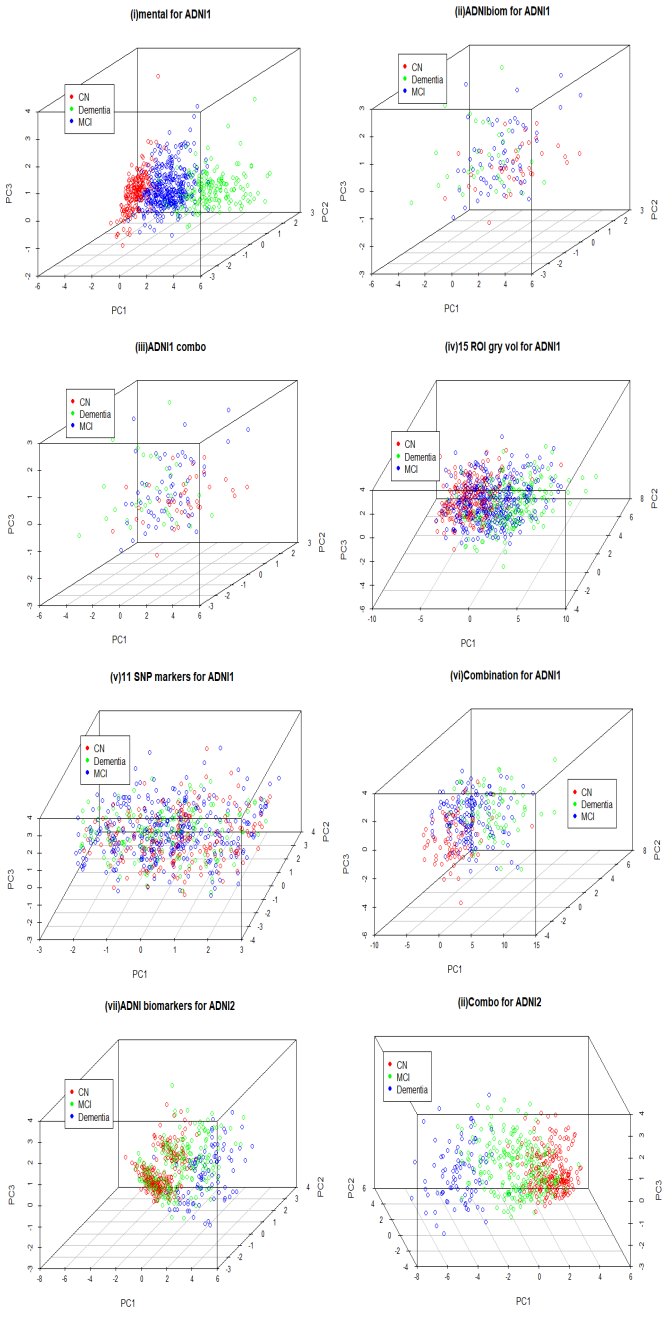


Figure 2.6: 3D Plots for PCA from models i-viii



2020a; S. Liu et al., 2015; D. Zhang et al., 2011). The goal of imaging genetics is to understand how psychiatric illnesses develop and progress through the combination of biological components, with strength of diagnosis resulting from

objective measurable traits rather than only mental outcomes. Nonetheless, many of these mental illnesses continue to be classified into the disease stages by the neuropsychological test results alone. How can scientists truly understand the biological basis if the diagnoses are guided by tests that have already shown late onset and high variability?

The idea of classification is to use patterns among the features of the data matrix to separate the data into groups by the natural relationships observed. First, the data are randomly divided so that 3/4 of the data are used for training, then linear discriminant analysis (LDA) is carried out on this portion of the data. The results are then tested on the remaining 1/4 of the (test) data, and the accuracy with which the testing data was divided into the categories is derived based on the actual DX values. Thus, LDA is considered a supervised learning approach (whereas PCA and ICA are unsupervised), where linear combinations of the features are derived while creating a new latent variable for each that accounts for the multiple classes. The number of linear combinations are equal to the number of predictors or the number of groups minus one, whichever is smaller. These are the discriminant functions that are formed after removing the redundant variables, so that the first function maximizes the differences between the groups on that function. The second function is also maximal but is uncorrelated to the first, and so on. Similar to PCA, LDA preserves interpretation since each discriminant is assigned a score that determines how well it can predict group placement. In this way, structured correlation coefficients may be derived and each predictor's weight in the linear combination is the standardized coefficient. Furthermore, the eigenvalues are the characteristic roots of each function, and the magnitude of the eigenvalue displays how well that function differentiates the groups.

For this research, the goal is to use the neuroimaging, genetic and mental data as the features in order to establish a pattern by which subjects may be divided up into the diagnosis categories, CN, MCI and AD. Six data sets were used: (i) mental data, (ii) ADNI1 biomarker data, (iii) i and ii combined, (iv) the 15 ROI grey volumes, (v) the top 10 Alzgene SNPs with APOEϵ_4 and two TOMM40 allele frequencies and (vi) iv and v combined. Five quantities were derived from the classification procedures that will describe how well the features in the data are able to appropriately group the subjects into the diagnosis categories, accuracy, a 95% confidence interval (CI) on the accuracy, sensitivity, specificity, and area under the curve (AUC). Finally, the receiver operator curves (ROC) are plotted.

LDA is traditionally a binary classification tool, however we have three classes. While it's possible to extend this to observe three classes, the classifi-

cation tasks are broken up into three categories, CN-AD, MCI-AD, and CN-MCI, using data sets (i)-(vi), to see if any one binary classification is more informative than the others. When deriving the results, a confusion matrix is computed, with the number of true positives (TP) and false positives (FP) in the first row and the number of false negatives (FN) and true negatives (TN) in the second row. The five quantities are derived from this table. The accuracy is defined as the number of correct predictions over the total number of predictions (TP+TN)/(TP+FP+FN+TN). Sensitivity is the true positive rate, or the propotion of those correctly identified with the disease, TP/(TP+FP). Similarly, the specificity is the true negative rate, or those accurately identified without the disease, TP/(TP+FN). The AUC is best described by understanding the ROC. The ROC plot is of the sensitivity (TP rate) plotted against 1-specificity, the FP rate. AUC is then the phyiscal 2D area under the ROC.

Table 2.8: LDA classification results for CN-AD, MCI-AD, CN-MCI stages.

Stage	Stats	Mental	ADNI Biom	Mntl + Biom	15 ROI vol	Top Alzgene	ROI + Alz
CN-AD	Accur	1	.818	1	.808	.657	.838
	95% CI	(.96,1)	(.73,.89)	(.96,1)	(.72,.88)	(.55,.75)	(.75,.91)
	Sensit	1	.711	1	.725	.478	.721
	Specif	1	.885	1	.864	.811	.929
	AUC	1	.867	1	.911	.689	.889
MCI-AD	Accur	.912	.699	.926	.699	.684	.691
	95% CI	(.851,.954)	(.614,.774)	(.869,.964)	(.614,.774)	(.568,.734)	(.583,.747)
	Sensit	.756	.348	.886	.304	.044	.315
	Specif	.979	.878	.946	.900	.967	.902
	AUC	.98	.685	.974	.679	.504	.666
CN-MCI	Accur	.979	.759	.972	.745	.614	.703
	95% CI	(.941,.996)	(.681,.826)	(.931,.992)	(.666,.814)	(.529,.693)	(.622,.776)
	Sensit	.989	.791	.956	.821	.717	.795
	Specif	.964	.704	1	.621	.391	.561
	AUC	.999	.828	.995	.836	.654	.772

The classification results may be found in Table 2.8 and the plots viewed in Figure 2.7. It is important to think about what these results mean in order to understand how well LDA classifies the diagnoses using the features as inputs. The overall accuracy is, of course, very important, but this is not all that should be considered. A high sensitivity in this case means that a high proportion of those who are classified in the advanced stage of AD are correctly identified, while a high specificity means that a high proportion of those who actually have the advanced stage of AD are identified. It is ideal if both are high, but there is somewhat of a trade-off. Which is most important? When one considers possible treatments, I believe it is better to identify the largest proportion of actual

positives, meaning less AD patients go undetected and can receive treatment (specificity). We expect the mental exams to obtain the highest results of all the data sets since this is the actual diagnostic criterion. However, if a combination of mental exams and biomarkers or any combination of the biomarkers achieves similar results, then this is excellent evidence for using biomarkers in addition to mental exams as classifiers.

As expected, the largest differences of all the features would be most noticeable between the CN-AD classes. It turns out that data (i) and the combination, (iii), have equivalent results. The ADNI1 biomarkers alone achieve an accuracy of 81.8% which is within the range of suggested accuracy provided by the ADNI research groups (M. W. Weiner et al., 2015; M. W. Weiner, Veitch, Aisen, Beckett, Nigel, et al., 2017). However, the 15 grey matter ROI volumes yield a similar classification accuracy, with the added Alzgene data from (vi) achieving 83.8%. The specificity of this data is excellent with 92.9%. When moving to the MCI-AD category, (iii) gives slightly better results than (i) in terms of accuracy, while the specificity of the mental exams is higher. Surprisingly, the ROIs perform poorly overall, with a low accuracy and sensitivity yet still a high specificity at 90%. In fact these results are similar for Alzgene and (vi). It's possible that the change in biomarkers occurs more drastically during earlier stages of AD, which would explain this drop in performance. Of most interest to me is the ability to detect the change from CN-MCI, because this would display the ability to detect the earliest diagnosis progression. The AUC for (i) is .999, with very high sensitivity and specificity. We see a drop in the ADNI1 biomarkers to 75.9%, not as close as was observed for the CN-AD category. The combination of mental and biomarkers yields similar results to the mental exams, but the specificity rises to 100%. The ROI results are similar to (ii), and both (v) and (vi) could use improvements.

There are some drawbacks to the LDA approach. First, these percentages reflect the accuracy with the actual diagnosis, which is somewhat subjective and can change within the first year. Second, LDA assumes multivariate normality for each level of the grouping variables. When comparing the diagnosis with the predictors in Section 2.2.3, often the nonparametric method was used due to a lack of normality in the data. Much of the data are skewed within the different disease categories. Also, LDA performs best for higher levels of collinearity. While some of the predictors revealed moderately high correlations among the features, some were more moderate-to-low. Finally, random samples are assumed, yet we have already mentioned the unavoidable volunteer bias that is likely present. For the use of larger data, violations of these properties must be

accounted for. Past research suggests that more flexible forms of discriminant analysis are probably best for a multi-class MMNG classification.

2.2.7 Longitudinal Modeling

The longitudinal collection of multi-source data allows researchers to track the progression of AD by observing changes in neuroimaging phenotypes over time. Again, this neurodegeneration is referred to as brain atrophy. There is much research to support the fact that the brain degenerates throughout the course of AD, so that the overall brain volume in the data diminishes as AD symptoms worsen. It is well understood that certain ROIs are responsible for multiple functions and that, often, various regions also work together to achieve a response. In this section, 15 separate ROIs are examined for the change in brain volume over the first year of the ADNI1 study. Then, a linear mixed model with random subject-specific and time-specific effects is used to study how the volume of the top brain ROI changes over the course of the first two years. In order to understand how the brain atrophies in patients over the course of the first year of study, I calculated the change in brain volume between baseline and the 12 month measurements for all 15 grey matter ROIs from the dataset *uscd_vol*. That is, the phenotypic response variable is the difference in brain volumes, with the baseline volume subtracted from the 12 month volume. The explanatory variables considered are all six neuropsychological exams, the top 10 significant genes from the Alzgene meta-analysis, the poly-T variant TOMM40 allele1 and allele2 frequencies, FDG-PET imaging average, CSF tau protein, gender and age.

Table 2.9: Overall MANOVA results on volume of 15 grey matter ROIs.

Variable	WLMS	CDRSB	ADAS13	MMSE	RAVLTi	RAVLTp
Approx F	1.4	1.47	1.47	0.94	0.87	1.02
Pr(>F)	0.161	0.132	0.131	0.519	0.600	0.441
Variable	AGE	GENDER	DX	APOE4	FDG	r1179
Approx F	4.78	4.43	0.95	1.25	2.87	1.18
Pr(>F)	0.000	0.000	0.553	0.186	0.001	0.246
Variable	r6000	r5444	r4650	r8361	r0139	r0932
Approx F	1.38	1.47	0.93	0.83	1.16	0.69
Pr(>F)	0.100	0.065	0.571	0.724	0.270	0.888
Variable	r4373	TOMM1	TOMM2	TAU		
Approx F	1.21	0.55	0.79	2.07		
Pr(>F)	0.223	0.907	0.682	0.017		

For this basic example, a multivariate regression model was run. Each of the genetic markers were taken as factor variables to keep track of which gene expressions would predict an increase or decrease in brain volume. It is suspected

that the presence of two minor alleles in the 0,1,2 coding will yield a greater impact on brain volume over the course of the first year. In addition, the diagnosis variable, DX, is modeled as a factor with three categories CN, MCI, and AD. It will also be interesting to see how the test statistics of the biomarkers will compare with the mental exam results. Keep in mind that using a multivariate linear approach assumes independence among the ROIs, which we know is not accurate. Table 2.9 yields the Type II MANOVA results. This tests the overall impact of the predictors on the 15 response variables. The highest test statistics are observed on variables age and gender. FDG and CSF tau, reflections of functional neuroimaging, quantities are associated with the brain volumes controlling for all other variables in the model.

However, when one breaks this study up into the 15 separate regression models, the results change drastically. A heat map of the p-values can be found in Figure 2.8. These results begin to uncover a relationship with the ROIs (on the x-axis). The darkest blue reflects the lowest p-value, which matches two minor alleles of APOEϵ_4 with the right hippocampus. This is not surprising, given the background research on hippocampal volume and AD progression. We also notice that some of the mental exams, such as CDR-SB, ADAS-13 and MMSE are associated with different areas of the brain, hippocampus, medial temporal, and left inferior temporal, respectively. The TOMM40 alleles carry weight in the hippocampus and right entorhinal ROI.

Based on the preliminary analysis in this paper and past research, the top ROI of consideration is the hippocampus. Degeneration in this area has shown to be associated with a progression in AD. Now, I model the hippocampal volume as it changes for subjects over the course of two years. Using a linear mixed model with random effects, we can allow a more flexible covariance structure among the predictors. Fixed effects included are the six mental exam scores, FDG-PET, $A\beta$, tau, and the factors APOEϵ_4, gender, DX, VISCOME (the time unit in months) and interaction effects of DX*VISC and DX*APOEϵ_4. A subject-specific random intercept is incorporated with unstructured covariance matrix, and a subject-specific random slope on time (VISCODE over months 0, 6, 12, 18, 24) is incorporated with the AR(1) covariance structure. These covariance structures were included because this yielded the lowest AIC when compared to the unstructured, compound symmetry, and spatial-Gaussian structures using SAS.

Table 2.10 shows the Type III Fixed Effects of the model. Four of the six mental exams show high F test statistics when controlling for the other variables, especially ADAS-13. The main effects of FDG, DX, time, gender and age are highly significant as well as the interaction of APOE with DX. Of further in-

Table 2.10: Type III fixed effects from LMM on hippocampal volume change over the course of two years.

Variable	WLS	RAVp	RAVi	MMS	ADS	CDR
F val	5.78	4.0	6.02	.00	14.2	.05
P(>F)	.017	.047	.015	.945	.000	.828
Variable	FDG	ABTA	TAU	APO_4	AGE	GNDR
F val	27.2	1.37	2.68	.98	132.5	50.7
P(>F)	.000	.243.103	.377	.000	.000	
Variable	DX	VISC	VIS∗DX	VIS*APO		
F val	4.86	31	.52	3.8		
P(>F)	.009	.000	.721	.005		

terest is the actual quantities of the estimates because this breaks the factors up into the categories, allows one to see whether the effect is positive or negative on brain volume and gives the average brain volume increase or decrease per unit change in the explanatory variable. Interaction plots of DX*VISC and DX*APOEϵ_4 may be found in Figure 2.8. The plot on the left shows how the hippocampal volume declines over the course of the study and the different colored lines represent the disease statuses. Although CN experiences a decline in volume, the overall volume begins at a lower point for both MCI and AD. By year two the greatest decline is observed for MCI. The plot on the right shows the significant interaction among the diagnosis categories and the APOEϵ_4 allele frequencies. The volume is closer together for 0 minor alleles, with CN having the highest volume and MCI with the lowest volume. The decline in volume from 0 to 1 minor allele is more drastic for those with AD, and only slighty lower for MCI. Finally, the largest difference is observed for two alleles, and the lowest with MCI. Future research should include more ROIs into the picture, and later protocols (with less missing data) will be employed.

Figure 2.7: ROC curves for multi-class LDA classification.

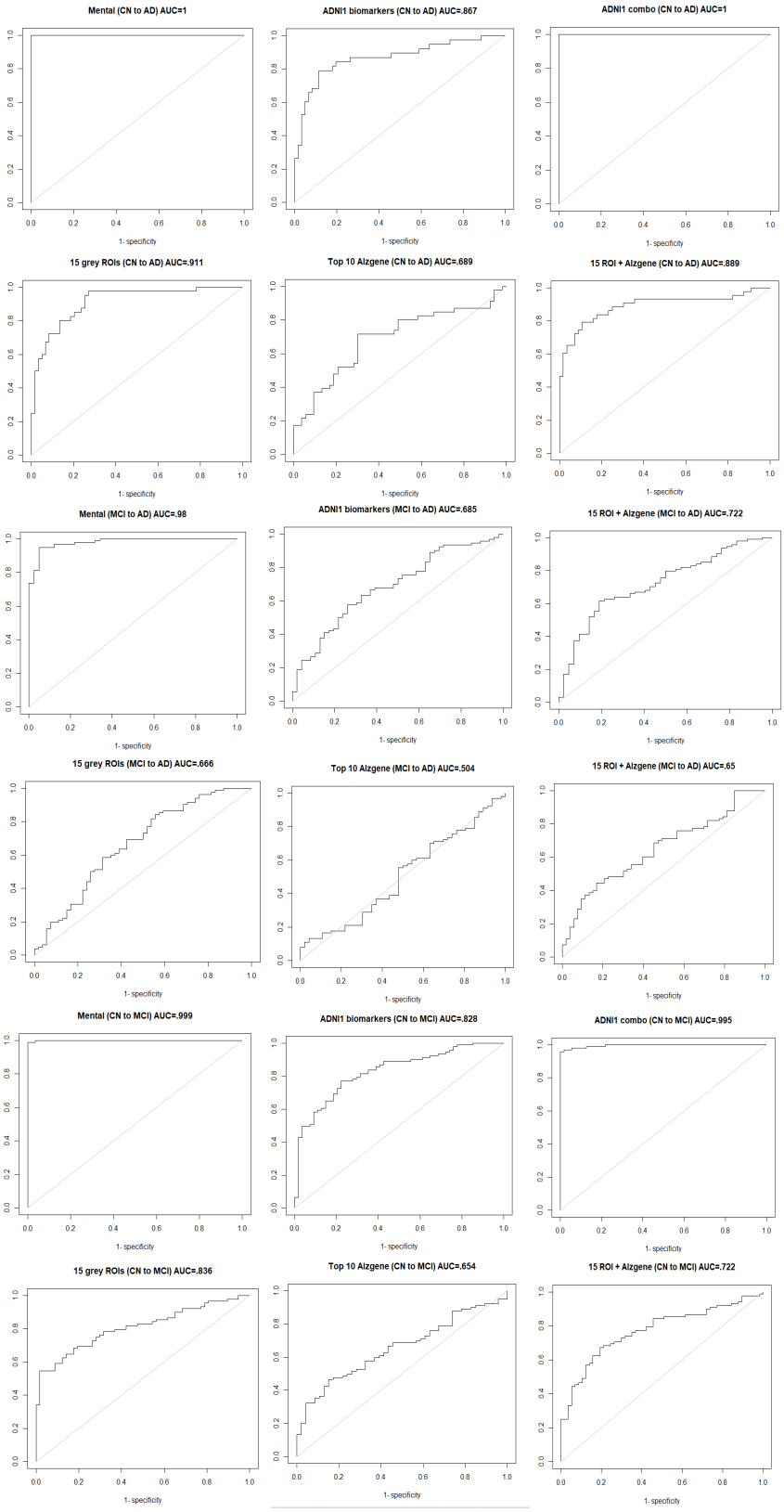


Figure 2.8: Heatmap of p-values for 12 month change in 15 ROIs.

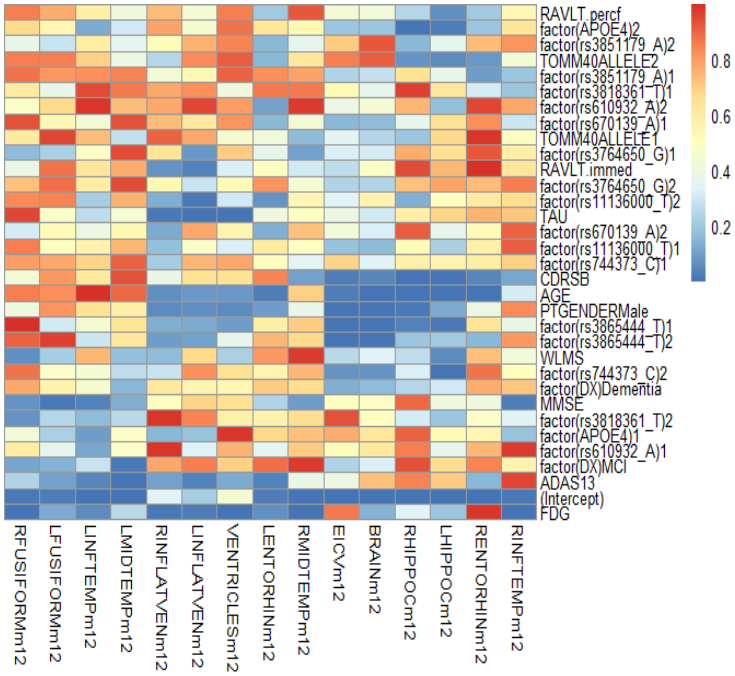
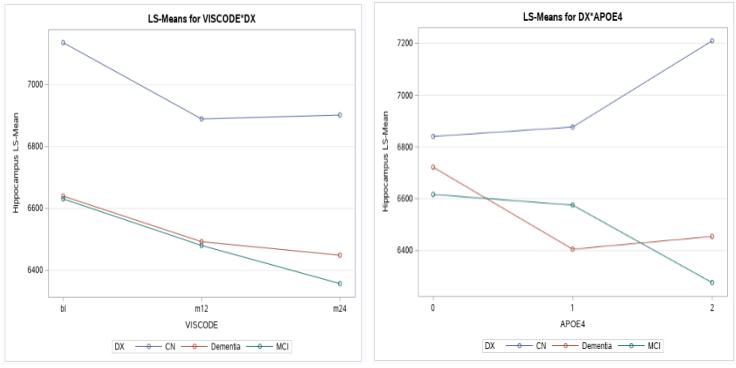


Figure 2.9: Hippocampal volume change over 2 years: DX by VISCODE (left); DX by APOE4 (right).



Chapter 3

Multi-Modal Methods

Thus far, the case for multi-modal methods have been presented and past models discussed and introduced with motivation for imaging genetics applications. After data collection, pre-processing and cleaning, there are three important steps to the methodological process of MMNG in the ICA framework: model pre-processing, the fusion framework, and the optimization approach. While recent literature have focused on developing the fusion framework, much of the first and third steps in this process retain traditional methods that may not scale well to the high-dimensional MMNG scenario. Although some studies have included genetic data, the fusion framework is mostly focused on neuroimaging data. Past researchers have not assessed the statistical reliability of the component loadings when adding the new diversity of genetic data. In addition, the optimization approaches used at large were introduced early in ICA history, prior to this data revolution and without facing the big data problem.

In this chapter, I begin by diving into the mathematics that underscore past multi-modal methods and discuss the theoretical implications of these methods in the imaging genetics scenario. Two mainstream multi-modal decomposition frameworks are compared, jICA and paraICA. Novel to this research is the formulation of parallel tIVA with mCCA as an alternative to allow the complexity of various imaging modalities to be modeled in unison with genetic material. I then present a new class of unconstrained multi-modal data fusion optimization techniques, involving linear and nonlinear approaches. The overall aim of the methods I propose is to leverage the strengths of previously proposed fusion strategies, while introducing a neural network optimization into the multi-modal ICA framework. The goal is to ease the complexity in computation and improve statistical reliability in a high-dimensional environment.

Then, a multi-way simulation setting will be built to assess the statistical reliability of the algorithms. Also, as a new addition to MMNG research, three dif-

ferent modalities are simulated, SNP, sMRI and rsfMRI to test the algorithms in the three most common data sources of MMNG problems. Finally, with all these aims in sight, the creation of a generalizable MMNG Fusion Pipeline is built. Important issues such as data source selection, data manipulation and preprocessing, as well as the computational implementation are discussed in detail within this pipeline. Future researchers may follow such a strategy for future analyses of ADNI data or for the study of other neuropsychiatric diseases.

3.1 Formulation of Multi-Modal Fusion

Multiple data fusion strategies have been discussed with particular focus placed on two schools of multi-modal analytics, one being decomposition methods with the other approach being deep learning. Here, I will present the theory underlying joint ICA and parallel ICA and discuss any accompanying pitfalls. Then I will extend the parallel ICA method to include tIVA with mCCA, introducing a novel structure to a three-way MMNG problem. Then, in the next section, the algorithms used for optimization are described with the inclusion of a new extension to the multi-modal setting that permits unconstrained optimization, containing similar properties to neural network algorithms and imposing sparsity for dimension reduction. The full model development is discussed at length to describe how the techniques pull together the advantages from past models to produce a more robust, algorithmically efficient and generalizable multi-modal analysis.

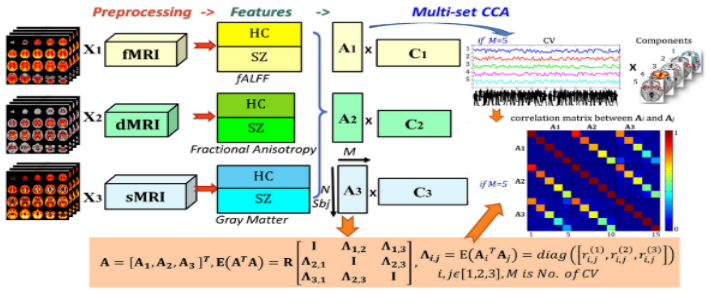
3.1.1 Pre-Processing of Multi-way ICA

Prior to introducing each step of MMNG ICA, I'll introduce a traditional but effective method for inferring information on the cross-variances among different modalities. Given the normalization and whitening that occurs with ICA (to be discussed in the next few paragraphs), the resulting structure of the data permits a standard canonical correlation analysis (CCA) in a multi-source setting (mCCA). This concurrently applies a symmetric data-driven decomposition method that is fully derived from the patterns extracted from the interlinked modalities. As mentioned earlier, jICA derives one unmixing matrix to form the independent components of the fused data. However, the assumption of equivalent covariation among the separate disease groups is unrealistic when pulling data from at least two distinct data sources. This assumption is softened by applying mCCA prior to deriving the component analysis. One major difference between jICA and paraICA is the order in which ICA and

mCCA methods are implemented. Thus, paraICA allows a unique mixing matrix for each modality and then performs mCCA to properly fuse the data sources. This may be the desired choice for highly diverse sets such as mixing imaging with genetic data, but this case must be explored with real data.

Under the mCCA framework, blind separation is performed by decomposing the data into canonical variants (CVs) for each of the modalities, utilizing shared information between the modalities. This may be done in a pairwise manner for two groups at a time to later allow for group-level inference in post-hoc component analyses. That is, each of the data sources obtains its own set of CVs, or linked variables, that are derived from maximizing the inter-subject covariation across two features at a time. This method works best when the values of the CVs are truly distinct. The CVs are then fused and used as input into the jICA model, further decomposing the remaining mixtures into independent components.

Figure 3.1: Flowchart of data fusion using mCCA, p.797 (Sui et al., 2015)



The flowchart in Figure 3.1 shows this process of mCCA+jICA Sui et al decomposed the three imaging modalities into features separated by schizophrenic and healthy groups, from which the mixing coefficients A_i were derived. The matrix $\mathbf{A}$ is the pairwise correlation matrix between each of the $i = 1, ..., M$ modalities, among the disease groups. The matrix $\mathbf{C}$ are the components, or i spatial maps, by which the pairwise modality correlations of A_i and A_j $(i \neq j)$ were maximized to form the elements $\Lambda_{i,j}$ (Sui et al., 2015). These matrices were obtained by maximizing the sum of squares of the correlation between the demixing vectors $\mathbf{w_i}$. Now, consider the extension of CCA to n-way modalities. This is formulated as a two-stage process for optimization of $g = 1, ..., G$ groups (Y. O. Li et al., 2009):

$$stage(1) : \{\mathbf{w_1}^{(1)}, \mathbf{w_2}^{(1)}, ..., \mathbf{w_M}^{(1)}\} = \arg\max_{\mathbf{w}} \sum_{i,j=1}^{M} |r_{i,j}^{(1)}|^2$$

$$stage(2) : \{\mathbf{w_1}^{(g)}, \mathbf{w_2}^{(g)}, ..., \mathbf{w_M}^{(g)}\} = \arg\max_{\mathbf{w}} \sum_{i,j=1}^{M} |r_{i,j}^{(g)}|^2, \tag{3.1}$$

$$\text{such that } \mathbf{w_M}^{(g)} \perp \{\mathbf{w_1}^{(g)}, \mathbf{w_2}^{(g)}, ..., \mathbf{w_M}^{(g-1)}\}.$$

The application of mCCA in equation (3.1) is now extended to a multimodal decomposition analysis. Prior to applying mCCA, PCA is applied to the modalities individually. After this calculation, each modality has it's own diagonal and square matrices consisting of diagonal eigenvalues and column-wise eigenvectors, $\Lambda_k = \{\lambda_{k1}, \lambda_{k2}, \lambda_{kN}\}$ and $\mathbf{B}_k$, respectively, of the covariance matrix, $E\{X_k \cdot X_k^T\}$. In the whitening process, the top M_k eigenvectors with the largest λ_{ki} are selected s.t. Λ_k' is of size $M_k \times M_k$ and $\mathbf{B}_k' : M_k' \times N$ (Sui et al., 2010). For the resulting whitening matrix, $\mathbf{D}_k$, the whitened PCs are given by

$$\mathbf{Y}_k = \Lambda_k^{-\frac{1}{2}} \cdot \mathbf{B}_k' \mathbf{X}_k = \mathbf{D}_k \cdot \mathbf{X}_k, k = 1, ..., M. \tag{3.2}$$

Under jICA, the CVs are derived using the whitened PCs from each of the data sets as inputs. This is achieved through linear transformations on $\mathbf{Y}_k$ that remove the between-set cross-correlations, resulting in the canonical variants (CV) (Sui et al., 2010), defined such that:

$$\mathbf{c}_k = \mathbf{C}_k \mathbf{Y}_k, \text{ s.t. } E\{\mathbf{c}_k \mathbf{c}_k^T\} = \mathbf{I}, k = 1, 2$$

$$E\{\mathbf{c}_1 \mathbf{c}_2^T\} = E\{\mathbf{c}_2^T \mathbf{c}_1\} = \mathbf{L} = \ell_1, \ell_2, ..., \ell_M, M = min(M_1, M_2) \tag{3.3}$$

for canonical transformation matrices, $\mathbf{C}_1$ and $\mathbf{C}_2$ that are calculated by the eigenvalue decomposition problem:

$$\mathbf{C}_1 \cdot E\{\mathbf{Y}_1 \mathbf{Y}_2^T\} E\{\mathbf{Y}_2 \mathbf{Y}_1^2\} \cdot \mathbf{C}_1^T = \mathbf{L}^2 \tag{3.4}$$

$$\mathbf{C}_2 = \mathbf{L}^{-1} \mathbf{C}_1 \cdot E\{\mathbf{Y}_1 \mathbf{Y}_2^T\} \tag{3.5}$$

$$\rightarrow E\{\mathbf{c}_1 \mathbf{c}_2^T\} = E\{\mathbf{C}_1 \mathbf{Y}_1 \mathbf{Y}_2^T \mathbf{C}_2^T\}. \tag{3.6}$$

For M modalities, we apply mCCA so that this is done in the two-stage approach given in equation (3.1). Each resulting CV is the cross-correlation vector with every other modality. In other words, there is one vector of canonical variants for each feature within the modality subtracted by 1, $[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{M-1}]$. Although this is a well-known and effective approach for deriving correlation links among the data sources, it falls short in separating sources with close or approximate correlation coefficients, as is often the case across neuroimaging techniques. For this reason, jICA is applied on the extracted CVs in order to carry out the multi-modal data fusion process. For the case of pICA, the ICA step will occur prior to the mCCA derivation above.

3.1.2 Multi-modal Independent Component Analysis

Now, I will show the derivation for multi-modal independent component analysis, further formulating jICA and paraICA. An extension of single ICA (ICA applied to a single modality) is transposed independent vector analysis (tIVA). Past research has not extended this to MMNG data, nor has it been followed up with mCCA. Therefore, I propose parallel tIVA with mCCA (p-tIVA+mCCA). In this dissertation, the proposed framework will be compared with mCCA+jICA useing two traditional optimization techniques and two novel techniques. In this derivation, mCCA fits into both of these models in unique ways. Start with the data from $i = 1, ..., M$ modalities, $\{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M\}$, whether genetic, imaging, or the like. As previously mentioned, there are steps to apply to each data source prior to applying single-sourced ICA. However, these same pre-processing techniques are not applied in the IVA setting. Consequently, the data are scaled and centered. Then, the algorithm must incorporate dimension reduction in order to be applicable to the multi-modal framework. Algorithms will be discussed in Section 3.1.3.

For tIVA, each individual data set may be expressed as an instantaneous mixture of latent sources, $\mathbf{s}_i$, per voxel, time point or genetic marker,such that

$$\mathbf{X}(v) = \mathbf{A}\mathbf{s}(r), \;\; 1 \leq v \leq V, \;\; \mathbf{X}(v), \mathbf{s}(v) \in R^N. \tag{3.7}$$

Note that $j = 1, ..., N$ subjects are collected in each of these modalities, for the moment, and there are $k = 1, ..., K$ sources, or components, selected from each of the M modalities. The matrix $\mathbf{A}$ is a full rank mixing matrix, where

$$\mathbf{U}(v) = \mathbf{W}\mathbf{X}(v), \;\; \mathbf{W} = \mathbf{A}^{-1} \text{ s.t. } \mathbf{s}(v) \equiv \mathbf{U}(v). \tag{3.8}$$

In the IVA formulation (3.8), $\mathbf{W}$ is the unmixing matrix that connects the k-dimensional noiseless signals, $\mathbf{s}$, to the N subject-specific weights estimated.

That is, the magnitude of the row-wise weights provides a summarization for each subject along with the influence at each column-wise feature (component) extracted. Thus, decomposition simultaneously reduces the data to a smaller dimension while extracting the features most relevant to the data fusion process. The interpretability is then preserved with feature- and subject-related weights that have a direct expression with the most biologically meaningful entities of the data.

Joint ICA with mCCA

With jICA, the data matrices $\mathbf{X}(v)$ are replaced with $\mathbf{C}_i, i = 1, ..., M$, defined in equations (3.3-3.6), maximizing the independence among the correlations to further parse out similar variants. That is, the k ICs extracted with this method are obtained by the following transformation with the original data $\mathbf{X}_i$ such that

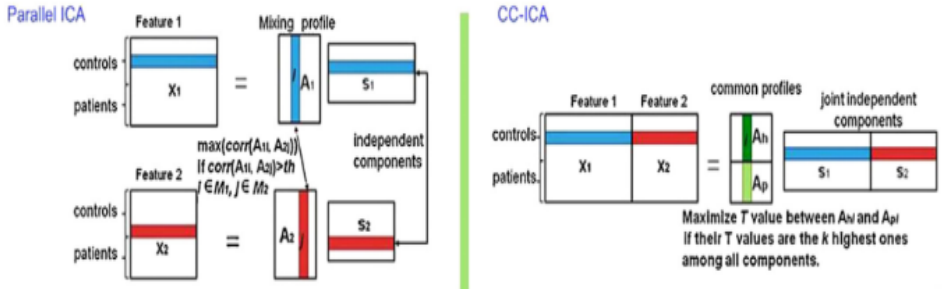
$$\mathbf{s}_k = (\mathbf{W}\mathbf{C}_k\mathbf{D}_k) \cdot \mathbf{X}_i = \mathbf{W}\mathbf{X}_i\mathbf{A}_k = (\mathbf{W}\mathbf{C}_k\mathbf{D}_k)^{-1}, k = 1, ..., K \quad (3.9)$$

From there, one may then sort each of the components by the strength of the correlation with the other ICs, $r_1, r_2, ..., r_M$. Thus, the resulting decomposition is a $N \times K$ matrix with vectors that represent the maximally interrelated structures among the neuroimaging and SNP data with rows equal to the number of subjects. These results may be treated as a sufficiently fused data set and are ready for use in downstream analyses. For example, how do these correlation relationships relate to the diagnosis outcome? May we also use the matrix of linkages to predict disease status while controlling for important subject characteristics, like sex and age? In theory, yes, one may use these vectors in a downstream analysis involving any of the above: a linear mixed model, a longitudinal analysis, a classification approach, or even clustering.

It is important to keep in mind the trade-off that occurs in any dimension reduction and feature extraction process. For jICA, there are 4 levels of pruning the data until they are fully decomposed: PCA, whitening, mCCA and then jICA. Two final steps will reconstruct the data from these reduced components by first multiplying the component mixing matrix with the dewhitened matrix, known as back reconstruction, and then scale the components back to the units of the data by using multiple regression of the components with the original data and flipping the signs depending on the signs of the beta weights from the model (Sui et al., 2010). While the fused data are relayed back to the original data, it may be true that "too much" reduction will have taken place, or the model may be overestimated. Thus, mCCA+jICA may not be the best choice for models

with distinct pathways that are jointly expressed from cross-relationships among the modalities. For this reason, a new model will be produced that permits more flexibility in the fusion framework.

Figure 3.2: Semi-blind data-driven decomposition methods, p.234 (Calhoun & Sui, 2016)



N-way Parallel IVA with mCCA

Instead of deriving one unmixing matrix as in jICA, paraICA derives one unmixing matrix for each modality and then performs correlation analysis between the two. This doesn't force a shared covariance matrix, and likely permits more diversity in the modalities that may be incorporated into multi-modal data fusion. Especially when moving outside of neuroimaging modalities and including very different data sets, this flexibility may be necessary. The fusion structure discussed so far may be visualized in Figure 3.2 as a flow diagram of the two semi-blind approaches introduced by Calhoun et al (paraICA on the left) (Calhoun & Sui, 2016). Coefficient-constrained ICA (ccICA), on the right in Figure 3.2, is closely related to mCCA+jICA in that it concatenates the modalities for data fusion while controlling for group separation. This is done under semi-blind constraints where components are prioritized based on those that show significant group differences (Sui et al., 2009). On the left, the paraICA approach derives unique mixing matrices, $\{A_1, A_2\}$ for the case of two modalities presented in the diagram. The cross-correlations are then analyzed using the resulting sources, $\{s_1, s_2\}$.

Although n-way jICA was introduced by Sui et al (Sui et al., 2013), 3-way tIVA has not been built into the paraICA framework. I first describe an n-way pICA, then extend this to p-tIVA. Following closely with the mathematical formulation introduced for parallel-ICA by Liu et al (J. Liu et al., 2008), let $i = 1, ...M$ represent the M modalities. Then the constraints are imposed such that $i \neq i'$, for mixing matrices $A_1, A_2, ..., A_m$ and $S_1, S_2..., S_m$ independent

components for M modalities:

$$X_M = A_M S_M;\ Z_M = W_M X_M;\ W_M = A_M^{-1};\ \text{then}\ Z_M = S_M; \tag{3.10}$$

$$\max\{H(Y_M)\} = -E[\ln f(Y_M)];\ Y_M = (1+e^{-U_M})^{-1};\ U_M = W_M X_M + W_{0M} \tag{3.11}$$

$$\max\{H(Y_1) + H(Y_2)... + H(Y_M) + \sum_{i \neq j} Corr(A_i, A_j)^2\} =$$

$$- E[\ln f(Y_1)] - E[\ln f(Y_2)] - ... - E[\ln f(Y_M)] + \sum_{i \neq j} \frac{Cov(A_i, A_j)^2}{V(A_i)V(A_k)}. \tag{3.12}$$

Constraints (3.10) and (3.11) are maximized in parallel using gradient maximization, following (J. Liu et al., 2009). The right-hand side of (3.12) is optimized using the steepest descent method with the step size calculated at each iteration, and a learning rate λ for each modality. This typically employs the Infomax Algorithm, allowing nonlinearity in the model but remaining constrained in the optimization approach. Another algorithm of popularity is FastICA, which maximizes the neg-entropy function. More on these algorithms will be discussed in Section 1.1.3.

The process of an n-way paraICA model is tied to the theory of independent vector analysis (IVA), without the restriction of model pre-processing as in paraICA. Instead, tIVA was built for multiple data sets in order to encompass a more diverse class of signals in the feature construction process. It is worth noting that aNyway-ICA was built as a three-way IVA approach that extracts subspace component vectors (SCV) prior to applying a group-based component analysis, but has not since been applied to real data (Duan et al., 2020) or considered at the subject-level. In addition, aNyway-ICA does not perform fusion of the different modalities into one matrix. In this body of work, the minimization of mutual information is carried out with IVA-G. Instead, in the method presented here, I add a post-processing step, mCCA, on the outputs of the p-tIVA model to concatenate the IC vectors by deriving CVs that eludicate the maximally shared information among the modalities. In addition, rather than perform tIVA after PCA and whitening is performed, I instead present two new algorithms to allow efficient dimension reduction techniques that also permit subject-level decomposition. These optimization procedures will be presented and discussed along with the traditional optimization procedures of FastICA and Infomax in Section 3.1.3.

Now, I will build the framework of the concatenation procedure for this novel method. Extending this to M modalities, k_i components per modality, and $v = 1, ..., V$ voxels or genes, we can rewrite (3.7) with

$$\mathbf{X}^{[M]}(v) = \mathbf{A}^{[M]}\mathbf{s}^{[M]}(v),\ \ 1 \leq m \leq M,\ \ 1 \leq v \leq V_i \tag{3.13}$$

for N samples (or subjects) s.t. $\mathbf{A}^{[M]} \in \mathcal{R}^{k_i \times k_i}$. In this case, the rows are the sources, leading to the definition of transposed IVA (tIVA) that will be considered in this research. These rows are denoted as the k_i^{th} SCV,

$$\mathbf{s_{k_i}(v)} = [\mathbf{s^1(v_i)}, \mathbf{s^{[2]}(v_i)}, ..., \mathbf{s^{[M]}(v_i)}]^{\mathbf{T}} \in \mathcal{R}^M,$$

where the dependence of the modal-specific features are taken into account by allowing $\mathbf{s}$ to be written as a function of v_i (Adali et al., 2015). This dependence is considered through minimization of the mutual information rate, or entropy rate, r so that

$$\mathcal{I}_r(\mathcal{W}) = \sum_{j=1}^{N} H_r(\mathbf{u}_i) - \sum_{i=1}^{M} \log |\det\Big(\mathbf{W}^{[M]}\Big)| \tag{3.14}$$

$$\rightarrow \mathcal{I}_r(\mathcal{W}) = \sum_{j=1}^{N} \big(\sum_{i=1}^{M} H_r(u_i^{[m]}) - \mathcal{I}_r(\mathbf{u}_n)\big) - \log |\det\Big(\mathbf{W}^{[M]}\Big)|. \tag{3.15}$$

for $\mathcal{W}$ is a block diagonal $MN \times MN$ weight matrix and $H_r(u_i)$ is the unique modal entropy rate evaluated at $H_r(u_i) = \lim_{v_i \to \infty}[H[u_i(1), ..., u_i(v_i)]/v_i]$. For more detailed mathematical descriptions, see (Adali et al., 2015; Michael et al., 2014). Then, fusion takes place by applying the mCCA method described above as a post-processing procedure (J.-h. Lee et al., 2008). That is, rather than reduce dimensionality by taking the data through 4 steps (PCA, whitening, mCCA, then ICA), the data are reduced in only two steps, p-tIVA and mCCA.

3.1.3 Optimization of Multi-Way Decomposition

The MMNG fusion techniques already developed have been instrumental in uncovering novel findings in imaging genetics as well as confirming past knowledge about structural-functional pathways. However, little emphasis has been placed on monitoring statistical reliability when including genetic information. In addition, many of the same optimization techniques have been used regularly, such as Infomax and FastICA, often without testing the stability of the estimates or the use of various modalities. There are theoretical limitations that should be taken into account for the algorithms that are used to derive the weights of the sources that make up the ICs. Therefore, after forming a new structure for MMNG problems through p-tIVA+mCCA, I now present optimization procedures that leverage the efficiency of neural networks with the dimensionality reduction that results from imposing sparsity. In this section, I will introduce the traditional, constrained optimization techniques and then present two algorithms that extend alternative procedures, writing a new structure into MMNG that requires fewer data summarizations. This is in hopes of

preserving the natural complex structure within and between modalities while also creating more computationally tractable procedures.

Constrained Linear and Non-linear

The optimization techniques commonly used in ICA aim to derive estimates of the non-Gaussian property based on higher order statistics to impose independence on the extracted features. Many of these methods were developed when data sets typically had a much lower dimension for $\mathbf{X}$ and for single data sources at a time. Some problems may occur when the input data are high-dimensional and in the presence of multiple sources of big data. I will explain the underlying mathematics for these derivations so that one may understand the properties of the algorithms. Then, the accompanying optimization methods and associated computing techniques will be discussed.

First, independence is a desirable property for data because it becomes tractable to derive joint probability distribution functions (pdf). That is, for two independent densities, say y_1 and y_2, it follows by definition that y_1 and y_2 are independent if and only if the joint pdf is factorizable as $p(y_1, y_2) = p(y_1)p(y_2)$. Although PCA uncorrelates the data, it does not permit independence when the data are non-Gaussian. Take a normally distributed random variable, y. Then, the kurtosis is a measurement of normality, equal to $kurt(y) = E(y^4) - 3[y^2]^2 = 0$ for exactly normal distributions. The drawback to using kurtosis for a non-normal distribution is that the 4th order expectation will be sensitive to outliers. Cleverly, independence is imposed by considering higher-order statistics that are estimates of non-Gaussianity. FastICA and Infomax are algorithms historically used for implementing the independence assumption by using different objective functions for estimation. Optimization of these algorithms usually requires gradient descent or some constrained technique that becomes computationally expensive for big data. Nonetheless, few multi-modal extensions stray from these well-known optimization approaches.

Another representation of non-normality uses an information-theoretic approach to estimate a nongaussianity objective function, called negentropy. The entropy of a random variable (rv) is conceptually defined as the degree of information that the observational variable gives; the more random behavior observed, the higher the entropy (Hyvärinen & Oja, 1999). For the sake of simplicity, let's now denote y as a unimodal, non-normal and continuous rv with values that are centered and whitened. Then, the entropy $H(y)$ with density $f(y)$ and negentropy (or differential entropy), $J(y)$, are

$$H(y) = -\int f(y)\log(y)dy;\;\; J(y) = H(y_{gauss}) - H(y), \tag{3.16}$$

respectively, for $y_{gauss} \sim Normal(0, \Sigma)$, where Σ is the covariance matrix of y. Thus, the negentropy, $J(y)$, is always non-negative, equal to 0 if and only if $y = y_{gauss}$, and is invariant for invertible linear transformations. Thus, we have presented a measure of nongaussianity that does not contain the use of kurtosis directly (Hyvärinen & Oja, 1999).

In order to approximate the negentropy term, FastICA follows the "maximum-entropy principle", which estimates the quantity of $J(y)$ from

$$J(y) \approx \frac{1}{12}E\{y^3\}^2 + \frac{1}{48}kurt(y)^2 \tag{3.17}$$

$$\rightarrow J(y) \approx \sum_{i=l}^{k} k_i[E\{G_i(y)\} - E\{G_i(z)\}]^2, \tag{3.18}$$

where $z \sim Normal(0, 1)$, y is symmetric, and $G(.)$ is a nonquadratic function. Allowing G to take on these common quadratic functions, $G_1(u)$ and $G_2(u)$:

$$G_1(u) = \frac{1}{a_1}\log(cosh)a_1u;\; G_2(u) = -\exp\left(-u^2/2\right) \tag{3.19}$$

for $1 \leq a_1 \leq 2$, then $J(y) \propto [E\{G(y)\} - E\{G(z)\}]^2$. For FastICA, the algorithm uses a fixed-point or approximate Newton iteration procedure to maximize the nongaussianity of the projection matrix/vector, $\mathbf{W}$ (Hyv, 1999). In order to derive this maximization, take the derivative of the quadratic functions in (3.19), so that $g_1(u) = tanh(a_1u),\;\; g_2(u) = u\exp(-u^2/2)$. Then calculate an initial weight vector $\mathbf{w}$, iteratively solving

$$\mathbf{w}^+ = E\{xg(w^Tx)\} - E\{'(w^Tx)\}\mathbf{w};\;\; \mathbf{w} = \mathbf{w}^+/||\mathbf{w}^+||, \tag{3.20}$$

until convergence. Note that ICs may be solved up to a multiplicative sign since $\mathbf{w}$ and $-\mathbf{w}$ are definitively the same direction vector. This is then extended to a weight matrix $\mathbf{W} : N \times K$, as denoted in (3.8), such that the k columns correspond to $\mathbf{w}_1, ..., \mathbf{w}_k$ ICs. Row-wise, $\mathbf{w}_1, ..., \mathbf{w}_n$, are subjective-specific loadings.

Unlike the gradient-based algorithms, the FastICA approach may be applied by selecting a nonlinearity choice for g. The convergence is cubic or, at the very least, quadratic for faster optimization. Components may even be es-

timated one-by-one following a projection pursuit technique that is useful for exploration of the IC quantities (Hyv, 1999). Furthermore, the fixed-point algorithm permits maximum likelihood estimation, a seamless derivation for data sizes of $n > p$. Potential pitfalls to this iterative scheme will be addressed later.

Nongaussian directions may also be solved through one of two closely tied principles: minimization of mutual information and maximum likelihood. For m rvs, $y_1, ..., y_m$, mutual information is defined equivalently to the Kullback-Leibler divergence of the joint density $f(y_1, ...y_m)$ with the product of the marginal densities, $f(y_1) \times f(y_2)... \times f(y_m)$,

$$I(y_1, y_2, ..., y_m) = \sum_{i=1}^{m} H(y_i) - H(\mathbf{y}). \tag{3.21}$$

The interpretation behind mutual information may be explained as the quantity of information gained when switching code length. In particular, the reduction that is obtained by coding the whole vector length (obtained jointly), $H(\mathbf{y})$, instead of coding them separately, i.e. a product of marginals, expresses the maximal amount of information simultaneously gained by using the minimal amount of data (Hyvärinen & Oja, 1999). This is similar to the underlying theory of finding a sufficient statistic. Similar to the previous method, this approach would become null and void under independent rvs, $y_1, ..., y_m$.

Since this is a measure of the dependence between rvs, we can represent the mutual information as an invertible linear transformation, $\mathbf{y} = \mathbf{W}\mathbf{x}$, so that (3.21) becomes

$$I(y_1, y_2, ..., y_m) = \sum_{i} H(y_i) - H(\mathbf{x}) - \log|\det \mathbf{W}|. \tag{3.22}$$

However, when the densities are uncorrelated and of unit variance, as they are after PCA and whitening, then $\mathbf{W}$ must be constant by:

$$E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{W}E\{\mathbf{x}\mathbf{x}^T\}\mathbf{W}^T = \mathbf{I} \tag{3.23}$$

$$\rightarrow \det \mathbf{I} = (\det \mathbf{W})(\det E\{\mathbf{x}\mathbf{x}^T\})(\det \mathbf{W}^T). \tag{3.24}$$

That means the mutual information becomes some constant subtracted by $\sum_i J(y_i)$, and, therefore, the minimization of the mutual information is equivalent to the negentropy approach used in FastICA, shifted by a constant (Hyvärinen & Oja, 1999). Minimization of mutual information is effectively described by maximization of the log-likelihood function, L. The invertible linear trans-

formation $\mathbf{y} = \mathbf{W}\mathbf{x}$ may be written as $p_x(\mathbf{W}\mathbf{x})|\det \mathbf{W}|$ (Bell & Sejnowski, 1989). It follows that the log-likelihood is written as

$$L = \sum_{t=1}^{T}\sum_{i=1}^{n} \log f_i(\mathbf{w}_i\mathbf{x}(t)) + T\log|\det \mathbf{W}| \tag{3.25}$$

for realizations of x, $x(t), t = 1, ..., T$.

Now, take instead each realization of x as the input whose outputs are of the form $g_i(\mathbf{w}_i^T\mathbf{x})$ for non-linear scalar functions g_i. When these non-linearities are the cumulative distribution functions of densities $f_i = g_i'$, (Cardoso, 1997), this is equivalent to the maximum likelihood density above and is known as the Infomax Principle with joint likelihood

$$L_2 = H(g_1(\mathbf{w}_i^T\mathbf{x}), ..., g_n(\mathbf{w}_i^T\mathbf{x})). \tag{3.26}$$

This procedure is often computed using a stochastic gradient descent method that may be slower than FastICA (Bell & Sejnowski, 1989). However, the ease of use and clean underlying maximum likelihood theory are advantages to this algorithm.

Unconstrained Linear and Non-linear Considerations

The iterative procedures required for the optimization of the objective functions using FastICA and Infomax have been introduced. Although these methods have proved useful in single modalities and with rather small to moderate data sizes, problems may become exposed with larger and multiple data sets. Keep in mind that the non-Gaussian estimates derived in ICA involves an orthonormality constraint. That is, the rule imposed on the unmixing matrix, $\mathbf{W}\mathbf{W}^T = I$. For projected gradient descent, this involves the iterative solving of $\mathbf{W} := (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}}\mathbf{W}$. Recall that this matrix contains N rows and K columns corresponding to the desired number of components, effectively smaller than the original data matrix. However, this orthonormalization restricts the dimension of the components so that they may not exceed the number of observations. This failure to allow for overcomplete representation is a limitation, especially when data collection may be expensive and there are limits on the number of subjects we may collect data from. Even more so, when the dimension of the features exceeds the number of subjects (the classic $n << p$ problem), deriving the features only up to the number of subjects could greatly reduce the amount of information we glean from the big data.

Several ICA algorithms, including those mentioned above, begin decomposition with all columns still intact in order to utilize the most amount of data

as possible and then determine K dimensionality as a follow-up. With big data, this manipulation involves factorization of large matrices under the model constraints and becomes a challenging computational problem when done linearly and iteratively. The algorithms may necessarily run for days or weeks, and other issues of convergence can arise. In addition, studying these models becomes a difficult task for anyone even remotely removed from a solid programming background. Given the large amounts of data and the GPU power needed to optimize the objective functions, these constrained optimization procedures must be programmed and run on a computational cluster to allow for the big data. Along with the goal of extending these approaches to a multi-modal setting comes the necessity for multi-disciplinary collaborations. The algorithms must be able to handle large data and be integrated into a pipeline that is approachable for those outside of a data science realm. Furthermore, future studies must link this pipeline to the clinical process of diagnosis, reiterating the need for generalizable, practical, and interpretable protocols.

In addition, it seems counterintuitive to necessitate pre-processing of the features with PCA, when the argument behind ICA is that PCA is not a valid approach for deviations from normality. After PCA occurs and the resulting dimension is sorted by maximal variance, whitening is then performed to fully normalize the data. While this effectively reduces the dimension prior to IC decomposition, it may result in the loss of too much information. However, this has become standard practice to handle the $n << p$ problem and ease of computation of the ICs and is then usually followed by one of the traditional optimization techniques. Unfortunately, this pre-processing is unavoidable when taking the mutual information or maximal negentropy approaches. The data are, therefore, sensitive to these dimension reduction techniques and may result in overestimation due to the multiple steps of this process. Following decomposition and back reconstruction, the data have been pushed through 4-5 stages of reduction and summarization. How much information are we losing through this process? Given any one of these issues mentioned, further algorithms should be explored so that efficient and more robust decomposition may be performed.

First, in order to create a more robust optimization technique, I propose a replacement for the orthonormality constraint with a soft reconstruction penalty term. This approach has been referred to as Reconstruction ICA (rICA) (Guillén, 2017; Le et al., 2011), where the replacement allows for an overcomplete basis in the dimension reduction procedure and is equivalent to the cost function in sparse coding and autoencoders (Bruno & David, 1996; Olshausen & Fieldt, 1997; Vincent, 2010). With high-dimensional data, especially neuroimaging and

genetic data, we are attempting to extract the most biologically relevant features among a deep pool of heterogeneous connections among the features. There may be many more redundant features throughout the raw data than there are useful features. This idea of sparsity has been used in previous statistical models through the introduction of shrinkage terms, or Lasso, procedures that include a penalty term in order to push the redundant features towards 0 and highlight the meaningful vectors. Several ICA models have included this idea of sparsity through sparse reduced-rank regression, Bayesian group-sparse multi-task regression, and hybrid models with a SVM step for classification (Greenlaw et al., 2017; Hao et al., 2020b; Sun et al., 2018; Vounou et al., 2010), respectively. Nonetheless, this has not been formalized for the combination of neuroimaging and genetic data. Therefore, as a second optimization novelty, I propose the use of an L_2 norm sparsity function and refer to this optimization procedure as sparse filtering ICA (sfICA).

Now, I will formulate the mathematics of this orthonormality constraint in the context of $i = 1, ..., M$ modalties with $k = 1, ..., K$ extracted features for $j = 1, ..., N$ subjects. Recall $g(.)$ is a non-linear convex function. For a single modality, the unlabeled data are $\{x^{(i)}\}_{j=1}^{N}$ for $\{x^{(i)}\} \in R^N$. Then the optimization problem is given by

$$\min\{W^{(m)}\} = \sum_{i=1}^{M}\sum_{j=1}^{N} g(W_j)x_i^{(M)} \text{ subject to } WW^T = \mathbf{I}. \qquad (3.27)$$

The orthonormality constraint in (3.27) prevents the bases in the unmixing matrix W from becoming degenerate and is the key element in ICA optimization for $N < d$, thereby restricting the optimal number of components so that $K \leq n$. It is when $n << p$ of the feature space exceeds the samples (or subjects) that the standard optimization techniques require PCA pre-processing because (3.27) no longer holds. Even the fixed-point iterative method presented in FastICA may become slow and require fine-tuning for each new data set. This makes generalizations to various data sets within the same modalities a challenge, because the local minima that are solved change with each run of ICA and each bootstrap "shuffle" of the data.

Sparse encoding, on the other hand, includes a penalty term in the cost function that adjusts for sparsity in the complex data source. Take the minimization function $\mathcal{A}$ as the sum of the negentropy term subtracted by the product of the sparsity parameter λ with a measurement a_i that accounts for the sparsity distribution among the linear superposition of basis functions $\phi_i(x)i^{(m)}$). The

maximization of negentropy or equivalent minimization of mutual information may be expressed as in (3.22), then the estimate of the sparse encoder is

$$\mathcal{A} = -\sum_{x,y}[I(\mathbf{x},\mathbf{y}) - \sum_{i} a_i\phi_i(x,y)]^2 \tag{3.28}$$

Then, it is clear that the objective function is solved with respect to a_i while ϕ_i uses gradient descent averaged over each iteration of the convergence criteria (Bruno & David, 1996). This involves a local network algorithm that produces many possible local minimums, thereby allowing the data to drive the optimization. As data-driven and unsupervised procedures are useful when labels are absent from the data, this method allows the finding of an overcomplete basis and a non-orthogonal restriction in the process.

Sparse encoding is comparable to auto-encoders that are utilized in the modern neural networks within a deep learning framework. Under the input-output system of DL, the hidden layers introduced in Chapter 1 represent linkages and pathways among the biological networks that underscore psychiatric illnesses. An auto-encoder has two parts to this procedure: the encoder and the decoder. The auto-encoder takes the outputs and reconstructs the data by backpropagation. This compresses or denoises the data through an encode-decode expectation $E[||\text{decode}(\text{encode}(\tilde{x}))||^2]$, where decode(encode$(\tilde{x})$) is written from (3.28) for coding sparsity function $S(.)$ as

$$\min_{a} = \sum_{j} ||\mathbf{x}_j^{(m)} - \sum_{i} a_i^{(m)}\phi_i||^2 + \lambda \sum_{i} S(a_i^{(m)}) \tag{3.29}$$

where the first term is the reconstruction term that permits the new basis and the second term is the penalty that imposes sparsity (Guillén, 2017). Written for the adjustment of the orthonormality constraint, the formal definition of the rICA objective function is then:

$$\min_{W} \frac{\lambda}{N} \sum_{i=1}^{m} ||W^T W x_i^{(m)} - x_i^{(m)}||_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{d} g(W_j x_i^{(m)}), \tag{3.30}$$

for the encoding and decoding steps, $Wx_i^{(m)}$ and $W^T W x_i^{(m)}$, respectively, and nonlinear function $g(.)$.

The algorithm permits the use of an optimization technique known as the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS). This is a quasi-Newton optimization approach that is unconstrained, that is, not requiring a gradient descent or fixed-point algorithm, first introduced by (Fletcher, 2013).

Such optimization approaches are implemented regularly in machine learning models, thereby making the connection between the reconstruction penalty term present in neural networks with deep learning algorithms (Rafati & Marica, 2020). The idea behind this technique is to take up to a certain limit of iterations until the norm is less than a tolerance threshold, or the norm of the gradient at the iteration is less than a gradient tolerance times a scalar,

$$\tau = \max(1, min(|f|, ||g_0||_\infty)), \tag{3.31}$$

where $|f|$ is the norm of the objective function, and $||g_0||_\infty$ is the infinity norm of the initial gradient (Nocedal & Wright, 2006). That is, under rICA, the nonlinear convex function introduced in equation (3.30) is evaluated at the product of the unmixing matrix with the original data ($X * W$). This is to make the element-wise transformation to the data, achieving a "nearly orthonormal" weight matrix that minimizes the cost function, maps the data back to the inputs, then pulls the extracted features. (Nocedal & Wright, 2006).

While rICA beautifully formulates an unconstrained optimization approach within the ICA framework, it does so while retaining a linear formulation of the data transformation. Recent work has explained that linear restrictions in high-dimensional and complex biological data may not always be relevant to real-life data. One may, instead, seek to map the linear space to a non-linear one, in order to allow more flexibility in the feature extraction. This may be done by expanding a sparsity term to the optimization problem. In other words, we may allow both an unconstrained optimization approach that sifts through the data to filter out the sparse elements without a linear restraint. Sparsity may be induced by allowing $g(.)$ to become $g(u) = \sqrt{u^2 + 10^{-8}}$, a $L2$ sparsity term that is applied element-wise to $X * W$, in order to obtain the matrix F. Note that under the nonlinear, sparse function, $g(.)$ must be a smooth nonnegative symmetric function close to the absolute value function. In other words, this is applying an L^2 norm (Ngiam et al., 2011). The resulting column-wise normalized matrix is given by

$$\widetilde{F}(i,j) = F(i,j)/||F(j)||; \;\; ||F(j)|| = \sqrt{\sum_i = 1^N (F(i,j))^2 + 10^{-8}}. \tag{3.32}$$

Similarly, the rows are then normalized by the L^2 norm, resulting in the matrix of converted features, denoted $\hat{F}(i,j)$. A shrinkage that may be eluded more to a filtering process is considered with an additional λ term. Then, the objective function of the sparse filtering ICA (sfICA) is given by

$$h(W) = \sum_{j=1}^{k} \sum_{i=1}^{N} \hat{F}(i,j) + \lambda \sum_{j=1}^{k} w_j^T w. \tag{3.33}$$

Four optimization algorithms with different cost functions have been introduced that are either constrained or unconstrained and linear or non-linear. Novel to multi-modal research is the addition of both rICA and sfICA. These methods utilize the L-BFGS quasi-Newton optimizer, while rICA assumes a linear space and sfICA is nonlinear. Similarly, the traditional and most popular algorithm techniques are FastICA, a linear, fixed-pointed approach and Infomax, a nonlinear gradient descent approach. While rICA and sfICA have been used for single data sets, they have not been applied to neuroimaging or genetics data or other high-dimensional settings. Nor have these approaches been written into a multi-modal fusion framework. Similarly, the idea of sparse-autoencoders has not previously been extended to the decomposition framework (Guillén, 2017; Le et al., 2011), nor have these ideas been applied to multiple modalities. Thus, in the new algorithms, I expand the idea of reconstruction penalty terms and L^2 norm sparsity terms in rICA and sfICA, respectively, as the introduction of novel approaches scaled to MMNG problems. Table 3.1 lists the traditional algorithms in the "constrained" column and the newly proposed algorithms as "unconstrained". The inclusion of both a linear and nonlinear approach will make these new methods comparable to the FastICA and Infomax algorithms, respectively.

Table 3.1: ICA optimizations considered

	Constrained	**Unconstrained**
Linear	FastICA	rICA
Nonlinear	Infomax	sfICA

3.2 Multi-way Simulation Setting

Before applying the multi-way models to real-life AD data, it is important to test the properties of the algorithms using unbiased data, or data that are unrelated to the outcome variable, AD diagnosis status. In this section, a multi-way simulation stage is set up that incorporates simulated functional and structural brain data along with simulated SNP data. The assessment for algorithmic reliability and statistical stability of multi-modal pipelines for all possible combinations of modalities are discussed and will be carried out in Chapter 4.

3.2.1 Data for Testing

After a thorough literature search, it appears that past researchers have not simulated the multi-modal scenario specifically for imaging genetics problems utilizing 3 or more modalities. Although, this has been done for strictly neuroimaging settings (Duan et al., 2020; Groves et al., 2011; Qi et al., 2019) and imaging genetics problems of a smaller scale (J. Liu et al., 2009; Silver et al., 2011). In order to do this, I have simulated sMRI, rs-fMRI and SNP data of varying sample sizes (20, 50, 100, 300). This is a novel simulation setting that will help researchers understand the statistical implications that decomposition and deep learning optimization will have on the study outcomes.

It is not a trivial task to simulate data in a MMNG setting, because there are unique data collection processes for each modality. For example, the spatial noise present in structural MRI data as well as the temporal correlation among fMRI data must identify with "natural" procedures within the scanners. Given the extensive genome collection available that are representative of large populations, SNP data are the least cumbersome. Simulations are generated for SNP, sMRI, and rs-fMRI data for samples with 20, 100 and 300 subjects for 567 SNPs, 256x256 and 704x704 voxels for each modality, respectively.

rsfMRI Data

The simulation of fMRI is not new in the literature (Welvaert & Rosseel, 2014). However, fewer studies have considered the simulation of resting-state fMRI. The reason for this is that task-based fMRI analyses have been studied for decades, and resting-state is a more recent study design. However, resting-state fMRI (rsfMRI) has already made leaps and bounds of progress in imaging research, especially related to neuropsychological illnesses. Resting state fMRI data gathers slices of the brain over time where the blocking design is simply ignored, represented a task-free fMRI design. There are two methods which have allowed me to simulate the rsfMRI data, and I initially tested the data using both procedures.

The first simulation method involves a Matlab toolbox, called SimTB, that simulates functional data and can construct a blocking scenario for event-based inference. This method has been used widely for testing of multi-modal imaging methods(Allen et al., 2012; Duan et al., 2020; Michael et al., 2014; Welvaert & Rosseel, 2014). The toolbox allows one to simulate data from specific regions within the brain by selecting components from a spatial map (SM) and adjusting for location, rotation and size parameters. The grid size of the voxels may be chosen for the slices, and I proceeded to derive a 220x220 grid to remain

consistent with the dimensions of the ADNI rsfMRI data. Thus, the simulated data are representative of a full slice of the brain for the "subject".

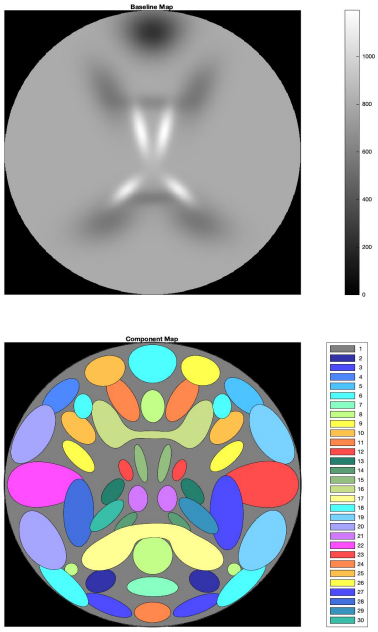


Figure 3.3: Baseline and spatial maps for SimTB.

The maps are selected from the SM sources under the assumption of spatiotemporal separability, meaning the data are expressed as the product of the time courses (TC) and SMs (Erhardt et al., 2012). This is the basis of the simulation, which constructs a linear combination of amplitude-scaled and baseline-shifted TCs and SMs, denoted $\mathbf{Y_i}$, for subjects $i = 1, ..., n_i$ for $i = 20, 50, 100, 300$. The baseline intensity is the gray image in Figure 3.3. The plot below the baseline image in shows the spatial maps. Each color symbolizes one component, or region, in the brain, with 30 in total. For each experimental simulation, the components desired are selected. Then, one "draw" is taken per subject to represent one brain slice per individual. Each simulated observation allows variability by applying noise appropriate to fMRI data. The temporal noise, as is evident in fMRI data, is then included into the construction by specifying a contrast-to-noise ratio (CNR) that is consistent with Rician noise. For more information on this process, see (Erhardt et al., 2012).

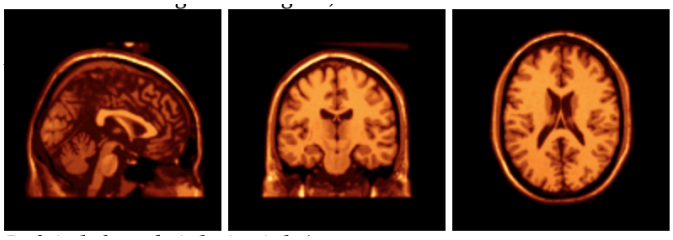
The second method used to simulate rs-fMRI is generated from the R package neuRosim, which provides simple simulation scenarios as well as 4D MRI strategies (Welvaert & Rosseel, 2014). These simulations lean on the idea that fMRI signals are the outcome from a Fourier transformation of k-space, re-

sulting in complex-valued magnitudes (Lindquist, 2008). With neuRosim, the authors break up the simulation process into two steps: the design of the study (be it resting or task-based) that yields the activation, and the necessary temporal noise. Unlike the SimTB method, this one begins with the hemodynamic response function (HRF) which the user may choose based on the distribution provided, gamma, double-gamma, or the balloon model. Instead of overlaying this as a linkage to brain regions, the dimensions of the model act as the template of a matrix from which the simulated HMF is mixed with the selected CNR structure. Options of the noise relation include an autoregression correlation, Gaussian random field and Gamma random field. Once this is done, the time series is then developed and the matrix is "filled i". For more details on this procedure, see (Welvaert & Rosseel, 2014).

sMRI Data

The simulation of voxel-wise, structural brain data proved to be even more of a challenge, as it was difficult to uncover a standard method within current literature. Several options presented themselves from coreMRI, BrainWeb, and from SimTB. Although the latter was designed for fMRI simulation, one may adjust this for structural data by extracting only one TC and then adjusting the noise based on "typical" volume structure using Gaussian noise (R. Y. Zhang et al., 2020), in theory. However, the program would not allow me to extract just one time point. On the other hand, coreMRI and BrainWeb utilize MRI data phantom scans with added noise to simulate the voxel-wise data on a 256x256 Gradient Echo grid (Kwan et al., 1999; Xanthis & Aletras, 2019). Both approaches worked fine, but can only produce one subject at a time.

Figure 3.4: The T1-weighted phantom image used in the BrainWeb MR Simulator.



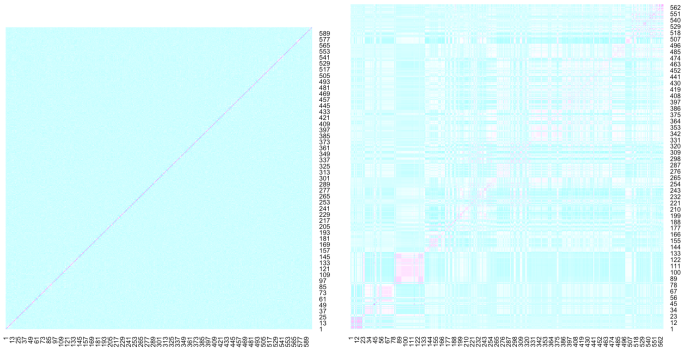
Due to ease of access and processing speed, the best choice was to do one of two things: simulate MRI grey matter data from the BrainWeb database or randomly select MRI data from healthy individuals in ADNI or even a UKBiobank. (Sudlow et al., 2015). I chose not to use the MRI data from healthy individuals, because the sample would still be biased by age and volunteer bias. Literature

suggests that MRI simulations are most helpful when considering two different forms of spatial noise construction. Using the BrainWeb, the phantom scans are pulled from the online database assuming a set of T1-weighted data with a SFLASH (spoiled FLASH) sequence of TR=22ms, TE=9.2ms, and flip angle=30 degrees. Therefore, I pulled this data in Matlab, converted it to 3D matrices and then resampled within the data and added CNR Gaussian noise for i observations. Each resampling is representative of a patient in the study.

Genetic Data

The SNP data are gathered using the R package sim1000G, developed for simulation of SNPs for unrelated individuals and family-based designs (Dimitromanolakis et al., 2019). Due to the expansive knowledge of the human genome, and the ease and accuracy with which genetic imputation are performed, it is a rather straightforward approach to simulate SNP data when compared with imaging data (Howie et al., 2009). Thus, the data are pulled from the well-known HapMap Phase II genetic map from build 35 to GRCh37, an extension of the HapMap project. Recombination of the genetic encoding is then built into the algorithm that more-or-less shuffles through the SNP placements randomly but realistically. All of this is computed using a 2GHz processer and 4GB of RAM for extremely fast computation. It is important to note that the standard allele frequency distribution and LD structure within the genetic pathways are preserved for this simulation procedure (Dimitromanolakis et al., 2019). In Figure 3.5, we can see a heat map of the correlations derived from SNP data on healthy ADNI1 participants (left) after undergoing quality control (QC), and on the right we output the same for the simulated SNP data without any processing. Maintaining the natural correlation structure allows one to test how this impacts the results.

Figure 3.5: Heatmap of Linkage Disequilibrium for ADNI1 vs Simulated SNPs



When ICA is performed on SNP data alone, one must recall the properties of the data. The data are extracted from a DNA encoding in the form of letters, A, C, G or T, the most granular building block of the human genome. The conversion to 0, 1, or 2 is representative of the minor/major alleles present in the genetic encoding. However, in order to input into an ICA model, the data are scaled and taken as continuous variables, which is performed at the pre-processing step of PCA prior to the ICA runs. This plus the sheer volume and miniscule granularity of SNPs has brought up some controversy in the literature for whether ICA is warranted and or if other avenues should be explored (Kim & Lee, 2020). If prior information is known so that we may target specific genetic pathways, this apriori information may be used in reference ICA (Chen et al., 2014; J. Liu et al., 2012). Others have suggested that reducing the dimension of SNPs by first performing a GWAS also shows useful results when using ICA (Soheili-Nezhad et al., 2021). Due to the fact that this is an exploratory analysis of new methods, and given the pre-processing that has been performed on the SNPs, I proceed with including the SNPs at the most granular level. The question of interest is how the current ICA approaches may be expanded to allow genetic data in addition to the imaging data. Due to this goal, the simulations will seek to determine whether the new methods may improve the reliability of the cluster estimates. The following multi-modal sections will then analyze SNP data in conjunction with imaging data to see if these results scale well to the MMNG framework.

3.2.2 Simulation Measures

When discussing the fusion processes and the optimization algorithms that will be used, one may recognize a common theme among the ICA algorithms. ICA extracts components and assigns weights that make up the elements of the components, and these values are derived by optimizing one of the various cost functions discussed. However, the computation process results in the iterative solving of many local minima rather than one solution. Since there is not one exact solution, or global minima, for this data driven approach, one will receive different estimates for the values of the ICs for multiple runs of ICA even when analyzing the same data set (Himberg & Hyvxrinen, 2003). Given the various results that may be obtained, one major question researchers have about the validity of ICA is whether or not the subject-specific weights are reliable enough for downstream analyses. If results change within the exact same data set, how much variation will exist among different data sets? Does this mean the results of ICA and inconsistent and possibly change the overall outcome of the analysis? While ICA, in theory, preserves interpretation by the weights assigned

to the subject- and feature-specific elements of the ICs, how much does this variation change the conclusion? To date, current research fails to compare the widely used ICA algorithms with unconstrained approaches. In addition, the stability within imaging genetics problems need more exploration. Thus far, stability is mostly an idea of single modalities or, in a few cases, for the combination of imaging modalities. As more and more researchers are utilizing data from multiple sources, it is important to assess how the complexity of the joint information attributes to the findings.

ICASSO

In order to measure the stability or reliability of the IC estimates, one may observe how the results change when multiple IC runs are performed on the same data set. After obtaining a different set of components for each run, one may then perform clustering for each of the vector-wise components across the various runs to observe how similar the outcomes are. That is, collect all of the first components for each of the n runs, then perform a cluster analysis on the values. Then, repeat for all k-1 components. Note that this is performed after the optimal number of ICs has already been extracted. One measure of similarity is to quantify the cluster separability within ICs by deriving a compactness index, denoted I_q. This index provides a number between 0 and 1 that provides the average intra-cluster similarity subtracted by the average extra-cluster similarity measured. This well-known procedure, referred to as ICASSO (Himberg et al., 2004; Himberg & Hyvxrinen, 2003), additionally provides a visualization technique for assessing the reliability of the ICA procedure. The idea is that the more compact the clusters are after multiple ICA runs, or the closer I_q is to 1, the more reliable the estimates are.

The mathematical formulation of I_q builds on the statistical concept of between and within variances, where now we are looking to derive the difference between the average distance within each cluster and the average distance between the clusters. Keeping in mind that each component source will have it's own cluster, this means we want there to be low separability within the clusters and higher separability between the clusters so that they are non-overlapping and distinct. In essence, this will be found by taking the absolute value of the mutual correlation coefficients of the h^{th} ICA run and the i^{th} subject, r_{hi}, yielding similarity matrix $\sigma_{hi} = |r_{hi}|$. This may then be transformed into a dissimilarity matrix by subtracting by 1, $d_{hi} = 1 - \sigma_{hi}$. Next, consider the disjoint set of all estimates gathered from the K components, $\bigcup_{k=1}^{K} C_k$. The clusters are selected by agglomerative hierarchical clustering so that the highest level of the "tree" make up the clusters, but this allows the researcher to explore the

smaller subsets of clusters, or the "branches". The agglomeration link strategy for ICASSO uses a group average link, and the following formula will yield I_q (Himberg & Hyvxrinen, 2003):

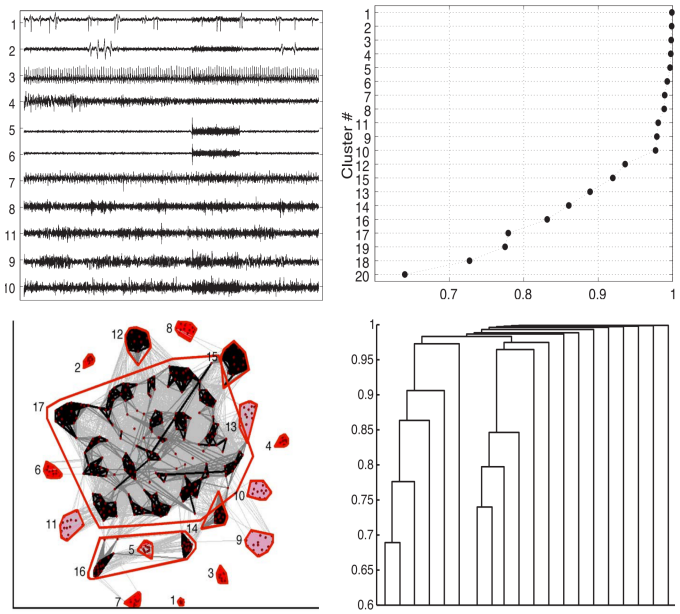
$$I_q(C_k) = \frac{1}{|C_k|^2} = \sum_{hi \in c_k} \sigma_{hi} - \frac{1}{|C_k||C_{k'}|} \sum_{h \in C_k} \sum_{i \in C_{k'}} \sigma_{hi} \tag{3.34}$$

ICASSO runs multiple ICAs on the same data set, but it may also be used across data sets to see if the findings are consistently replicated. In addition, one may bootstrap within a data set for each new ICA run or randomly draw the initial values at each run. Changing the initial values for various runs analyzes the algorithmic stability of the ICA approach, while bootstrapping from the data on each rerun or changing the data set altogether reveals the statistical reliability. In addition, this tells us a lot about the generalizability of the ICA model. If the cluster compactness is much lower than 1, than the validity of the IC estimates may be in question. The centrotypes of the clusters may then be selected as the final ICA elements of the K components. ICASSO is a strong tool for assessing the reliability and compactness, because these metrics may easily be visualized. Estimated sources corresponding to the centrotypes of the clusters may be visualized, a graph of the stability index for the number of clusters, a similarity graph of the clusters will display the distances within and between clusters and a hierarchical dendogram, all seen below in Figure 3.6, respectively , using the Matlab program *icasso* (Hyvärinen & Oja, 1999)).

Minimum Description Length

Furthermore, assessing component compactness may additionally provide a method for sorting and selecting the number of components chosen in an ICA model. PCA has the advantage of providing a relatively straightforward approach for selecting the number of components once the maximal or desirable variance is explained within the data. Several methods have been suggested in past research, such as the Akaike Information Criterion (AIC) or the Kullback Information Criterion (KIC). However, the information theoretic criterion are based on a likelihood derivation that assumes identically distributed random variables (iid), which is most certainly not the case with MMNG data (Y. O. Li et al., 2007). Another approach that has been widely accepted is to derive a renewed version of minimum description length (MDL) (Calhoun et al., 2001; Grünwald, 2019). Originally defined as a likelihood method as well, the derivation with respect to the number of components, k, is given by

Figure 3.6: Visualizations extracted from the ICASSO procedure in Matlab, (Himberg et al., 2004).



$$MDL(k) = -V(NT-k)\mathcal{L}(\hat{\theta}_k) + \frac{1}{2}(1 + kT + \frac{1}{2}(k-1))\ln(V), \quad (3.35)$$

where V is the number of voxels/SNPs, N is the number of subjects, T is the number of time-points for fMRI data and equal to 1 otherwise. $\mathcal{L}(\hat{\theta}_k)$ is the ratio of the geometric mean of the $LN - k$ smallest PCA eigenvalues to their arithmetic mean. The mathematical form is given by

$$\mathcal{L}(\hat{\theta}_k) = \ln \frac{(\lambda_{k+1} \cdot \ldots \cdot \lambda_{TN})^{TN-k}}{\frac{1}{TN-k}(\lambda_{k+1} + \ldots + \lambda_{TN})}. \quad (3.36)$$

One can see that the formulation above is applicable to neuroimaging or genetics data, and the same concept may be extended to cognitive measures, or any other modalities that may be considered. MDL is estimated for each modality separately for ICA decompositions. However, for jICA, the same number of components must be selected for each modality. In order to strengthen the case for the number of components selected, this may be coupled with ICASSO. The number of components may be derived using MDL, and then the K components may be sorted in descending order of the I_q values. Alternatively, one

may pre-select an acceptable cluster separability and not allow components less than this number. In this work, I select the latter method, because the goal is to assess how stable the model is, and this would eliminate any of the unstable findings. In addition, MDL is the standard approach used in many of the past MMNG models, and I would like to have a consistent comparison for these methods when using unconstrained optimization.

Simulation Plan

There are three parts to the simulation plan that I will perform in Chapter 4. First, I will consider the simulated SNP, sMRI and rsfMRI and test the algorithmic stability and statistical reliability using ICASSO. The maximum, mean and standard deviation of I_q will be compared for sample sizes of 20, 50, 100 and 300, comparing the cluster compactness over multiple runs of all four algorithms: FastICA, Infomax, rICA, and sfICA. In addition to examining the stability of the algorithms, using simulated modal-specific data allows us to test the extracted sources with the ground truth of the modalities. The measurements I use are two common approaches for testing recovery of signals for BSS, the minimum distance (MD) index and the Amari Error (AE) (Miettinen et al., 2017). Both measures provide values between $(0, 1)$, with lower values of MD and AE as indicators of good separation performance. MD has an interesting asymptotic property related to the variance of the estimation for the unmixing matrix. For ICA, since $\hat{W} \rightarrow (A)^{-1}$, it follows that $\sqrt{n}\text{vec}(\hat{W}A - I_i) \rightarrow N_{i^2}(0, \Sigma)$ (Nordhausen et al., 2011). On the other hand, AE measures the average mutual information between the sources and the ground truth. Let $\hat{G} = \hat{W}A$, the elements of which are $\{g_{ij}\}$, and $\mathcal{C} = \{C \text{ each row and column of C has exactly one non-zero element}\}$, then the two separation estimates are given as follows, for $i = j = 1, ..., K$ number of extracted components:

$$MD(\hat{G}) = \frac{1}{\sqrt{K-1}} \inf ||C\hat{G} - I_K||, \tag{3.37}$$

$$AE(\hat{G}) = \frac{1}{2K(K-1)} \left(\sum_{i=1}^{K} \left(\sum_{j=1}^{K} \frac{|\hat{g}_{ij}|}{\max_h |\hat{g}_{ih}|} - 1 \right) + \sum_{j=1}^{K} \left(\sum_{i=1}^{K} \frac{\hat{g}_{ij}}{\max_h |\hat{g}_{hj}|} - 1 \right) \right) \tag{3.38}$$

Next, I will compare the algorithms for single data sources pulled from ADNI data. The modalities of sMRI, FDG-PET and rsfMRI will be considered. This is to compare three very different types of neuroimaging data against the four optimization techniques: structural MRI, functional summarizations of FDG-PET, and time-

varying rsfMRI. The dimensions vary from each modality, as seen in Table 3.2. In addition, with the application of ICA to each modality, I pull brain images of the components and discuss disease-specific interpretations that may be gleaned from these images. Finally, I perform MMNG data fusion under two frameworks and four algorithms. The common fusion structure of mCCA+jICA will be compared with the novel fusion framework of p-tIVA+mCCA for FastICA, Infomax, rICA and sfICA optimization techniques. Three multi-modal data combinations will be considered SNP+sMRI, sMRI+FDG and sMRI+FDG+SNP. Group comparisons will be made using the components across and within modalities with visualization aids. In addition, the overall statistical replicability is tested through simultaneous random iterations of the initialization parameters and bootstrapping of the data for all of the ICA approaches.

3.3 Imaging Genetics Data Fusion Pipeline

The most time-consuming effort of this multi-modal data analysis was not the method development itself, although this was certainly a challenging endeavor. Any imaging genetics problem begins with a neuropsychiatric disease that is best expressed by multiple sources of data, often biological. This may include blood biomarkers, brain imaging signals, genetic material, regulatory diagnostic measures of the human body, or even clinical measures such as memory, cognition and intelligence. Even prior to analysis, other challenges include downloading the information, manipulation and storage of the high-dimensional data, and pre-processing of the various modalities. Then, as analysis begins, one must remain organized in keeping track of the files, data, and output, and the researcher must have a clear path forward for the fusion procedure. As a data scientist, one of the biggest contributions one may make to the field of imaging genetics is to provide a generalizable pipeline that will aid other researchers in similar analysis pursuits in the future. Assuring replicability is vital to move scientific efforts forward in this big data realm. Unfortunately, one studying imaging genetics will necessarily face issues with open science as proper reproducible procedures are not always in place. In this section, I will outline the process that I followed in order to perform MMNG fusion with emphasis on the importance of participating in and contributing to open science: collection of data sources, data pre-processing, data organization and manipulation, computational tools, interpretation and visualization.

3.3.1 Collection of Data Sources

Prior to collection of the data sources, one must begin with a thorough literature review focusing on at least three areas: underlying science/biology of the disease, clinical questions that remain unanswered, and methodology that have already been applied to this area. Once a solid scientific question is established, one must identify or select

the modalities of interest. In part, this may be selected by the data sources available. However, it is important to create a question that is as biologically meaningful as possible in the field of interest. After the data sources have been selected, be it ADNI, UKBiobank, Enigma, Neuromark, etc, there are qualifications that must be passed in order to gain access to the data. For ADNI, I filled out a Data Usage Agreement that required an explanation for why I need the data. Each year I renew this approval. The researcher has an ethical obligation not to violate this agreement or to share the data with anyone else outside of the restrictions provided therein. Prior to download, one should become familiar with the study design and understand the background effort. There are likely going to be problems with data download or locating the proper data sources. With the ADNI database containing such a large quantity and variety of information, I found it very challenging to find the data I need and to understand the variable names. In addition, much of the information guides on the ADNI website were out of date and did not include instructions for the more recent protocol, ADNI3. Fortunately, ADNI provided two resources for asking questions, both which I found very helpful, providing quick and resourceful answers. One may contact Ask the Experts (http://adni.loni.usc.edu/support/experts-knowledge-base/ask-experts/) to ask a pointed question for one of the many Cores, such as the MRI, PET, Clinical, Data Sharing and Publications, Biostatistics, and more.

Once the organization of the data are well understood, the researcher will need to take into consideration the sheer volume and noisiness of the data. The complexity of the imaging data is best understood with a description of how the data are obtained. For an example, consider how fMRI data are collected on one individual. Upon entering an MRI machine, the patient must hold still for the course of the experiment, usually 30 minutes. Solid positioning is vital to ensure equivalent images are taken across time points. The MRI creates snapshots of the brain over time from various directions or imaging planes called slices. These slices are further broken up into over 4000 voxels (on a 64×64 grid), or volume elements (Lindquist, 2008). At each voxel, the measured data are the MR signal as it evolves over the time course. Thus, the data on one subject consist of potentially hundreds of thousands of time series, with hundreds of time points on each series (Lazar, 2008). Multiply this data by the number of subjects in the study and the number of measurements for longitudinal studies, and the issue magnifies. The data collected on each subject are then pre-processed to reduce noise that would otherwise prevent an accurate analysis. Noise, or alterations, in the signal may be induced from natural fluctuations in the magnetic signal, subject-related motion, environmental background noise and more (Lazar, 2008). To complicate matters, MRI data are highly correlated spatially and temporally. Temporal correlation occurs when the strength of the signal at time point t is affected by the signal at the earlier time point, $t - 1$, inflating the within-subject variation. On the other hand, spatial correlation occurs between subjects due to the heterogeneity in brain structure and function voxel-by-voxel (Lindquist, 2008). Noisy and highly-correlated neuroimaging data mandate pre-processing that can vary from study to study. Consequently, the

data can take many distributional continuous forms that may alter the results obtained if identical processing steps are not taken from researcher to researcher (Lazar, 2008).

Contrast this with human genetic material, which are not continuous and are considered count-data as the measurements taken are the number of allele frequencies per SNP of a gene. These data are usually expressed by possible values of (0, 1, 2) per SNP to indicate the homozygotic major, heterzygotic, or the homozygotic minor allele frequencies, respectively (Nathoo et al., 2019). Instead of collecting hundreds of thousands of data from individuals across multiple patient visits, the genetic data are obtained only once at the baseline measurement. A DNA sample is drawn by one 10 mL EDTA tube of whole blood and is shipped to a gene repository for genotyping (Saykin et al., 2015b). For ADNI1, they used Illumina BeadStudio 3.2 software to generate SNP genotypes from bead intensity data. After performing sample verification and quality control bioinformatics, the genotype data for 818 ADNI participants was made available to qualified researchers. Reprocessed array data was made available in 2010 using an updated version of BeadStudio, Illumina GenomeStudio v2009.1, for all 818 samples. The genomic DNA is extracted and sent to the geneticists responsible for sample verification and quality control. Each subject may have as many as 100-500k SNPs in a genome-wide analysis; multiply this by the number of subjects in a study and one can understand why genetic material also presents a big data problem.

The ADNI sMRI data are collected from over 60 sites and then sent to the MRI Core for MPRAGE pre-processing. Although the raw images are provided by the MRI Core, I chose to use the standardized data sets for the analyses so that both new and confirmatory findings may not be skewed or otherwise attributed to differences in processing (Wyman et al., 2013). The MRI acquisition plane is Sagittal with 3D acquisition, a PA coil, with field strength of 1.5 tesla and flip angle $8°$. The manufacturer of the scanner is Siemens, with T1 weighting and IR/GR pulse sequence of the symphony model. The matrix dimensions (in pixels) are $(X, Y, Z) = (192, 192, 160)$, with pixel spacing 1.3 mm and slice thickness 1.2mm. Other parameters are: TE=3.6ms, TI=1000.0ms, and TR=3000ms. The manufacturer of FDG-PET is Siemens, model 1093, and the radioisotope is F-18 with 18F-FDG radiopharmaceutical injection. The matrix dimensions per subject are $(X, Y, Z) = (336, 336, 81)$ with pixel spacing $(1.0)^2$mm, with 2.0mm slice thickness. The rsfMRI data are collected from the ADNI2 protocol, as this was not a known method during the time of data collection in ADNI1/GO. This information is processed by the Mayo Lab, the MRI Core. The manufacturer of the scanner is the Philips Medical Systems with 3.0 tesla field strength and flip angle 80°. The matrix size per slice is 64×64 with 3.3mm pixel spacing in both X and Y directions, 3.3mm slice thickness and 6720 slices per subject. The mfg model is Achieva with pulse sequence GR, TE=30 ms and TR=3000 ms (M. W. Weiner et al., 2015; M. W. Weiner, Veitch, Aisen, Beckett, Nigel, et al., 2017; M. Weiner & ADNI, 2013).

Due to the sheer volume of the data, downloading the data may prove to be very time-consuming, as some of the files may be very large. It is important to check the space on your own personal computer first and to have a plan for where the data will

be stored. As an example, I found that some of the imaging modalities could produce data as large as 100+GB, especially when looking at the modality across time points, by voxel, disease category, etc. When trying to download such a large volume of data, the computer could time-out or not have the RAM to complete the download. I found that breaking the data up into chunks by disease status, ADNI protocol and modality helped me to more efficiently run the data. I downloaded the data directly to my laptop, then stored the data on an external server. This would clear up the space I used for that data, then I could download the next chunk, and so on. Those who have very limited space on their personal or work computer may look into downloading from the data website or source itself directly into a cloud space. In addition, I highly recommend establishing a back-up routing at the very beginning of downloading. Also, it is important to keep the same exact file organization in at least two locations, where the files are backed up daily to the second location, externally to your personal computer or remaining in the cloud. In addition, one will save a lot of time if they discover what file format the data must be in to be able to directly use as input to the analysis tools. When one is combining data from various modalities, the data will be unbalanced across the samples, the data may need separate organizations for different programs or for different places in the pipeline, and the data may or may not need imputed. All of these possibilities are common data issues one will probably face when managing MMNG data sources. Even with the superior work from the software engineers and researchers at the TReNDS Center, the file organization was not entirely clear to me from the GIFT and FIT manuals, and I spent a large amount of time restructuring the data and file organization in order to fit to the Matlab program.

Table 3.2: Simulated and ADNI Data Organization

Modalities	Source	Data Dim	N
i. FDG	ADNI1	336x336x81	135
ii. sMRI	Sim/ADNI1	192x192x160	varies/135
iii. SNP	Sim/ADNI1	113,459	varies
iv. rsfMRI	Sim/ADNI2	448x448x197	varies/74
vi. sMRI+SNP	ADNI1	192x192x160 113,459	135 135
vii. sMRI+FDG	ADNI1	192x192x160 336x336x81	135 135
ix. sMRI+FDG+SNP	ADNI1	192x192x160 336x336x81 113,459	135 135 135

Data considered in these analyses begins with unbiased, simulated data, as discussed in Section 3.2.1. I also downloaded real-life data, collected from ADNI1 and ADNI 2 protocols, for assessment of the fusion strategies and optimizations. The ADNI

data modalities selected from ADNI1 include sMRI, FDG-PET and SNP data, from ADNI2 we swap the FDG-PET with rsfMRI. Information about the dimensions, and different combinations to be considered are showcased in Table 3.2. The dimensions of each modality differs, meaning that the various data combinations also test the ability of these algorithms to perform on varying degrees of depth. The largest data dimensions are considered for the three-way data fusion, totaling over 15 million observations per subject.

3.3.2 Data Pre-processing

Upon the collection of biomedical data, the biological information are transferred from the physical realm to the numerical quantities that are used in the analyses. That means there are many steps to take between retrieving the biological response, be it a BOLD fMRI reading or a blood sample for genotyping, to computing the algorithms. Even small deviations in this process may make a large difference in results obtained due to the high-dimensionality of the data. Therefore, it may be important to secure consistently pre-processed data across studies when using the results to derive biologically informative conclusions. On the other hand, we may learn important information about the study by making slight changes to the pre-processing as well. It is also important to understand and carry out the pre-processing, because the data won't always be analysis-ready. For the neuroimaging data in ADNI, I accessed the pre-processed data. If I had time to do this over, I would also download the raw data and test my results after processing the data myself.

Regardless of the pre-processing procedure that is chosen for the MMNG analysis, another learning curve involves learning new data properties and the science that underscores this collection. Often imaging genetics problems are tackled by multi-disciplinary researchers who are analyzing at least one data source that is outside of their own field. This requires another literature review to learn about the properties of the data, common issues faced in data organization and analysis for that specific data type. A wise researcher would probably also consult an expert (or two, or three..) in that type of data for confirmation that the procedures followed and/or the assumptions made are consistent. Due to the goal of this research being to test a new ICA framework under various optimization strategies, I made the decision early on to follow standard protocol when processing and imputing the genetic data and to use the standardized imaging data sets.

Due to the diligence of the ADNI MRI Core, consistently pre-processed neuroimaging data are provided for all ADNI collection protocols. One must keep in mind how the pre-processing procedures have changed over the years due to advances in technology and computing. The imaging data pulled from ADNI are pre-processed by one of the Imaging Cores, depending on the modalities. In fact, ADNI collected the data, pre-processed the raw images, and uploaded this to the online database. Several publications are dedicated to described these procedures and encouraging researchers to

use the standardized and pre-processed ADNI data (Jack et al., 2008; Wyman et al., 2013). The sMRI data were processed by the MRI Core using MPRAGE processing on voxel-wise whole brain volumes for sMRI. Pre-processing of FDG-PET uwas performed by the PET Core at the University of Michigan. The steps involved coregistration, averaged and standardized images. The voxels were sized to have uniform resolution. The attenuation correction was measured at ACCT with an all-pass convolution kernel. Pre-processing was performed utilizing a combination of the Statistical Parametric Mapping (SPM5) software by the Department of Cognitive Neurology at the University College of London, UK), the Resting-State fMRI Data Analysis Toolkit (REST) v1.5, Data Processing Assistant for Resting-State fMRI (DPARSF) using software in MATLAB v7.11. Another step taken was to discard the first 3 volumes to obtain steady state magnetization, slice time correction, realignment, normalization to SPM5 EPI template, smoothing with 4 mm full-width half maximum Gaussian kernel, linear detrending to correct for signal drift, and 0.01–0.08 Hz bandpass filtering to reduce non-neuronal contributions to the signal fluctuations (Alzheimer et al., 2018). For more information of the processes involving the other neuroimaging modalities, please see the literature Jack et al., 2008; M. W. Weiner, Veitch, Aisen, Beckett, Nigel, et al., 2017; Y. Zhang et al., 2016.

The SNP data, however, are in need of processing. Usually this includes an evaluation of a minimum minor allele frequency, holding of the Hardy-Weinburg equilibrium, a minimum genotyping call rate and possible adjustment for effects of population stratification or duplicates. Fortunately, this process may be done in PLINK, a open-source computational $C/C++$ tool for whole-genome association and population-based linkage analyses Purcell et al., 2007. However, this involves another learning curve in order to gain familiarity with the program. The SNP data were downloaded as a PLINK file, since this would be the easiest download and provide the most consistent pre-processing with the literature. In order to understand the process of quality control, or pre-processing of the data, I read multiple papers on this process and followed the suggestions made in the majority of papers that use SNP data form ADNI. These papers included: (Saykin et al., 2015a; Stein, Hua, Morra, et al., 2010), Nathoo2019, Purcell2007. The following quality control (QC) procedure was taken on the 620,901 markers collected per individual: genotype call rate $< 95\%$, significant deviation from Hardy-Weinberg equilibrium $P < 5.7 \times 10^{-7}$, minor allele frequency < 0.10 and a quality control score of $< .15$, (Stein, Hua, Lee, et al., 2010).

3.3.3 Data Organization and Manipulation

The data organization and manipulation do not occur separately from the data pre-processing, but rather simultaneously. However, I've split this up into two sections because they involve very tedious courses of action. As an imaging genetics researcher, one often and appropriately begins with the data selection and downloading before really considering the analysis tools. However, I recommend that one check the programs

that will be used, the formats of the data that are required for each modality, and then to be sure to download this in a convenient way. For instance, there are multiple formats of imaging data that are used, including NiFTI (.nii), ANALYZE (.img/.hdr), BIDS, DICOM (.dcm), or meta files extensions, such as .json, .bvec, and .tvc. I made the mistake of downloading many GB of imaging data in NiFTI format when I realized one of the Matlab programs required the ANALYZE format. There are computing tools for transforming data to multiple data types, but I don't recommend doing so unless one is very familiar with this data type and has a clear protocol for validating this transformation. After genetic data were pre-processed in PLINK, I then downloaded these files into R through the command line as text files. However, then I learned that the Matlab program would require a utf-8 test format to be written in ASCII with an .asc extension, also done from the command line. I could have saved myself some frustration and time had I looked into the program prior to downloading the data.

One must then organize the data in the exact file format as requested by the computing software, which is often contrary to the organization from the actual data source. For instance the ADNI imaging data is downloaded with one folder per subject. If this subject has more than one data collection line, then each collection instance will show up as a subfolder. Within these folders, the data are saved in various formats depending on the level of pre-processing that has been done. Not to mention, the file names are all very long. I had to sharpen my command line skills by finding a way to extract the data files out from 6-8 subfiles, renaming each of the files, and then placing this in the parent directory needed. When you are comparing different disease categories and multiple modalities, the data must be reorganized for each of the studies. That is, each of the data combinations in Table 3.2 required a different file organization.

Next, one must consider the process of data imputation. For the neuroimaging data, I reduced the samples to subjects who retained data for all of the modalities. However, this reduced the sample size a lot, and future studies may consider imputation of imaging data (D. Lee et al., 2019; Wimley, 2017). The missing genetic data must necessarily be imputed, however these procedures are more straightforward. Although there are still expected to be missing SNP expressions, leaving "holes" in the genome. These will be imputed at the SNP-level using the HapMap3 panel with NCBI build 36/hg18 using IMPUTE2 (Howie et al., 2009). Derived from a meta-analysis of HapMap, this reference panel of haplotypes will identify the patterns expected within each haplotype block. Missing genotype data may be imputed with very high accuracy (Marchini & Howie, 2010).

3.3.4 Computational Tools

The whole process of performing a multi-modal data fusion necessitates computational tools at each step of the process. From data collection, to pre-processing, organization and manipulation, each of these steps involves computing of some sort, from database management to command line knowledge. These points have been made in the previ-

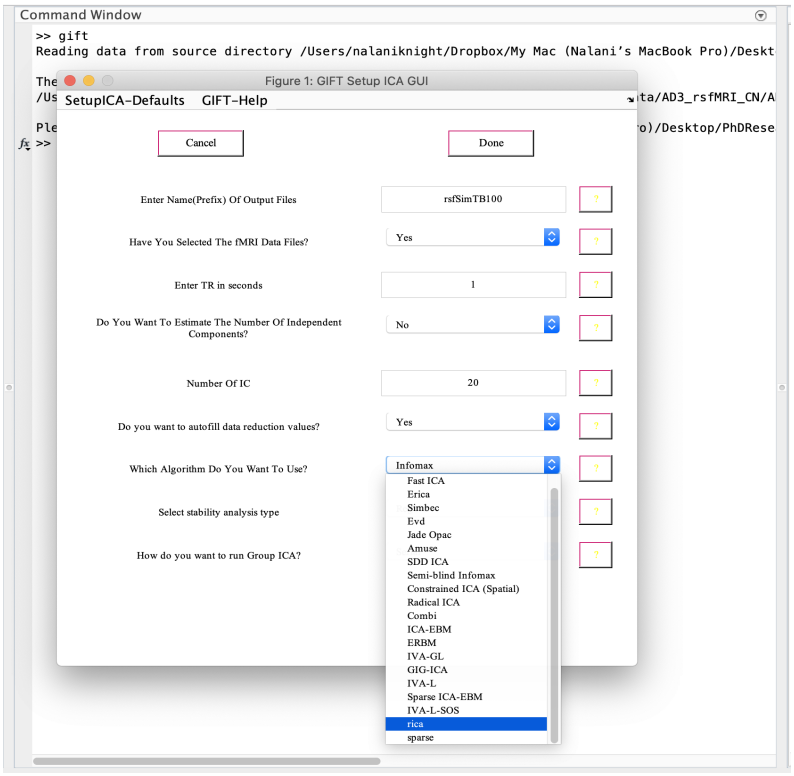
ous sections. The next stage in the Imaging Genetics Pipeline is to use computation for the fusion stage. Data fusion may be performed in Matlab, R or Python. For this project, I used mostly Matlab, since the tools for imaging data (and large data, in general) are established and well-documented. For multi-modal data fusion, I utilized packages and functions already written in Matlab. In addition, I built my own functions from some of the current open-source toolboxes. Along with this pipeline, I have created a GitHub page to guide one through the computational steps (https://github.com/knalani714/MultiModalDataFusion.git). Below, I will describe the code that was utilized and developed in Matlab, in addition to writing functions in R.

The Center for Translational Research in Neuroimaging and Data Science (TReNDS) website and GitHub provides the foundation for the code that will be used (https://trendscenter.org/software/FIT/; https://github.com/trendscenter). Group ICA/IVA software in Matlab, called GIFT, performs blind source separation using multiple ICA algorithms for imaging data, especially fMRI. Source-based morphometry, gICA for sMRI data, is also found within this code. Additional analyses may be done, including temporal and spatial network connectivity, Mancovan, and extensive display features enabling visual and quantitative group comparisons. The Fusion ICA Toolbox (FIT) in Matlab, also created by the same authors, is used to perform up to 3-way multi-modal analysis for imaging genetics, multiple imaging modalities and/or behavioral or EEG data. The methods employed in the toolbox include jICA, pICA, CCA, tIVA, and more with fantastic display and output features for summarization of the methods. The algorithms already employed in these toolboxes are: Infomax, FastICA, JADE, SIMBEC, AMUSE, ERICA, EVD, GIG-ICA, and more.

In order to develop the code for rICA and sfICA, I decided it would be most useful if I could write the algorithms into the open-source Matlab Toolboxes already available. Doing so would allow me to utilize the pre- and post-processing steps as well as the display features. The amount of time this took may be equivalent to downloading and preparation of the data. There are multiple Matlab folders within the files associated with both toolboxes, and I had to determine which files work together to achieve the output in order to know which scripts to rewrite. Rather than altering the Toolbox directly, I created new files with a subscript after the name of the novel method in order to keep this separate from the rest of the files and well-organized. This involved changing 20+ Matlab files, which amounts to shifting through and organizing thousands of lines of Matlab code. When I replaced the prior files that do not contain rICA and sfICA with the new files, indeed the ICA method that you select from the point-and-click GUI includes the new methods in the drop-down button. An example of what this code looks like written into the drop-down menus is provided in Figure 3.7. At the bottom, one will see the option for "rica" and "sparse".

There were some issues that I came across with the shared Matlab code from these packages. At first, the methods were not working correctly, because the website had not been immediately uploaded when the new verision was released. However, I was able to find the code I needed from the GitHub. Next, the example data was not placed in the

Figure 3.7: GIFT GUI in Matlab adjusted for rICA and sparseICA.



file location mentioned in the manual, and I had to search through another GitHub page to retrieve it. When testing some of the algorithms in the FIT package, I desired to perform the deep learning analysis which calls Python code from Matlab. However, the code was providing many different errors about incompatible packages. I contacted the software engineer, and he said that the package required the use of Matlab 2015a/b. This was a downfall in the FIT package. One may find a Docker container on the GIFT package which allows consistent analyses using groupICA for all future versions, but this does not exist for the FIT package. This is a downside to the FIT package and does not allow fully reproducible science. That being said, I do believe the software engineers are aware of the issue and are working to resolve this. For more information on writing code into a Docker container, please see: https://www.docker.com, (Boettiger, 2015; Rad et al., 2017).

Currently, there is not a multi-modal decomposition package in R. However, there are single-modality ICA packages that implement the traditional ICA algorithms, such as FastICA, Infomax, JADE, etc. There is a package, called MineICA that attempts to perform ICASSO on these analyses with one data source. When initially testing the stability of FastICA and Infomax, I discovered that the ICASSO method written into

this package only randomizes the initialization parameters and does not perform bootstrapping of the data, as it claims. Therefore, I wrote a new function in R that is derived from their original code that corrects the bootstrapping error. Also, the package they wrote was only applied to FastICA, but I needed to apply this method using the Infomax algorithm as well. Therefore, I wrote another function with the same structure that performs an ICASSO analysis using the Infomax algorithm. In order to run the repetitions on the new algorithms, I had to use Matlab. Matlab had pre-written L_2 sparsity filtering and reconstruction penalty functions that I used in writing my code. Going forward, it would be beneficial to write an R ICASSO package that includes all of the algorithms I compared. Even more so, a future idea could be to write a multimodal imaging genetics package in R.

As an early researcher, the learning curve was steep for downloading, cleaning, manipulating, and running computation of MMNG analyses. While I developed more efficient organization and big data skills along the way, many of the issues that I mentioned above were a result of vague instruction manuals, the lack of priority on reproducibility for some of the data or computation sources I used, or even blatantly incorrect code written (in R). This motivates the need for making open science a priority. Unfortunately, I was not aware of the available tools for practicing open science at the start of my dissertation work. Fortunately, I learned the necessity of implementing these practices along the way. For this reason, I am still in the process of turning my own code and this pipeline into a shareable format that will be open to the research community. Figure 3.8 provides the a flow diagram that I created for the Imaging Genetics Decomposition Pipeline that I followed. This flowchart need not be restricted to the ADNI data source and may apply to other MMNG decomposition analyses.

3.3.5 Comparison Metrics

Even prior to setting up the data and analysis plans, one of the biggest issues one must consider is this: *what is the end goal of the data fusion*? If this is to derive a greater understanding of the direct inter- and intra-relationships with the pathology and progression of AD, then *how does the multi-modal algorithm permit scientific translation?* Thus far, we have gained insight into the relevance and potential of multi-modal studies across disciplines. The architecture of two mainstreams of data fusion, extensions to component analysis and feature extraction by deep learning, were introduced. In addition, an example with real-life data was presented with step-by-step instructions for how to begin data fusion of multiple heterogeneous data sources. It is important to note that if we are to practice data fusion across disciplines, we must be pragmatic about the approach taken. The most desirable algorithm should not complicate the implementation or the end goal may be lost. Nor should the algorithm of choice be one that overly simplifies the problem, dumbing the data down to such a degree which may prevent new findings. Indeed, few researchers have looked at multi-modal data fusion through the lens of algorithmic reliability or statistical stability. These quanti-

tative matters then feed into vital conceptual matters to consider: interpretability and generalizability. *May these concepts be utilized across disciplines and replicable within the application field?* These sections further drive home the encompassing goal of this project: to create a reproducible MMNG pipeline that practices open science to the best of my ability.

Replication and Simulation

While past multi-modal data fusion techniques have already lent themselves to groundbreaking discoveries across multiple fields, in some ways, the formulation of a robust data fusion strategy is still in the stage of infancy. Depending on the problem, one may proceed with a decomposition method such as n-way parallel ICA or use a black box, machine learning, approach that improves classification under the assumption of well-labeled data. However, what brings this from concept to action, or from analysis to improving the way we do science, is by setting up the *data science* to permit **translational science**. How may one justify the methods for use in other fields? Is it possible to develop the theory to allow for replicability? These questions may be answered by assessing the algorithmic stability along with the statistical reliability.

When a method is replicable, the knowledge, programming and data are open to other researchers so they may obtain similar findings. This allows for proper testing of the methodology; when you use the same methods on different data but with the same research question in mind, you should obtain similar results. This idea of replicability regardless of the data at hand is known as *statistical reliability*, which may be measured using multiple data sets, bootstrapping from the same dataset and/or using a cross-validation method by splitting up your data into testing and training sets. Another important aspect of replicability looks at the optimization techniques used and assesses whether alterations in the initial parameters or underlying assumptions of the algorithm changes the outcome, known as *algorithmic stability*. In addition, one will want to use a method that is feasible, or does not take too long to run and does not necessitate costly computational resources. Therefore, Chapter 4 will present a measure of replicability for all of the algorithms, frameworks and data sets considered. Both simulated data and data pulled from ADNI will be used in the testing phases of the multi-modal methods. The simulations will display the reliability and stability at each sample size and with each modality, first singularly, then with each possible pairwise modalities, and finally with all three modalities.

Interpretability and Generalizability

Interpretability: how may the results be described in context of the data problem?

This leads me to the question: how may I interpret the results, outcome, and/or outputs of the analysis in the context of the data problem? The ability to conceptualize the outcome of the algorithm is vital to proceeding with inferential procedures. How

will the fused result relay back to the separate modalities as well as contribute to the understanding of the interactive processes among the modalities? Does the data fusion provide new knowledge or confirm concepts already known? All in all, what novel information is gained from the multi-modal analysis?

The process of data fusion requires that the data be manipulated in some way and then combined so that one joint data set may be used for any follow-up analyses. The resulting fused data have reduced heterogeneity and the patterns of the complex data sources parsed. Recall that both the blind and semi-blind decomposition methods reduce the data to ICs that maintain interpretability. That is, the components, or rows of the sources, $\mathbf{S}^{(M)}$, are observation-specific influences of the $\mathcal{M}$ features, where $\mathcal{M}$ represents the new dimensionality of the selected components. The columns of $\mathbf{S}^{(M)}$ are feature-specific with the weights of the values indicative of the impact of that feature on the model averaged over all observations. On the other hand, the outcome of the DL methods do not maintain feature-specific or observation-specific interpretations. Due to the end-to-end property of DNN, the "end" goal of the study must be incorporated into the model so that the data fusion may be directly used. For example, one may be interested in setting up a classification model. In this scenario, a DL fusion method would rely on the classification portion of the results to relay the fused result back to the raw data or scientific problem under study.

Generalizability: how does the method generalize to future analyses across and within disciplines?

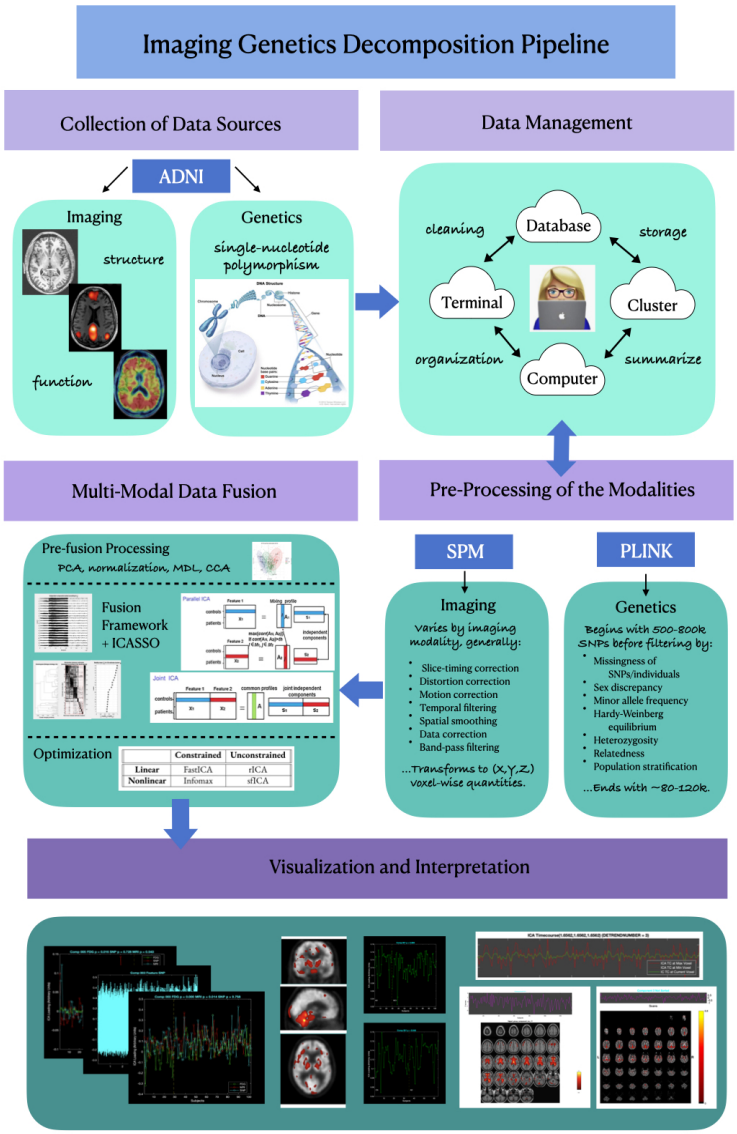
Another important aspect of a study to consider is the overall applicability of the procedure to a variety of problems. In other words, does the data fusion procedure have the ability to be generalized to diverse research scenarios? This answer may be assessed *globally* and *locally*. A global generalizability refers to the scientific community as a whole. Can the method itself be realistically implemented from discipline to discipline? The local generalizability refers to studies within one's own field. From study to study, can the algorithm be replicated in such a way that paves the multi-modal path forward in a particular discipline? Global generalization allows for a methodological impact on multi-modal data fusion to the entire scientific community. While local generalization permits replicable research within a discipline. Replicability is imperative to test the conclusions made from new studies and ensures that the ethical and biased decisions that go into a study are controlled for.

The interdisciplinary impact of the data revolution was supported with diverse examples earlier in this document. From a global perspective, multi-modal data fusion techniques must be generalizable across disciplines, otherwise the overall scientific gain from the influx of novel data will be sparse at best. Under this scenario, several disciplines will surely benefit from analyzing data simultaneously from multiple sources. However, the rate at which new scientific discoveries are made as a result of multi-modal resources will not be able to keep up with the data collections. This could create

other issues such as funding outcomes for grants. On the other hand, the quality of research is impacted by local generalization. If algorithms have to be re-trained for each study, then it will be challenging to create a pipeline for analysis within a discipline. Without a pipeline for implementation, the productivity of the research would slow. Furthermore, if it becomes a challenge for researchers to try the methods and test the new theories gained from the multi-modal analyses, then only minimal, and questionable, results will remain.

Generalizability as a whole is attached to how parsimonious the theory, computation, interpretation and end results are. This is quite a challenge when considering the complexity of the data to begin with! Decomposition methods have a concise and concrete implementation and interpretation to the procedure. The higher-order statistics used are the same from discipline to discipline, and the optimization algorithms, whether Infomax, FastICA or JADE are used, all maximize based on some estimate of non-normality. The interpretation of the ICs are observation- and feature-specific for each analysis. In contrast, DL methods input data into a "black box" that is not easily understood, nor is it easily implemented computationally. In addition, DNN must be re-trained for each new set of data sources. Only very similar data scenarios may allow the creation of a pipeline. The ability to utilize the ICs in a downstream analysis leave abundant possibilities for generalizability and interpretation. Thus, one could argue that decomposition through component analysis overtakes DL feature representation in both global and local generalizability. There will be caveats to this conclusion, and each discipline should assess the relative impact.

Figure 3.8: The Imaging Genetics Decomposition Pipeline follows this order: (1) data collection; (2) data management; (3) pre-processing of the modalities; (4) data fusion; (5) visualization/interpretation of the component results.



Chapter 4

Results

In disseminating the motivation, data challenges and multi-modal models throughout this work, I built the foundation for implementing a multi-modal analysis for the advancement of AD research. Now, the MMNG methodology is implemented for ADNI data in order to showcase the advantages of the novel fusion framework that I propose as well as to explore findings relating neuroimaging and genetic biomarkers to AD diagnosis. The chapter begins with results from simulated data on the four different optimization algorithms under consideration, providing a comparison of the traditional constrained algorithms with the newly proposed unconstrained algorithms. Then, the remaining sections compare the single-modal analyses of the various ADNI data sources before moving on to a multi-modal comparison. Model comparisons are considered between the new MMNG fusion framework proposed, utilizing the unconstrained optimization approaches, and previously proposed mCCA+jICA. These comparisons focus on how the reliability and stability of the estimates change for unconstrained optimization compared with more traditional approaches, a more flexible multi-modal fusion framework, under multiple combinations of modalities, and to see how stability is impacted when including SNP data.

4.1 Simulation Study of ICA Algorithms

In the first section, I present the simulation results on the simulated data for SNP, sMRI and rsfMRI data. The studies involving the simulated data are presented first in order to begin with a bias-free study on the algorithms used in a MMNG ICA analysis. We begin with the SNP data, simulated from the R package sim1000G, as mentioned in the previous chapter, (Dimitromanolakis et al., 2019). Next, neuroimaging simulations are carried out, simulated sMRI data, using neuRosim, and simulated rsfMRI data, created from the SimTB

Matlab toolbox. All three modalities are compared using the FastICA, Infomax, rICA and sfICA algorithms. The algorithmic and statistical reliability are assessed by running ICASSO and taking repetitions by randomizing the initializations, bootstrapping the data, and increasing the number of ICA runs per ICASSO set. Then I calculate the cluster compactness by looking at the behavior of I_q. In order to assess the overall replicability, I derive other measures of ICA performance so that we may determine how well the extracted features of the various methods are able to recover the ground truth.

4.1.1 Sim1000G: SNP data

Simulating the SNP data is relatively straightforward, due to the ease of the sim1000G package in R. This package was designed to allow one to specify whether the sample should comprise of unrelated individuals or family-based genetics. While this opens up the door for family-designed studies, which are often difficult to arrange, I focus on simulating genetic markers from unrelated individuals. The reason for this is that I want to assume that I am using a random sample, at least as much as possible. Samples of size 20, 50, 100 and 300 are collected, and the algorithmic and statistical reliability are assessed.

FastICA results on SNP data for the various sample sizes may be seen in Figure 4.1. These plots reveal how increasing the number of ICA runs may impact the statistical reliability of the estimates. This was assessed by bootstrapping the data 10 times each for an increasing number of runs, from 5 to 200 in increments of 5. This means that 400 runs of ICASSO were performed, and the maximum I_q was extracted for each run. Within each of the 10 repetitions, the average and standard deviations of the $\max(I_q)$ were calculated. The plots display the means for number of runs for each of the sample sizes. The number of components extracted for each run remained constant at 10. The goal of this simulation was to assess how reliable ICA estimates are. Not only does this assess the impact of using different data, but it also assesses whether or not an increased number of ICA runs will improve the stability.

For the SNP data, all sample sizes reveal similar results, beginning with a higher cluster compactness with 5 runs, then decreasing to around 0.30-0.35 by 35-40 runs. This reveals that no improvement is shown in the statistical reliability as you increase the number of ICA runs and the overall compactness is low. At the same time, the standard deviation of the compactness index is calculated to see how variable these results are for an increasing number of ICA runs. Figure 4.2 shows the standard deviation results over the same runs. We see a random fluctuation of variability between 0.02 and 0.06 for both n=50 and n=300. Therefore, the sample size of the data and the number of runs within

Figure 4.1: Mean of max I_q for bootstrapped SNP data for $n = 20, 50, 100, 300$ using FastICA.

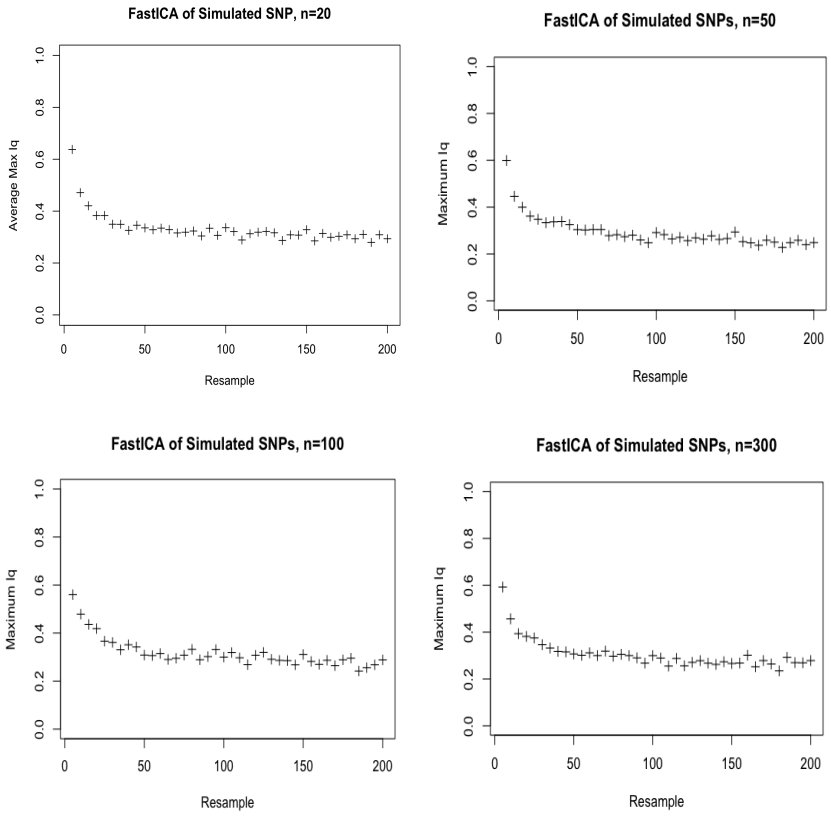


Figure 4.2: Standard deviation of I_q for bootstrapped SNP data for $n = 50, 300$ using FastICA.

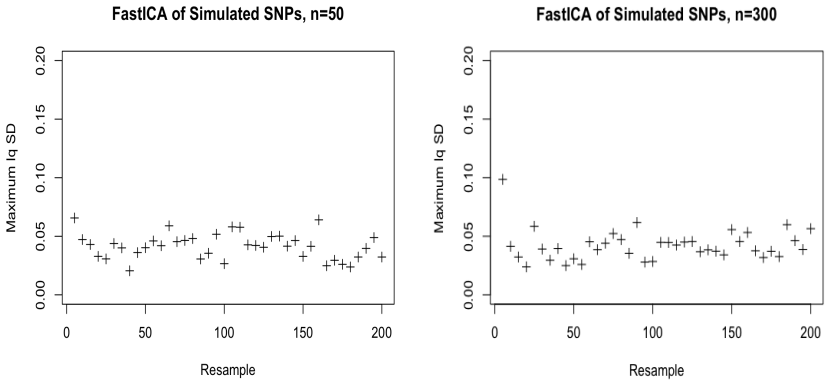
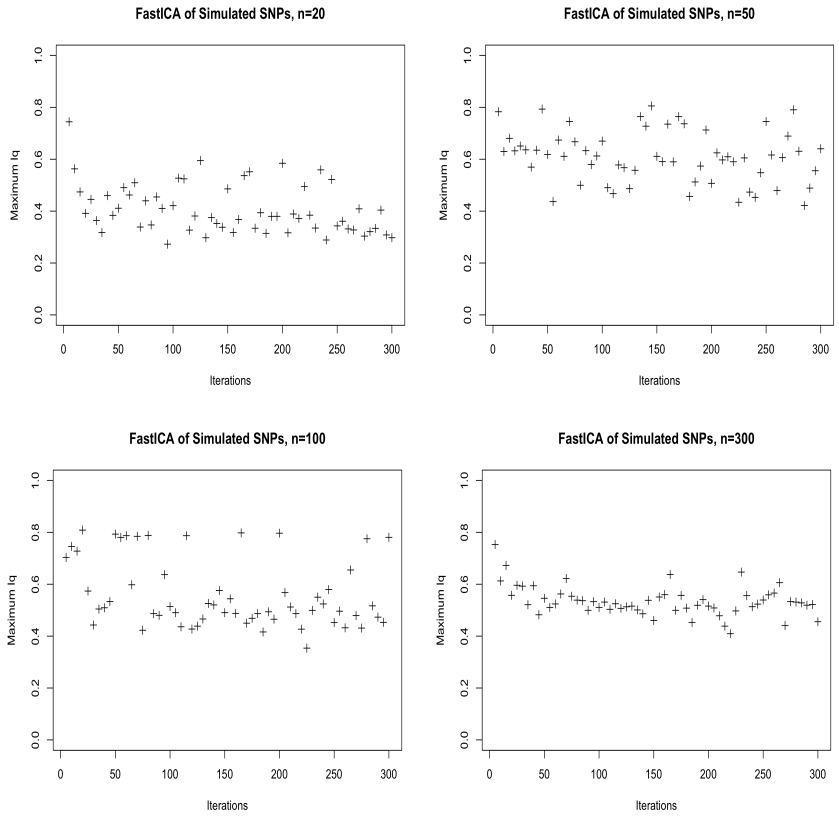


Figure 4.3: Max I_q for random initializations, simulated SNP data for $n = 20, 50, 100, 300$ using FastICA.

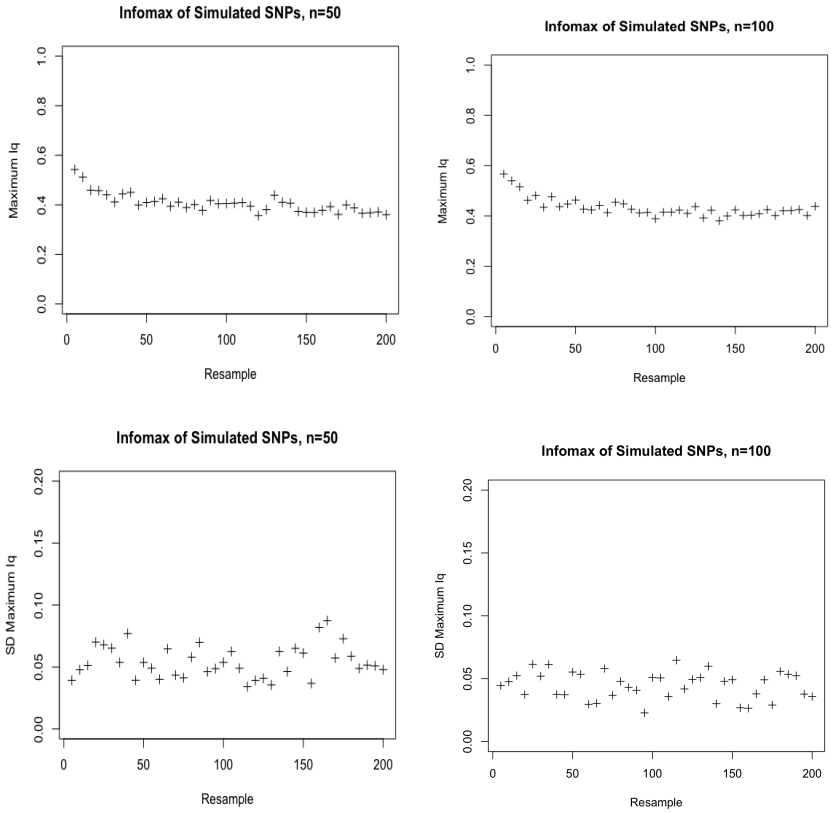


each ICASSO set has no impact on the variability. That is, we can pinpoint the variability we expect from run to run and do not worry about inflating the distance from the cluster means when introducing new data.

Figure 4.3 tells a different story. Instead of shuffling through the data, I alter the parameter initializations in the algorithm while increasing the number of ICA runs within each ICASSO. Sample size does have an impact on the algorithmic stability, as we see a more clear horizontal line appearing in the data by n=300. The overall spread of the mean maximum I_q is higher than the bootstrapped data; the center of the I_q scores appears to settle around 0.50-0.55, an improvement from the statistical reliability of the SNP data. In addition, when the sample size is small, the means tend to center around 0.4, then a slight increase is observed by n=50. For the FastICA algorithm, it took approximately 3 hours for each set of the 400 ICASSO runs.

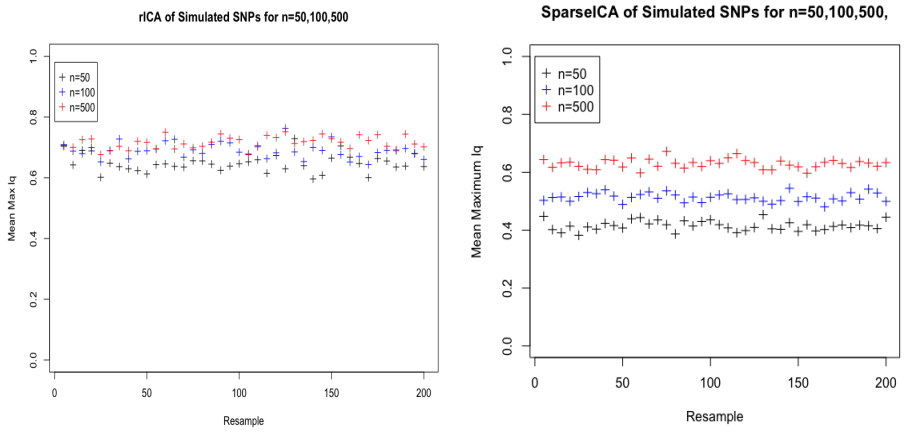
These results showcased the popular FastICA algorithm, bringing into question how reliable the IC estimates are from data set to data set as the overall compactness remained low. Introducing another traditional optimization, we can observe the impact of using a nonlinear gradient descent approach, Infomax. The Infomax runs took longer, at about 4.5hr on average, when performing the same number of bootstraps; that is, 400 ICASSO runs with increasing number of runs 5 to 200 in increments of 5. This is what one would expect for a slower, gradient descent algorithm. For this approach, the algorithm performs optimization of a convex function which promises convergence, thereby making the use of initializations unnecessary. However, one may observe the mean and standard deviation of $\max(I_q)$ for the bootstrapped, simulated SNP data in Figure 4.4. The plots display the results for n=50 and n=100. The compactness of the clusters remains consistent for both sample sizes, and the standard deviation is about .05, similar behavior to FastICA. However, the I_q values are just slightly higher than for FastICA. See Table 4.1 for a summarization of the I_q outputs.

Figure 4.4: Mean (top row) and sd (bottom row) of max I_q for bootstrapped SNP data for n=50,100 using Infomax.



With this goal in mind, the results of rICA and sfICA are presented. In order to run the new ICA algorithms using ICASSO, the code is written and run in Matlab. Due to the computing capability of Matlab, I consider sample sizes of 50, 100 and, additionally, 500. While the same number of repetitions were performed on the new algorithms, the nature of the repetitions vary. That is, instead of assessing statistical and algorithmic reliability separately, this is now done together so that each new run of ICA bootstraps the data and randomly iterates the initializations of the algorithms simultaneously. Analyzing the statistical reliability along with the algorithmic stability should reveal the overall replicability of a study. That is, for new data sets and across various runs set at different initializations, we desire to know how consistent the results are. This will tell researchers how well these methods will apply to new data sets.

Figure 4.5: Max I_q for bootstrapped, simulated SNP data using the rICA and sfICA algorithms on $n = 50,\ 100,\ 500$.



The unconstrained algorithms are plotted and placed side-by-side in Figure 4.5 for ease of comparison. The colors represent the sample sizes in ascending order, black, blue and red, respectively. At n=50, the rICA results show an increase in the overall replicability by obtaining a cluster compactedness around 0.65. This is an increase of about 0.20. For the larger data sets, we see a consistent increase, until we have an average maximum of 0.75 for $n = 500$. The reconstruction penalty appears to scale much better to the SNP data. When looking at sfICA, we see increases in the maximum I_q values as well, with clear separation between the sample sizes. However, it takes roughly n=500 to match the performance of rICA at n=50, at around 0.62-0.64. In addition, there appears to be a smaller spread between the data points, revealing that the new algorithms overall derive more stable results, both algorithmically and statistically. These algorithms were also written into the Matlab GIFT computational

tool in order to produce visualizations of the simulation results. Figure 4.6 showcases these tools for sfICA for 400 runs of ICASSO, extracting 50 ICs each for n=500.

Figure 4.6: ICASSO output using rICA for simulated SNP data, n=500

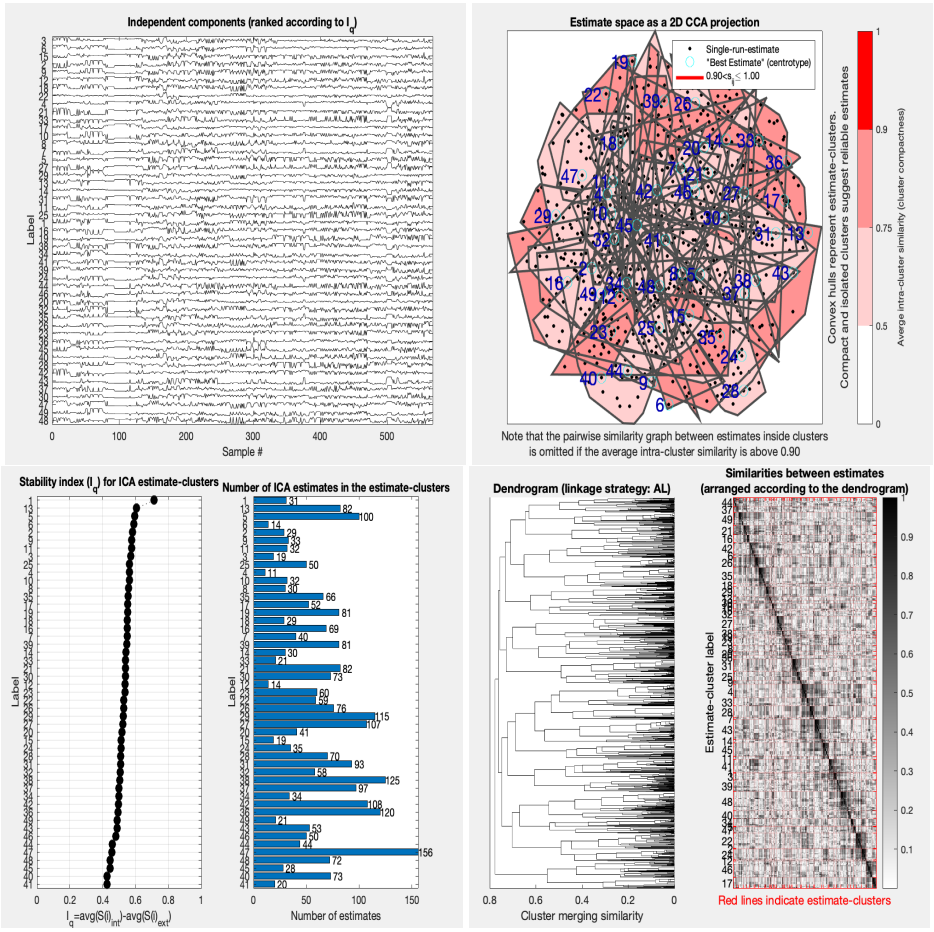


Table 4.1 summarizes the estimates of the cluster compactness, I_q, by providing the average maximum index over replications, then the mean and standard deviation of the first 10 components for each ICASSO run. Then the MD and AE are provided in order to compare performance with traditional estimates of ICA performance (Nordhausen et al., 2011). For each measurement, the best performance is indicated in bold. As a recap, we want a higher max and mean, a lower standard deviation, and lower values of MD and AE. The maximum mean cluster index is obtained by rICA, at 0.7554, followed with the largest mean. The minimum standard deviation and AE, derived over replications, are met by sfICA. Through the iterative procedure of gradient descent, Infomax is able to achieve the lowest minimum distance. However, the overall variability, measured through standard deviation and Amari, is lowest when using sparse

filtering. Altogether, the unconstrained algorithms of rICA and sfICA outperformed FastICA and Infomax by achieving the highest cluster compactness across replications and the lowest variability from run to run.

Table 4.1: Summarization of ICASSO for simulated SNP data, n=100.

Algorithm/Measure	Max	Mean	Sd	MD	Amari
FastICA	0.6968	0.595	0.0887	0.8766	0.3992
Infomax	0.5888	0.5701	0.0447	**0.7879**	0.4259
rICA	**0.7554**	**0.6819**	0.0337	0.7955	0.4012
sfICA	0.7114	0.6625	**0.0320**	0.7912	**0.3627**

4.1.2 neuRosim MRI

After a search of the literature for a sMRI simulation strategy, I discovered it was difficult to find a standard way to simulate volume-wise data as most simulations are focused on fMRI simulation. Instead of using BrainWeb as planned, I found a more simple approach. The simulated MRI data could have come from two sources discussed in the previous chapter, neuRosim and SimTB. Even though neuRosim was originally designed for the creation of data containing fMRI properties, one may alter the code to simulate MRI data. In order to do this, I simulated one time point and added Gaussian noise with magnitude to the data in order to use this as sMRI data. Samples were taken of size 20, 50, 100 and 300 and were saved as csv files to be used as input data sets into the ICASSO code (written in R). The other method was suggested by one of the SimTB creators who used SimTB for simulation of volume-based MRI measurements (Duan et al., 2020). However, instead of using SimTB for both imaging modalities, I wanted to be sure that we analyze diverse simulated data from different sources. Therefore, samples were taken using only neuRosim to be representative of sMRI data.

The same procedure was carried out for the simulated sMRI data as the SNP data. All together, 400 runs of FastICA, Infomax, rICA, and sfICA were performed and the mean $\max(I_q)$ plotted. Repeatedly bootstrapping the sMRI data reveals there may be some issues with statistical reliability using FastICA. Similar to the SNP data, for 5 ICA runs, the statistical reliability obtains a cluster index around 0.55-0.60 for sample sizes 20, 50, 100 and 300. However, this declines rapidly, and by about 50 ICA runs (per ICASSO), the average cluster compactness drops even below .20. Regardless of sample size, this same behavior occurs for each sample. In Figure 4.8, the standard deviations of I_q are very similar to that revealed for the SNP data, appearing randomly around 0.05. Fig-

Figure 4.7: Max I_q for bootstrapped, simulated sMRI data for $n = 20, 50, 100, 300$ with FastICA.

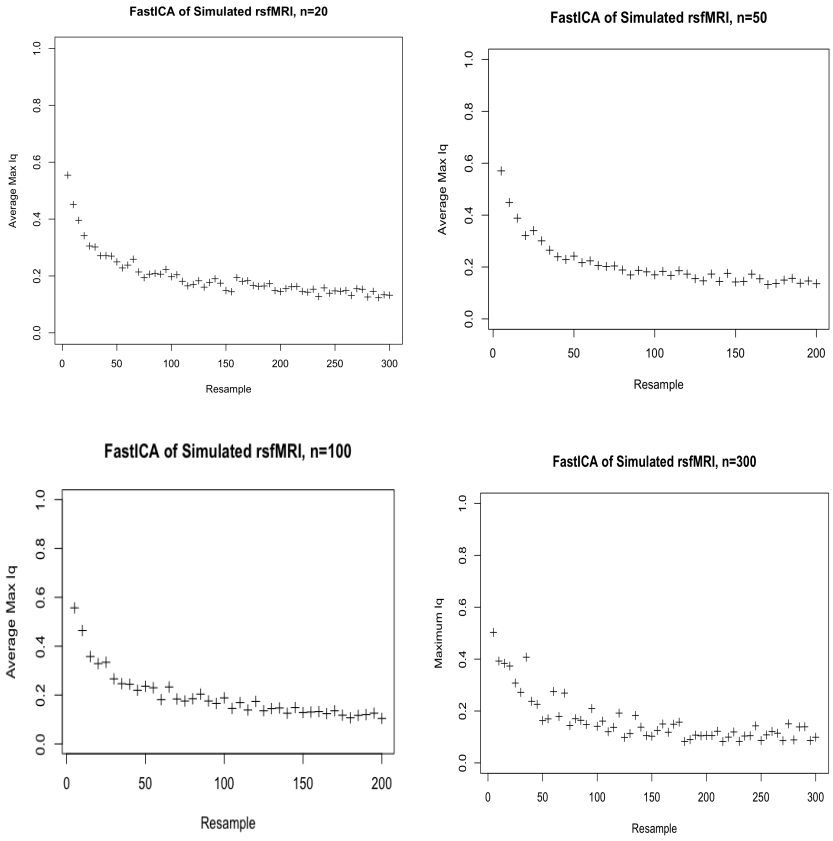
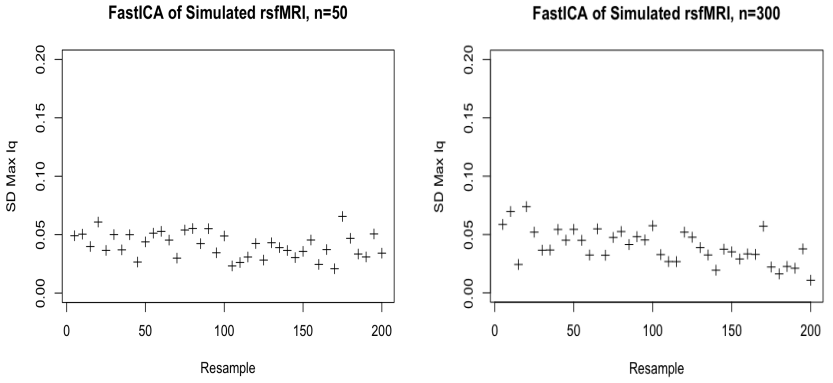
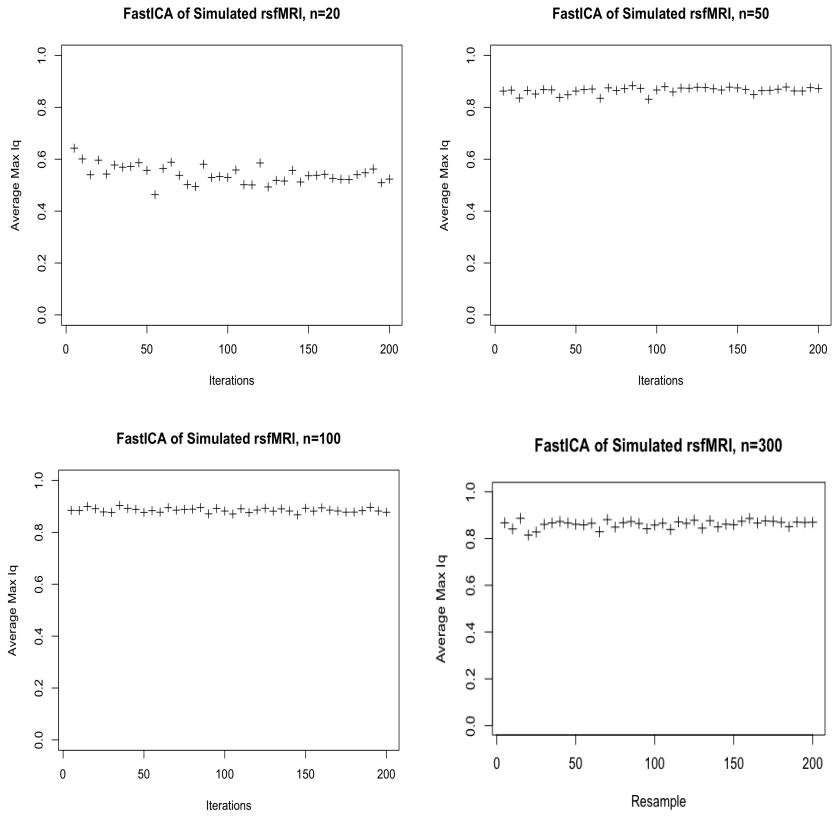


Figure 4.8: Standard deviation of I_q for bootstrapped sMRI data for n=50,300 with FastICA.



ure 4.9 displays the algorithmic reliability, showing an improvement in cluster stability and stays consistent across an increasing number of ICA runs. This remains between 0.55-0.60 for n=20, then increases and levels out around 0.80-0.90. Consistent with the SNP results, the algorithmic reliability is higher than the statistical stability.

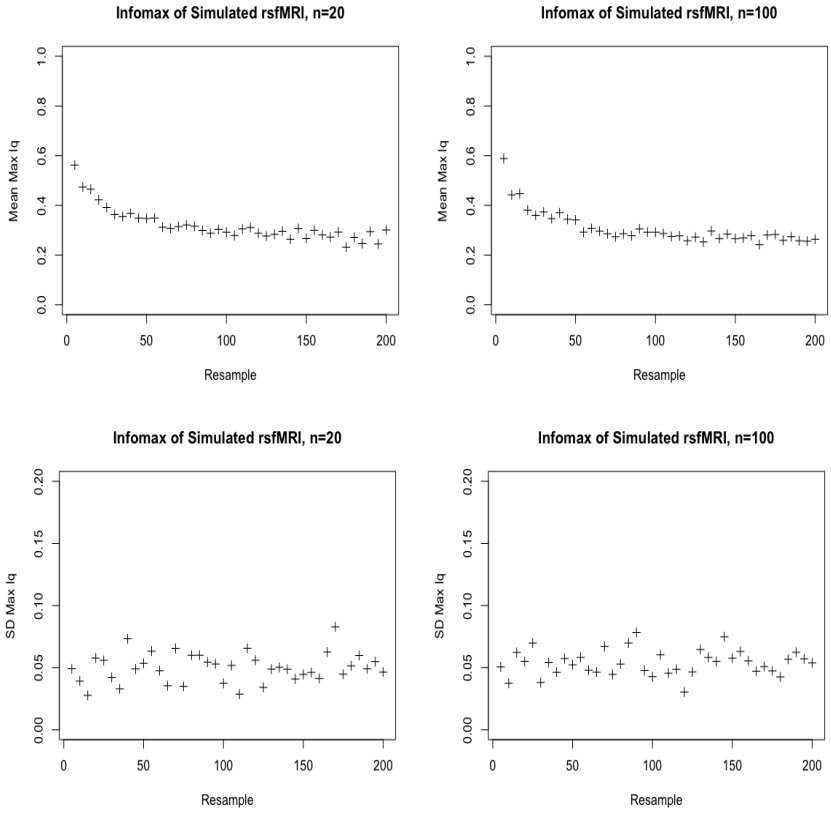
Figure 4.9: Max I_q for random initializations, simulated sMRI data for n=20,50,100,200 for FastICA.



Next, we can observe the nonlinear gradient approach of Infomax on the simulated sMRI data. Infomax results appear to improve the cluster compactness of the algorithm, but only a little bit as the average max I_q remains around 0.35-0.40. Initializations are not possible for Infomax, so we look only at reliability. The standard deviation is the same as received before. Overall, Infomax and FastICA don't show remarkable statistical stability, but the algorithms themselves are relatively reliable. Unfortunately, the major concern is for one to be able to rely on the estimates derived from ICA from data to data.

For the remaining algorithms, rICA and sfICA is run in Matlab as before. Due to the consistency of the variability across modalities, algorithms, and sample sizes, I gather the mean maximum I_q. Recall that this provides the first com-

Figure 4.10: Mean and sd of max I_q for bootstrapped sMRI data for n=20,100 using Infomax.



ponent of the multi-modal ICs, similar to the concept of the first component in PCA providing the optimal variance. In addition, for the new algorithms, it is of interest to see how the overall algorithmic and statistical reliability behaves and interacts under the same scenario. Therefore, for each run of ICA, the data are bootstrapped and the initialization is randomized simultaneously. Thus, the results showcase the overall replicability of the ICs under rICA and sfICA. Figure 4.11 places the rICA and sfICA results side-by-side for ease of comparison. Each point on the plot is the mean maximal I_q for each ICASSO run, separated by sample size in color. Although the separability between sample sizes are not as clear as with the SNP results, one may clearly see that n=20 (in black) provides the lowest cluster compactness, around 0.50 for rICA and 0.40 for sfICA. While the ICA method that allows for sparsity shows an increase for the larger sample sizes, we obtain similar mean indices for n=50 and n=100. A slight increase is then observed for n=300. Overall, these results show an in-

crease in the cluster compactness of the estimates over ICA runs, revealing that rICA and sfICA may improve the IC replicability.

Table 4.2 provides summarizations of the simulation findings. FastICA has the lowest AE error and sfICA shows the smallest standard deviation among the estimates, showing less fluctuation within the first ten components. However, the optimal mean I_q, max I_q and MD are reached by rICA. Given the max I_q of 0.4883, sfICA does not perform well for the simulated sMRI data. That is, sparsity does not show improvement for sMRI data, but rICA shows advantages over both FastICA and Infomax. Using a sparse auto-encoder optimization algorithm, the results show an improvement in the stability of the IC estimates. Figure 4.12 displays the ICASSO recovered sources, the CCA cluster similarity plot, the stability indices for each component and the dendogram with similarity matrix, providing a heatmap of the component similarities. Even though rICA outperforms the other algorithms, there is still a lot of overlap among the clustered components with an average compactness of 67% for the first ten components.

Figure 4.11: Max I_q for bootstrapped, simulated sMRI data for $n = 20, 50, 100, 300$ using the rICA algorithm.

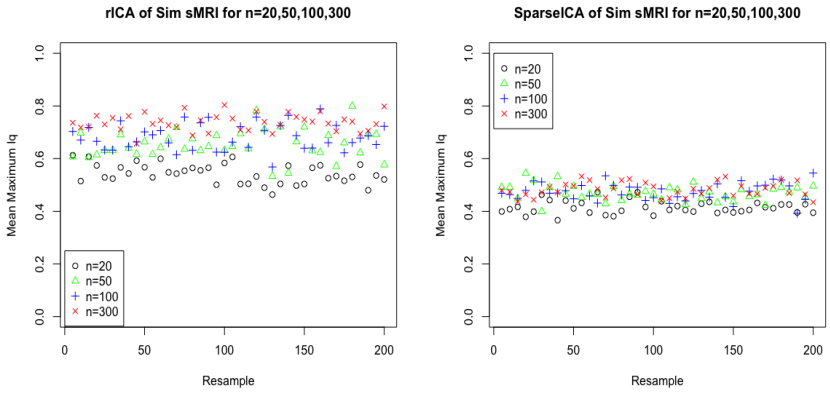
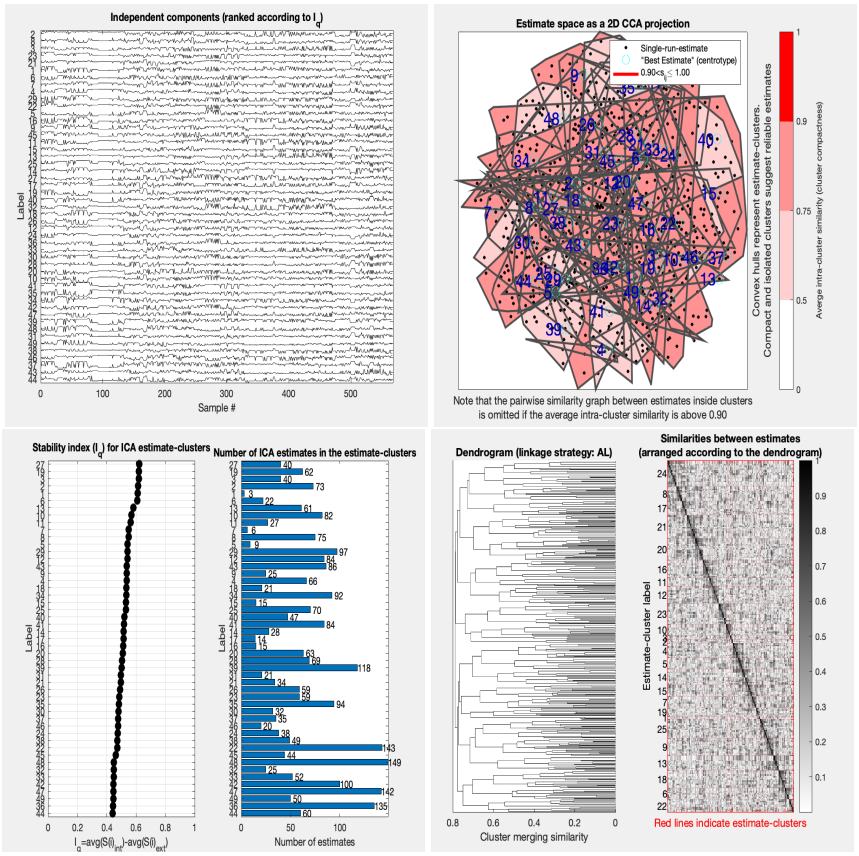


Table 4.2: Summarization of ICASSO results for simulated sMRI data.

Algorithm/Measure	Max	Mean	Sd	MD	Amari
FastICA	0.5874	0.5625	0.0239	0.8085	**0.3400**
Infomax	0.6287	0.6000	0.0268	0.8354	0.3678
rICA	**0.794**	**0.6735**	0.0676	**0.7893**	0.3592
sfICA	0.4883	0.4602	**0.0195**	0.8112	0.3612

Figure 4.12: ICASSO output using rICA for simulated sMRI data, n=300



4.1.3 SimTB rsfMRI

In order to simulate rsfMRI data, the data are collected using SimTB in Matlab and downloaded in NIfTI format in Matlab. These data are assumed to be resting-state instead of task-based, because the blocking effect and event-related portions of the design were set to 0 in the toolbox. The data extracted are representative of one "slice" of cortical matter. At first, I simulated 100, 150, and 200 time points, or time series (ts) for each "slice". This process took $11-14$ minutes for $n = 20$, $60-90$ minutes for $n = 100$, and 2hr 50min for $n = 300$ and $ts = 100$, and up to 3hr 45min for 200 ts. Due to the sheer volume of this data, and the common selection of $ts = 100 - 150$ for rsfMRI studies, I use only $ts = 100$.

The simulated rsfMRI data are analyzed using GIFT (also in Matlab), the group ICA package. The benefit of this simulation modality is that it allows us to visualize how well the components are "recovered". Using spatial ICA, the

various components pinpoint the anatomical components selected from Figure 3.3 in Chapter 3. In addition, one may run ICASSO directly within the package itself to also assess the cluster stability in this way. Using the data extracted for 100 time points and for the various simulated sample sizes, parameter initializations were included simultaneously with bootstrapping to assess the overall replicability that is measured for the data. Back-reconstruction is performed using spatial-temporal regression and the components are scaled as z-scores. It is important to note that the number of ICASSO runs are not increased at each iteration, as before. From the sMRI and SNP results, bootstrapping taken beyond 20-30 number of ICA runs did not offer much improvement, but rather stabilized. Therefore, these simulations focus on the stability of the various algorithms by performing 400 ICASSO runs, with 30 ICA reruns within each, extracting 10 components in order to be comparable.

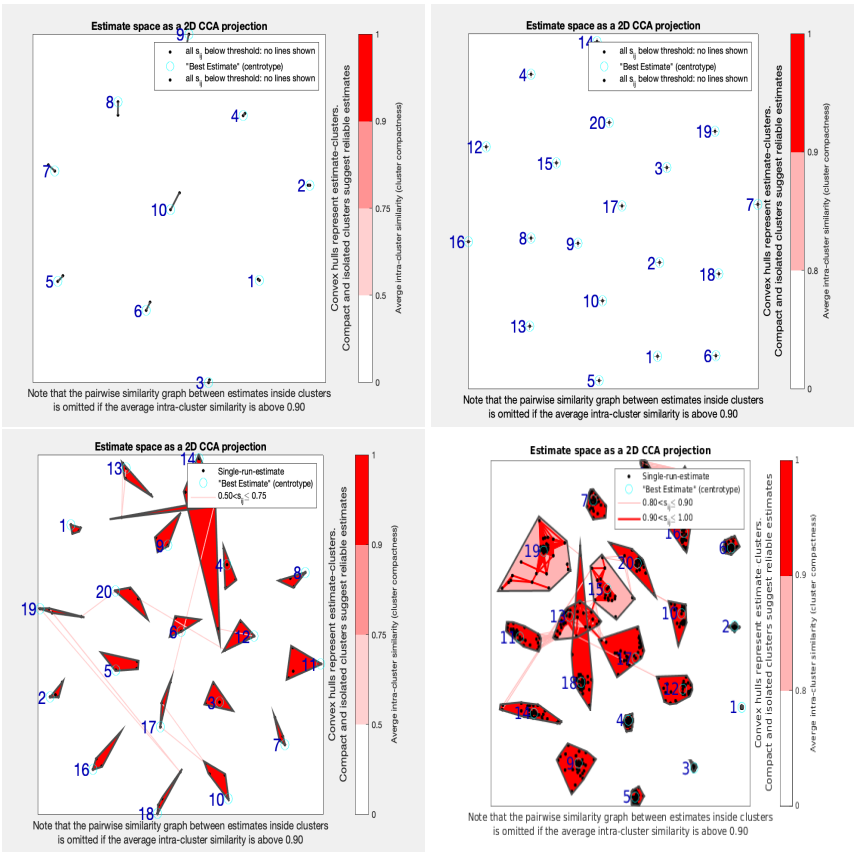
The FastICA, Infomax, rICA and sfICA algorithms were run on the data to assess the cluster stability across algorithms for the various sample sizes. Due to the sheer computational load that this requires using batch job submission, only sample sizes of 50 and 100 are assessed. When running the simulated data, even using cluster submission and setting GB=100, the job timed out. Therefore, using the Parallel Computing Toolbox in Matlab, I was able to submit the code, writing in multiple CPU cores on a single compute node by assigning 25 "workers" or CPUs per task. It's important to note that I wrote the rICA and sfICA algorithms into both the Group ICA and Fusion ICA toolboxes in Matlab. In order to test this, I ran the code on the GUI on my desktop for a much smaller sample size before submitting batch jobs in the cluster.

The graphs in Figure 4.13 reveal the ICASSO results for simulated rsfMRI at $n = 100$, and further results are summarized in Table 4.3. The FastICA graph on the left does not show any clusters in red. Instead, we see either a black dot or a small black line at the centroid of the graph. A line within clusters is shown when the IC connections represented are larger than 0.95 but the intra-cluster similarity is lower than this. However, nothing is shown if all connections within the cluster are above 0.90 (J. Zhang et al., 2012). Infomax offers an even closer intra-cluster similarity as shown by the "+" within each centroid and with a mean of 0.9901. While rICA and sfICA offer a close fit as well, we see more variability in clusters 10-20, ultimately showing that the strongest cluster compactness occurs within the first 1-8 clusters. Both traditional methods therefore show more consistent results, although these results show that the ICA algorithms overall perform well on the simulated rsfMRI data. It is important to note that SimTB is really simulated the source signals, rather than the full MRI process. This was discovered by corresponding with authors of the

following papers, (Duan et al., 2020; Erhardt et al., 2012; Vergara et al., 2014). For this reason, I decided to download rsfMRI data from ADNI2 in order to compare these results with real data (more on this in the next section).

Furthermore, the results in Table 4.3 confirm the plots that show higher consistency for the traditional algorithms. In fact, Infomax obtains the highest maximum, mean, standard deviation, averaged across runs. The minimum distance is also the lowest for Infomax. rICA derives the lowest AE, but is very close to the value for Infomax. Notably, all four algorithms perform well. Although the traditional algorithms provide the optimal statistical consistency, the novel unconstrained approaches don't fall very far behind. Overall performance is better using the simulated rsfMRI data through SimTB when compared to the simulated SNP and sMRI data.

Figure 4.13: ICASSO graphs of CCA cluster similarity projections for simulated rsfMRI using FastICA (top left), Infomax (top right), rICA (bottom left) and sfICA (bottom right) at $n = 100$.

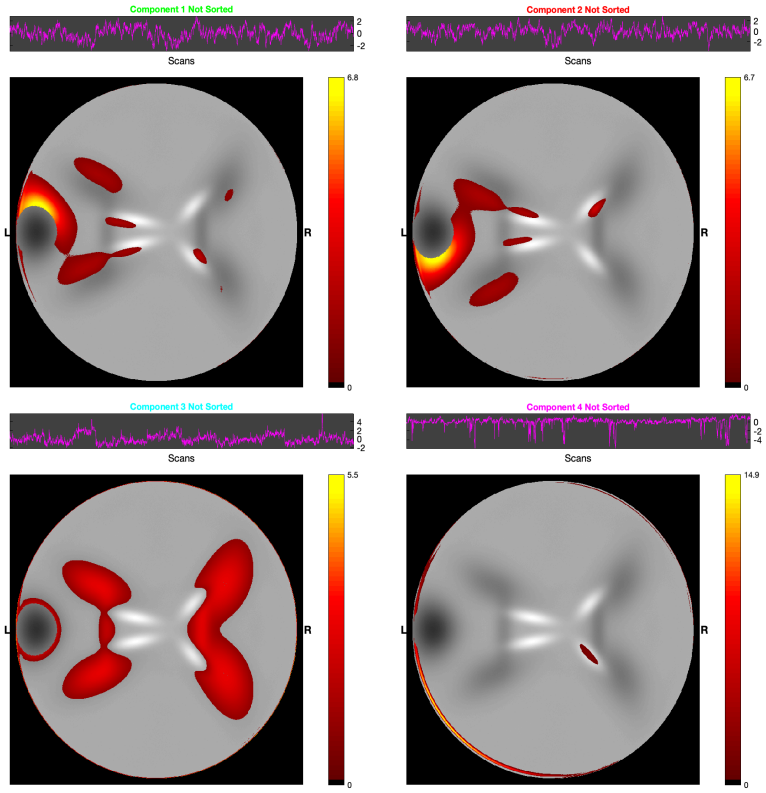


As mentioned, feeding the simulated NiFTI formatted data into the gICA toolbox in Matlab allows one to compare the simulated sources with the ground truth. Recall that each component should correspond to a different anatomical

Table 4.3: Summarization of ICASSO for simulated rsfMRI data, n=100.

Algorithm/Measure	Max	Mean	Sd	MD	Amari
FastICA	0.9963	0.9841	0.0150	0.7722	0.4131
Infomax	**0.9968**	**0.9901**	**0.0057**	**0.7712**	0.4004
rICA	0.9887	0.9436	0.0327	0.7843	**0.3987**
sfICA	0.9856	0.9194	0.0359	0.8251	0.4359

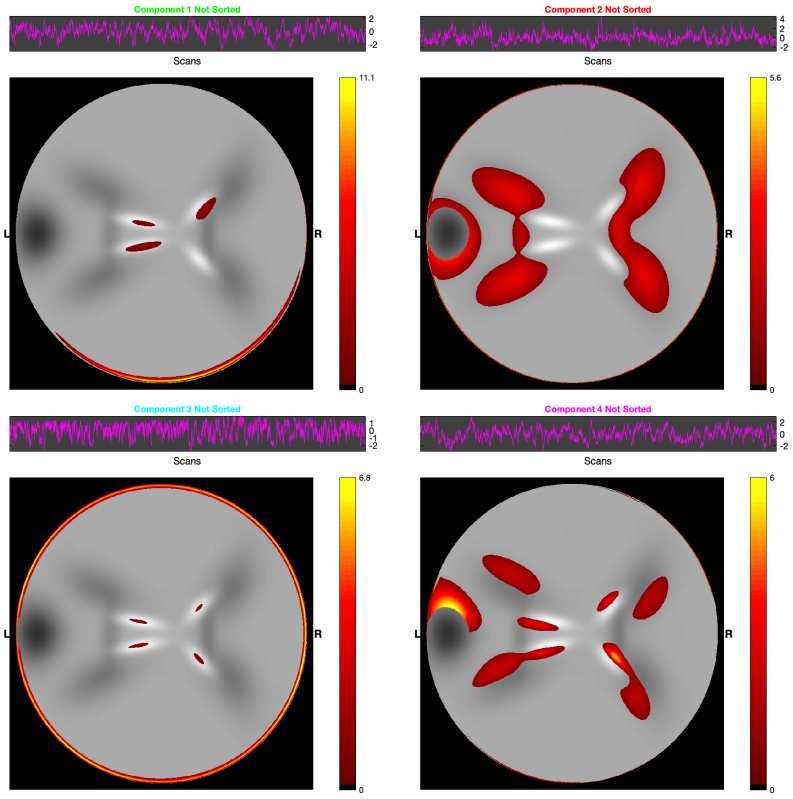
Figure 4.14: Simulated rsfMRI brain sources recovered from GIFT for n=100 using rICA.



region or, in the simulation case, a replication of this region. Figures 4.14 and 4.15 show the recovered "brain" sources for the rICA and sfICA algorithms, respectively. The ordering of the components is not in any particular order (hence, "not sorted"). For example, rICA seems to have the strongest visualization of a source in component 3, while sfICA has a very similar looking recovery in component 2. One may expect the intra-cluster similarity to be high for these respective components, and this is confirmed by the results with an I_q of 0.9887

for the first and 0.9856 for the second, equal to the max cluster index for the algorithms. This confirms that the intensity of the cluster similarity has a direct and meaningful anatomical significance. The graphs above the simulated source images in Figures 4.14 and 4.15 (the purple lines) are the z-scores of the recovered signals, IC loading weights, corresponding to each source image.

Figure 4.15: Simulated rsfMRI brain sources recovered from GIFT for n=100 using sfICA.



Finally, it has not been fully explored how sample size and the particular data source can impact the stability of the IC estimates. Although imaging and genetic data are both high-dimensional, the dimensions vary from study to study and the covariation within the data may be highly diverse from one modality to another. In addition, it is necessary to explore whether the stability properties are different for the various algorithms or may be taken as general properties of BSS using a decomposition analysis. See Table 4.4 for a summary of these values, providing the mean $\max(I_q)$ for each algorithm, sample size and simulated data set. Recall, rsfMRI was only considered at $n = 50$ and $n = 100$. The SNP data shows consistent values regardless of sample size for FastICA and Infomax, with a similar occurrence in the sMRI data. However,

rICA and sfICA seem to be influenced by the sample size, as we see a gradual inflation of maximum values of I_q with an increase in sample size. There is a slight difference in the cluster similarity for rsfMRI data, but the general result for this data source is that the IC components are very closely recovered from the ground truth. Overall, the SNP data showed the best perform for rICA regardless of sample size. This is also true for the sMRI data. However, there is a notable jump to 0.64 for $n = 300$ of the SNP data. The next section will explore how the algorithmic and statistical reliability will improve with combinations of these modalities.

Table 4.4: Mean $\max(I_q)$ using ICASSO compared among the algorithms for all simulated data and sample sizes.

Algorithm/Data	n	**SNP**	**sMRI**	**rsfMRI**
FastICA	20	0.61	0.58	
	50	0.60	0.59	0.98
	100	0.59	0.59	1.0
	300	0.60	0.58	
Infomax	20	0.57	0.58	
	50	0.57	0.60	0.99
	100	0.59	0.60	1.0
	300	0.58	0.61	
rICA	20	0.60	0.61	
	50	0.64	0.73	0.97
	100	0.66	0.80	0.99
	300	0.67	0.80	
sfICA	20	0.45	0.43	
	50	0.45	0.53	0.96
	100	0.57	0.58	0.99
	300	0.64	0.57	

4.2 Analyses of Single ADNI Modalities

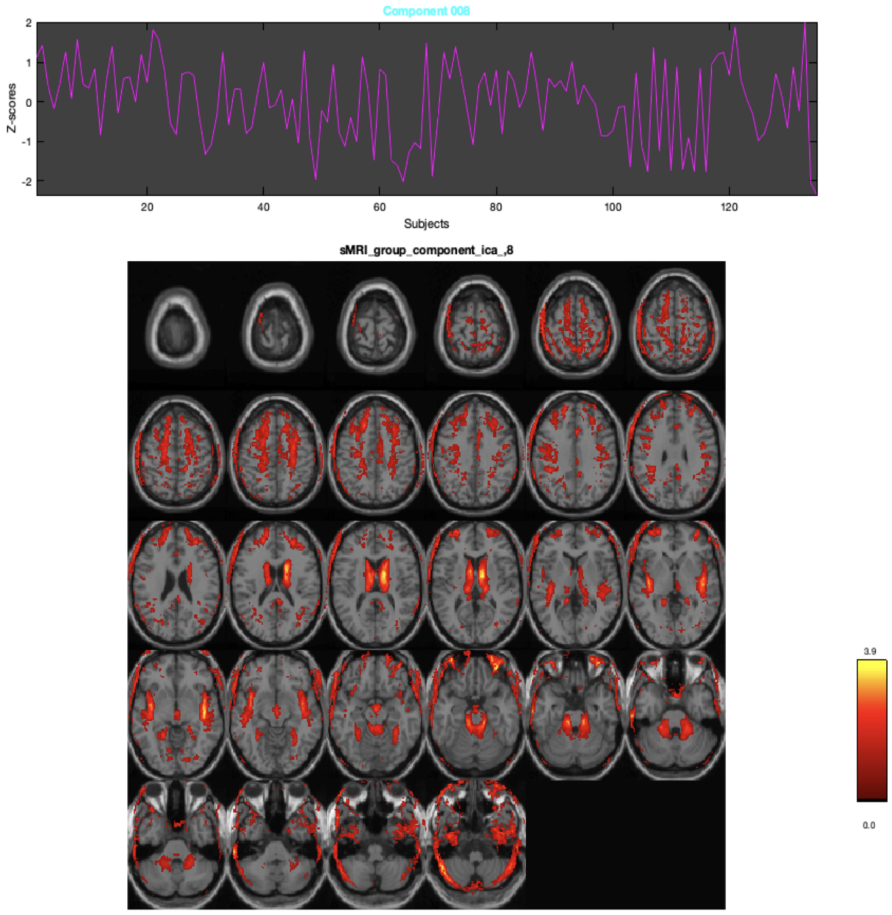
Next, I present ICA analyses of the single ADNI data modalities in detail to reveal the importance of including these data sources in a multi-modal analysis. For all cases, I summarize the stability of the algorithms as well as the statistical reliability. Visualizations are provided to further understand the biological significance of the modalities. Real-life AD data were downloaded from the ADNI database after obtaining access by committing to the terms of the Data Usage Agreement and continuing with this responsibility each year. Data are cleaned, processed and matched across subject IDs and brain-dimension sizes from ADNI1 (n=135) and ADNI2 (n=74). The diagnostic categories under ADNI1 are CN, MCI and AD, while the latter protocol splits the MCI category into EMCI and LMCI, and the healthy patients are either exclusively CN or are in the Significant Memory Concern (SMC) category. The purpose of this section is to showcase the use of these data sources in a decomposition analysis and to compare these outcomes using FastICA, Infomax, rICA and sfICA. One interesting aspect of this research is the ability to visualize the biological significance of the ICA approaches. While this is not so obvious in the genetic information at the SNP level, one can observe the direct relation of decomposition to brain structure and function through the extracted components. Thus, the goal of this section is also to elucidate the anatomical interpretation of the neuroimaging modalities using a novel optimization approach.

Structural MRI

Past research studies repeatedly show that the anatomical atrophy that occurs in those with MCI and AD tends to localize in the hippocampus, found within the temporal lobe of the brain. This fact confirms the classical knowledge that the medial temporal lobe houses the neural connections made during learning and memory tasks. Changes in cortical thickness and volumes have been identified by researchers in other areas of the brain, such as the parietal and frontal lobes. This structural deterioration tends to occur even prior to advanced stages of AD, thus motivating the inclusion of sMRI data into the analysis Gupta et al., 2019; Marenco and Radulescu, 2010. In order to assess the full-scale of the anatomical implication, voxel-wise brain volumes covering the whole brain are considered.

When ICA is applied to structural MRI data, the method is called structural-based morphometry (SBM), where the observations are at the subject-level and the sources recovered are weighted anatomical volumes. In this case, we apply voxel-based group ICA with the goal of quantifying the covarying gray matter brain patterns (Gupta et al., 2019; Xu et al., 2009). We may visualize these

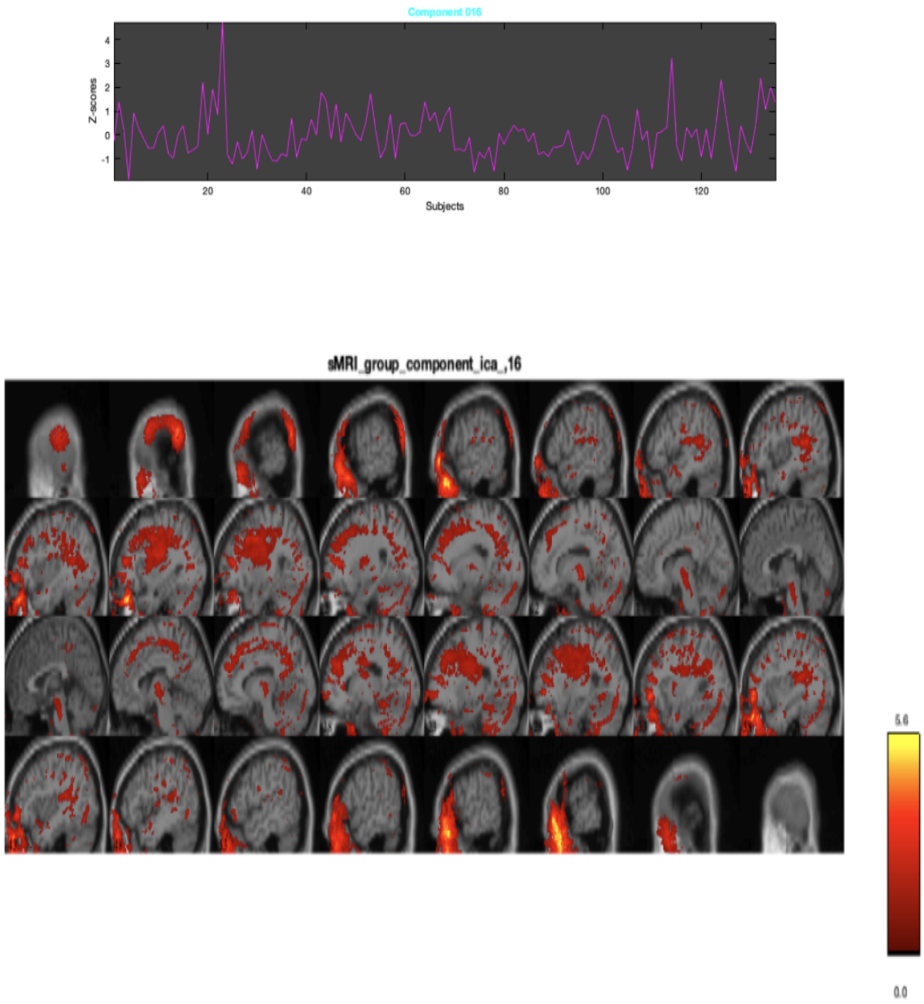
Figure 4.16: Components and their relation to anatomical regions of interest averaged over subjects of all diagnoses.



extracted components to see whether they are representative of brain regions of interest or structural brain networks (Calhoun & Sui, 2016; Qi et al., 2019). Two plots are seen in Figure 4.16. The top is the signal of the 8th IC, normalized as a z-score across all subjects of AD1, using the basic FastICA approach. One may imagine that the outcome of an IC model is a matrix of reconstructed features that can be viewed by overlaying the values onto the spatial structure of the brain over 28 slices of the brain from top to bottom. Thus, Figure 4.16 displays the relation of the column-wise features of the reconstructed data matrix to the anatomy of the brain. This overlay is averaged over the disease group of AD, displaying the congruous anatomical features across subjects and the disease

status. It's also possible to view these results in the sagittal view, from the side, left to right. This view is shown, with the same AD group represented, in Figure 4.17 for component 16. Notice the much stronger signal recovered in this second plot. Furthermore, this is run from the command line using the new rICA approach.

Figure 4.17: Sagittal view of sMRI ADNI1 data.



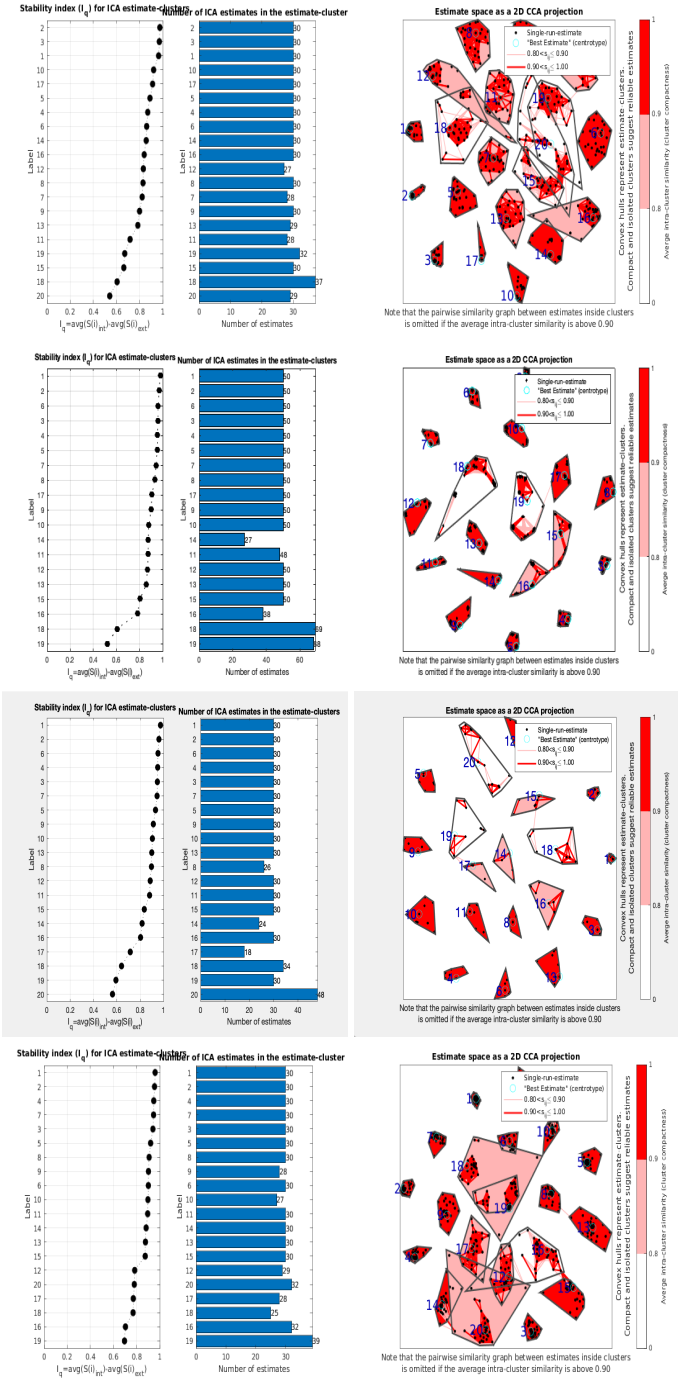
All four algorithms are applied with gICA, where the groups are separated by diagnosis status. Group comparisons are possible, but are not included until the multi-modal approaches are introduced. Using MDL, it is recommended to set the dimension of IC to 20 for restructuring of the sMRI data, thereby significantly reducing the large dimension of the data. Graphs of the stability indices (left) and CCA projections of the clustered components over 30 ICASSO runs

(right) are realized in Figure 4.18. The values in Table 4.5 reveal the maximum, mean and standard deviation of the I_q values for the 4 algorithms for this sample of ADNI1 subjects. The best values of the measures are presented in bold. sfICA performed the best with a mean and max higher than the rest of the methods. FastICA has the second highest max, however, it has the lowest mean and the highest standard deviation which shows that it has the lowest stability overall with a quick drop in performance after the first three components. Infomax has the second highest mean, with a very similar performance to sfICA. However, most of the clusters in the similarity graph of sfICA are dark red showing intra-cluster similarity measures between 0.95-1.0. rICA shows nearly identical results to Infomax; however, a dip in the results by the 14th component pulls the mean lower and inflates the variability of the cluster distances. It is worth noting that this shows considerably different, but also expected, outcomes from the simulation results of the sMRI data in the previous section. With the simulated data, rICA showed the best results and sfICA had the lowest. It appears that now rICA still performs well, but imposing sparsity seems to perform better for the larger dimensions.

Table 4.5: Summarization of ICASSO results with ADNI1 sMRI data.

Algorithm/Data	Max	Mean	Sd
FastICA	0.9757	0.8170	0.1214
Infomax	0.9590	0.9150	**0.0498**
rICA	0.9609	0.8664	0.0839
sfICA	**0.9766**	**0.9399**	0.0526

Figure 4.18: ICASSO results for structural MRI ADNI1 data for FastICA, Infomax, rICA and sfICA.



FDG-PET Uptake

Coupling the measurement of FDG-PET with the clinical diagnosis of patients have shown a stronger validation of AD than sMRI in recent literature (Vogel, 2020; Zhou et al., 2017). This is a change from the original belief that anatomical atrophy is one of the earliest steps in AD decline, typically occurring at the prodromal stage of the illness. However, functional decline may occur even earlier. It is now known that the presence of decreased levels of FDG correspond to a reduction in the brain metabolism of both glucose and oxygen, which is very common in AD patients (M. W. Weiner, Veitch, Aisen, Beckett, Cairns, Green, et al., 2017). Although FDG-PET is a functional measure, this is considered in a volumetric format for the sake of the analysis. This is because the ADNI Pet Core takes the average FDG uptake for each voxel in the brain. Therefore, it is not a value considered on a time scale as functional MRI data are typically done. The advantage or disadvantage to using this modality across the studies is that PET scanning has advanced and, therefore, changed drastically throughout the years of ADNI data collection. More recent studies involving PET scans focus on amyloid beta or tau-protein measurements, which may serve as surrogate endpoint biomarkers in the future. For this research, the PET data collected in ADNI1, for 135 subjects, are now analyzed using the 4 algorithms and assessing the statistical and algorithmic consistency together.

Similar to the presentation of the sMRI data, I present a select visualization obtained from the four different methods. These are run in Matlab using the SBM toolbox within GIFT. The images seen in Figure 4.19 are component 6 signals using the Infomax algorithm, run with the groups separated by disease status. Notice the color bar shows values ranging from 0 to 6.3. If the color indicated in the plot represents positive values, and the group mean of the loadings of that group is higher than the other group, then this means that the first group has a greater uptake of FDG, or the metabolism of glucose and oxygen. Since decreasing amounts of FDG are known to occur throughout the progression of AD, we expect larger positive quantities to be indicative of the healthy group and negative values to occur for AD. The yellow emerging towards the bottom and back portion of the brain in sagittal view shows the greatest amount of FDG uptake in healthy individuals. These results confirm what one will find in the literature between disease status and FDG uptake as well as with the general location of the greatest amount of uptake taking place in the cerebellum, which is responsible for muscular and metabolic function (Marcus et al., 2014).

Finally, I examine the overall replicability of IC features by assessing the cluster compactness of IC estimates over 30 ICASSO runs for all four algorithms on the real data. Table 4.6 and Figure 4.20 display the ICASSO results. The

optimal number of components run is 25, so each model considers this many components. The highest average maximum over the ICA runs occurs again at rICA. However, sfICA and Infomax are tied with the mean I_q, with very low and similar standard deviations. That being said, all four algorithms performed well. The highest cluster similarity (or lowest intra-cluster difference) goes to Infomax as it is consistently 0.98. This offers an improvement even over the performance of the others algorithms on the sMRI data. In general, this can mean that there are higher separability among the disease groups for FDG-PET, or the data are less complex. However, this question is beyond the scope of this dissertation.

Figure 4.19: Close-up of FDG-PET results using sfICA

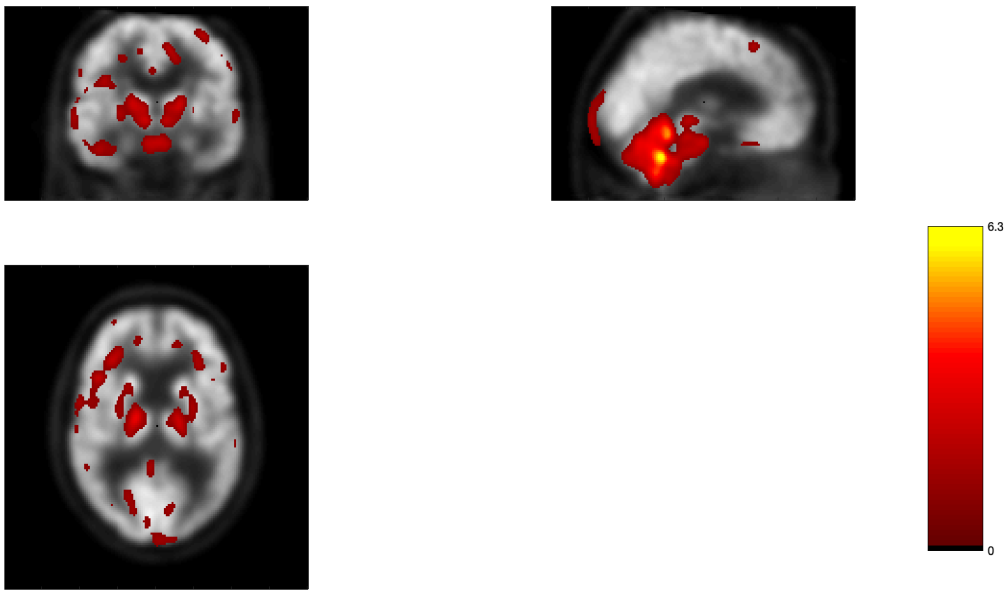
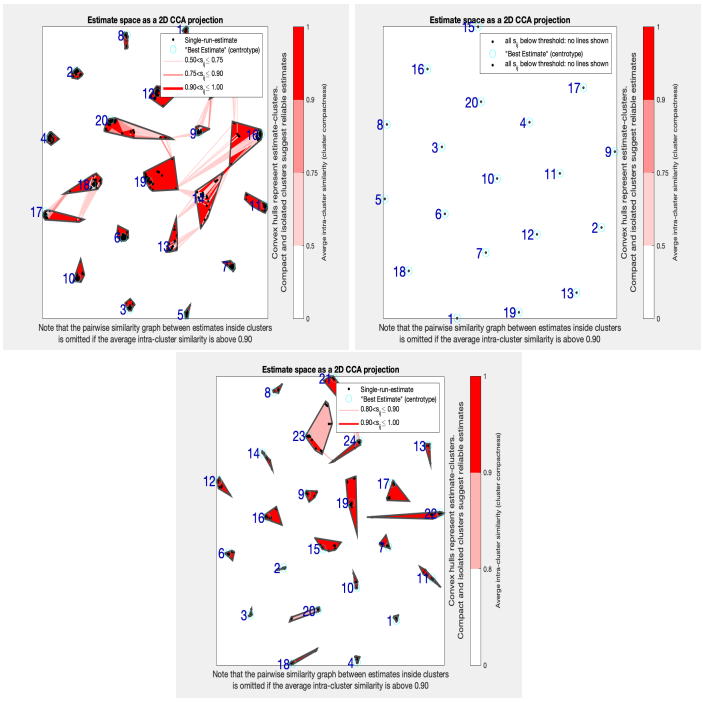


Table 4.6: Summarization of ICASSO results with ADNI1 FDG-PET data.

Algorithm/Data	Max	Mean	Sd
FastICA	0.991	0.9538	0.0443
Infomax	0.9836	**0.9811**	0.0032
rICA	**0.9964**	0.9808	0.0070
sfICA	0.9869	**0.9811**	**0.0031**

Figure 4.20: ICASSO results for structural FDG-PET ADNI2 data.



Resting-state fMRI

While FDG-PET measures the metabolic function in the brain, the average uptake is a volumetric measurement and is therefore analyzed using structural approaches, that is, with SBM. Additional advances in MRI technology have permitted the use of resting-state fMRI (rsfMRI), allowing a snapshot of the dynamic functional process in cortical regions of the brain. With this data source, the values in the data represent the hemodynamic BOLD response as a function per voxel in the whole brain. This is called resting-state because the measurements are taken in snapshots over time in the absence of any task- or event-related actions. When performing gICA on rsfMRI data, this is considered to be spatial ICA since the voxel-wise hemodynamic flow is evaluated as a function of time, revealing the natural dynamic brain activation. The decomposed and reconstructed features are functional connectivity pathways that are related to differences in the natural flow of blood through the brain. Furthermore, the ICASSO process clusters these pathways into consistent representations of the pathways that may be compared from data set to data set. This process is called dynamic functional connectivity (dFNC) and is run as a post-processing step after gICA is performed.

Figure 4.21: Resting-state fMRI brain plots for ADNI2 data: averaged signals over all subjects shown in 43 slices of the brain for two components.

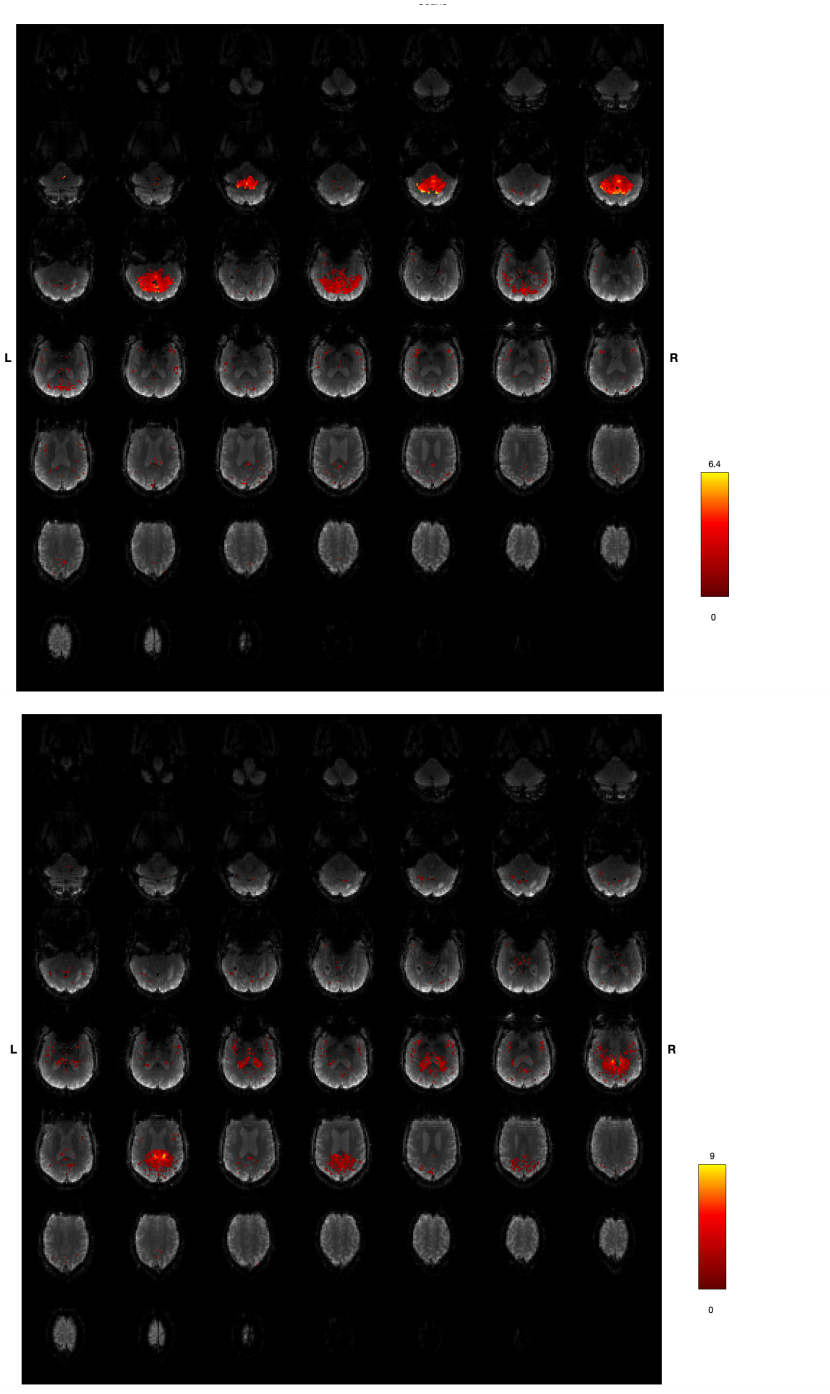
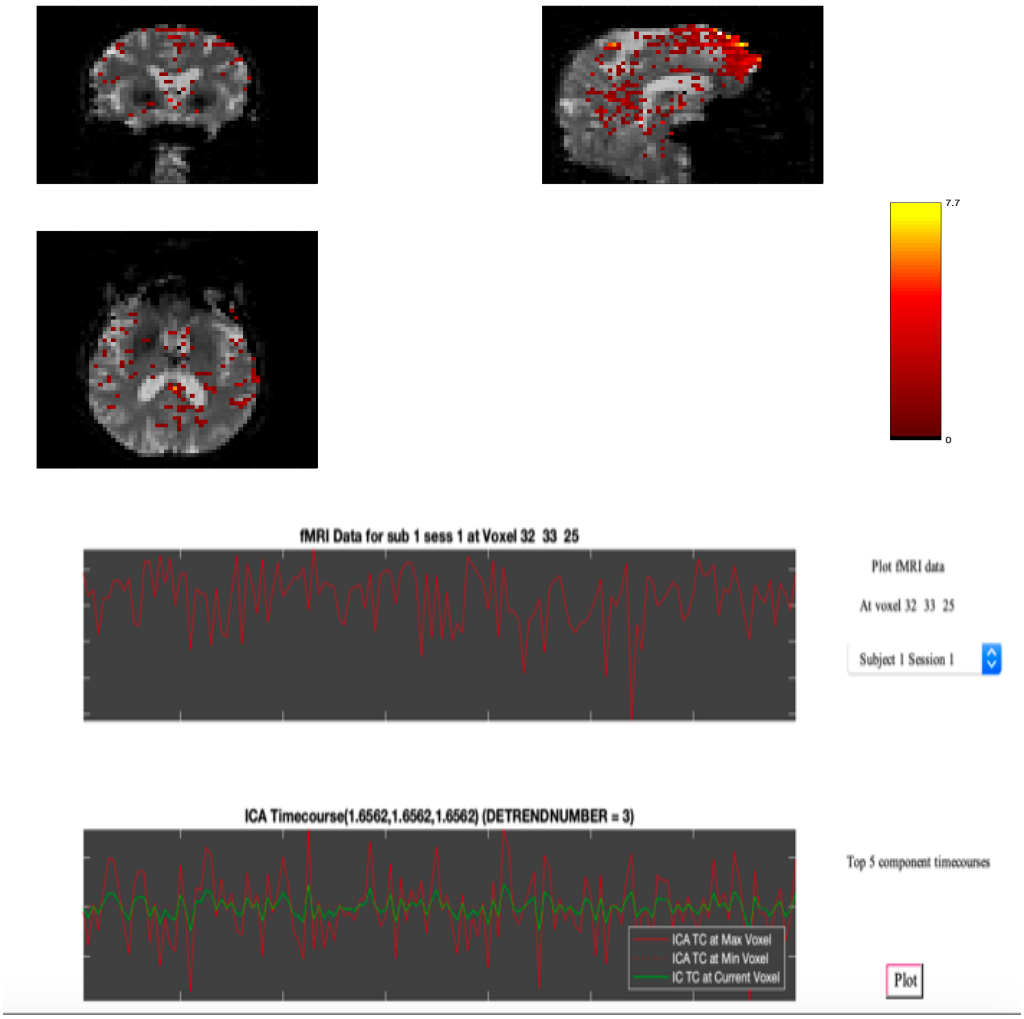


Figure 4.22: Brain images and associated ICA timecourse and IC signal for subject 1 at timepoint 1.



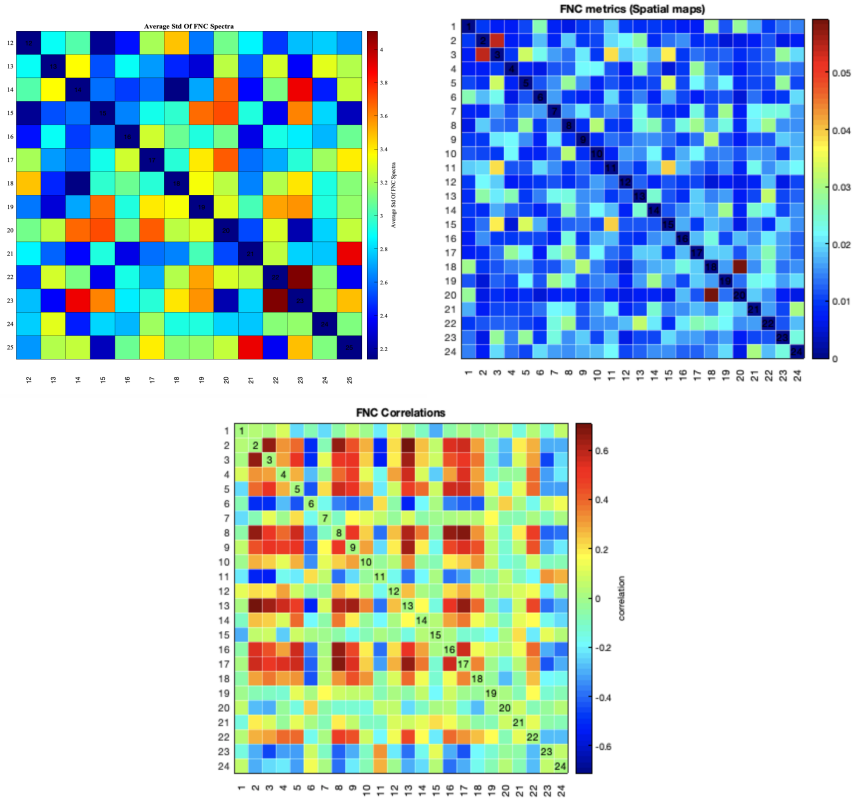
Anatomical plots of ADNI2 rsfMRI data may be seen in Figure 4.21. Recall that fMRI provides the bold-oxygen-level-dependent (BOLD) levels of the brain, which is directly tied to activation in the brain. Although task-based fMRI has been more common in the past, researchers are using this modality more and more in order to see the natural, or resting, flow through the brain. Therefore, using this modality in an analysis for AD will help researchers understand whether those with cognitive impairment have a lower level of blood

flow, as suggested in the literature. The brain images seen in this figure show the signals of two ICs averaged over all individuals in gICA across 43 brain slices, starting from the bottom of the brain and working up to the top. The highlighted areas in red are the components overlaying the brain that showcase whether there are functional differences among this ADNI2 cohort. The first image is from component 2 and the second is from component 10. Notice the scales are different for the two components, indicating that the different features are able to read different levels of blood flow. In addition, the slice location in the brain shows stronger activity lower in the brain for component 2 than for component 10. Consequently, the light orangish-yellow color shows that one of the groups has relatively higher blood flow than another. Given the high magnitude in the second plot, this is more strongly indicated in component 10. Figure 4.22 shows the sagittal, axial, and coronal planes of one subject with the associated source signals below for voxel location $(x, y, z) = (32, 33, 25)$. The sagittal view showcases activity in the frontal lobe with the strongest magnitude of 7.7. In addition, the axial perspective displays midbrain activity. The two signal plots below the brain plots in Figure 4.22 reveal the original fMRI signal (top) and the corresponding ICA timecourse (bottom), overlaid with the maximal voxel signal. This showcases the ability of ICA to provide subject-specific spatial and temporal information.

Research has developed a lot in the past 5 years or so in the areas of discovering the functional pathways in the brain that are related to certain neuropsychiatric illnesses. When functional pathways may be highly correlated with pathways in other locations in the brain, this defines a potential network in the brain. The process of discovering these linkages is called dynamic functional network connectivity (dFNC) and may be run as a post-processing step for rsfMRI data Damaraju et al., 2014; Iraji, 2021. The idea of finding networks within the brain is similar to the concept of ICASSO, but is applied at the voxel-level rather than the component-level. If you cluster the extracted features using cluster analysis, you may uncover networks. With the GIFT package in Matlab there is a dFNC post-processing step. After performing ICASSO with 30 repetitions, the rsfMRI features are grouped through the clustering analysis toolbox using k-means. This identified 24 pathways that are linked among each other after taking 100 repetitions. Heatmaps of the average standard deviation of the FNC spectral frequency and the corresponding spatial and correlation maps of these networks may be seen in Figure 4.23. The graph on the top left shows the average variation of the 24 spectra, revealing highest levels around pathways 19 and 20; the spatial maps on the top right correspond to the spatial distance among the pathways, highlighting frontal and midbrain activity; and

the bottom graph reveals the correlation maps among these pathways, captivating pathways from frontal to midbrain with positively strong relationships spread throughout.

Figure 4.23: Spectra standard deviation (left), spatial (right), and correlation (bottom) maps of 24 pathways recognized by dFNC using ADNI2 rsfMRI data.



In the literature, the idea of simulating or bootstrapping an rsfMRI data set is quite new, probably due to the fact of the recent boom in the use of this imaging modality. An analysis of the reliability of this data under the comparison of different ICA algorithms may offer a novel concept for the data. In Figure 4.24, I reveal the plots of the stability indices, the 2D CCA projectsions of the clusters and the dendrogram/similarity plots. I chose to showcase these two algorithms because they provide the most similar results and will offer an interesting comparison between constrained and unconstrained algorithms. The performance begins quite similar, but is higher for Infomax, with the decline a bit slower for sfICA. The biggest difference is among the last 5 ICs. Here, we see a quicker drop-off for sfICA. However, according to the similarity matrices, we can see that sfICA offers less off-diagonal similarity (a good thing). The lower standard deviation is probably attributable to this fact. With the

exception of FastICA, the algorithms provide great consistency. However, the mean is considerably lower with a higher standard deviation for the FastICA algorithm. This is interesting, because FastICA is most commonly used for fMRI data.

Figure 4.24: ICASSO results for rsfMRI ADNI2 data for Infomax and rICA.

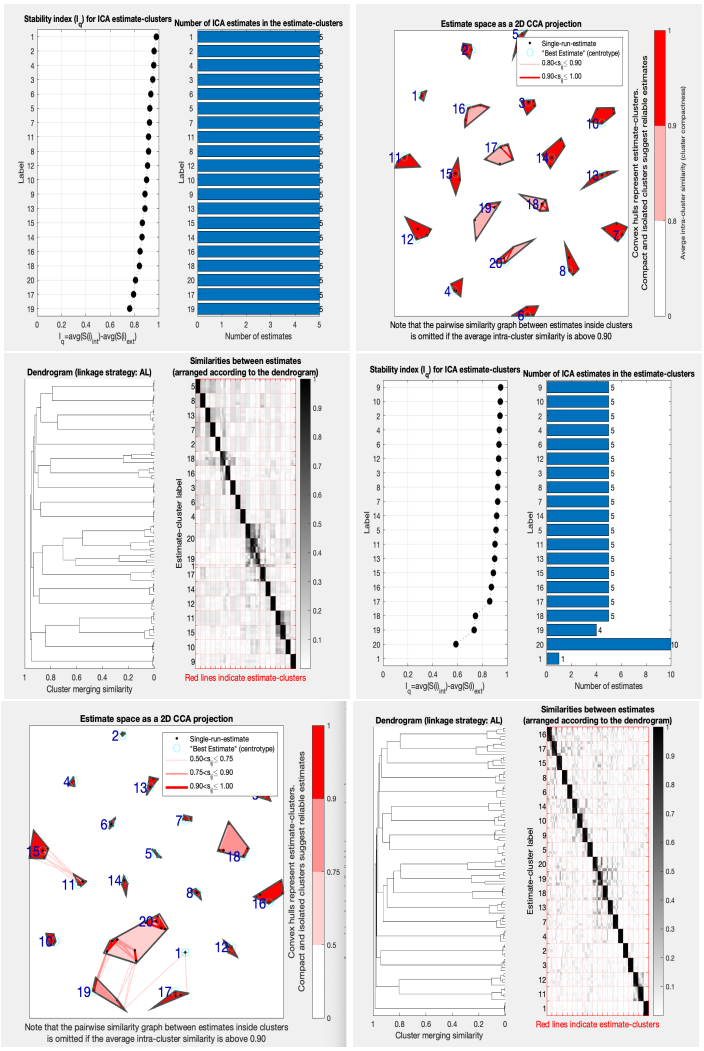


Table 4.7 shows the ICASSO results for ADNI2 rsfMRI. The maximum compactness is reached by FastICA, while Infomax has the highest mean. While one may presume that the traditional algorithms outperform the novel algorithms, keep in mind that the mean of rICA is close value to Infomax and also acquires the lowest standard deviation. Therefore, we may conclude that the real-life rsfMRI data are consistently optimized by both Infomax and rICA. In fact, rICA has the lowest variability, and sfICA also has a lower standard deviation than FastICA and Infomax, with maximum and mean I_q over 90%.

Although FastICA has the largest maximum, it shows the least stability for subsequent components. This reveals that for a higher-dimensional data set, unconstrained algorithms show smaller fluctuations across repetitions, while still achieving high statistical reliability in terms of the mean and maximum compactness. The novel algorithms show comparable results to the traditional algorithms for these data.

Table 4.7: Summarization of ICASSO results with ADNI2 rsfMRI data.

Algorithm/Data	Max	Mean	Sd
FastICA	**0.9797**	0.8911	0.0590
Infomax	0.9789	**0.9494**	0.0367
rICA	0.9590	0.9339	**0.0158**
sfICA	0.9468	0.9264	0.0198

SNP

The SNP modality contains information on the most granular form of genetic material for the subjects in the study. Unlike the neuroimaging material, this measurement is static and obtained only at baseline through a blood test. After pre-processing is done, as described previously, the data are uploaded as string of four possible alleles, A, T, C or G. Each genetic coding is representative of either 0, 1 or 2 minor allele expression. Thus, the data are uploaded in PLINK in DNA letter form, then converted to the minor allele frequency coding, thereafter pre-processed as discussed earlier in this work. In this section, I will provide a brief GWAS result on ADNI data and then express the need for ICA for this data source.

The subjects are coded as having MCI with a 0 and AD with a 1, and a GWAS is performed to tell if there are associations between the genetic coding and the presentation of the prodromal or full AD diagnosis. A Manhattan plot allows one to visualize the performance of association tests for diagnosis status and also points to the multiple comparison problem. Figure 4.25 reveals the Manhattan plot, colored by chromosome and displaying the results of $-\log_{10}(p)$. The p-value threshold is the blue line, with the dots above this indicative of significant SNPs. Of the genome, only 4 SNPs are recognized as statistically significant: rs2075650, rs11595021, rs7157639 and rs4886844. Table 4.8 names the chromosome position, major and minor alleles, the corresponding genes and how many publications exist listing the impact of the rs coding. The SNP with the lowest p-value, rs2075650, corresponds to the TOMM40 gene, which was used in the preliminary analyses in Chapter 2. This is an exciting confirmation of the role of TOMM40 gene in the progression of AD.

Indeed, this SNP has been listed in 132 publications which include confirmed associations with aging and brain trauma. However, the other three SNPs listed in Table 4.8 have little to no previous indication of relevance in past GWAS. It remains unknown whether these findings recover novel SNPs important in AD research or whether these SNPs are false positives. In addition, it is difficult to make group comparisons at this level of granularity.

SNP data are more easily compared in component format as the data are scaled and transformed to a continuous measure that may be compared by groups and/or other modalities. Graphically, the SNP results are not as useful as the imaging results. However, the magnitude of the weights remain interpretable. Contrast the GWAS strategy with the output of component 6 of the SNP modality, as seen in Figure 4.26. The signal oscillates and is challenging to read, but keep in mind that one SNP component summarizes genetic information across subjects. Rather than pinpointing one granular association, we can use the weighted loadings of the components in downstream analyses or combine with imaging data to study the other impact of MMNG applications at the subject-level.

Figure 4.25: Manhattan Plot identifying significant SNPs (p-value$<$ .01) between the MCI and AD disease categories.

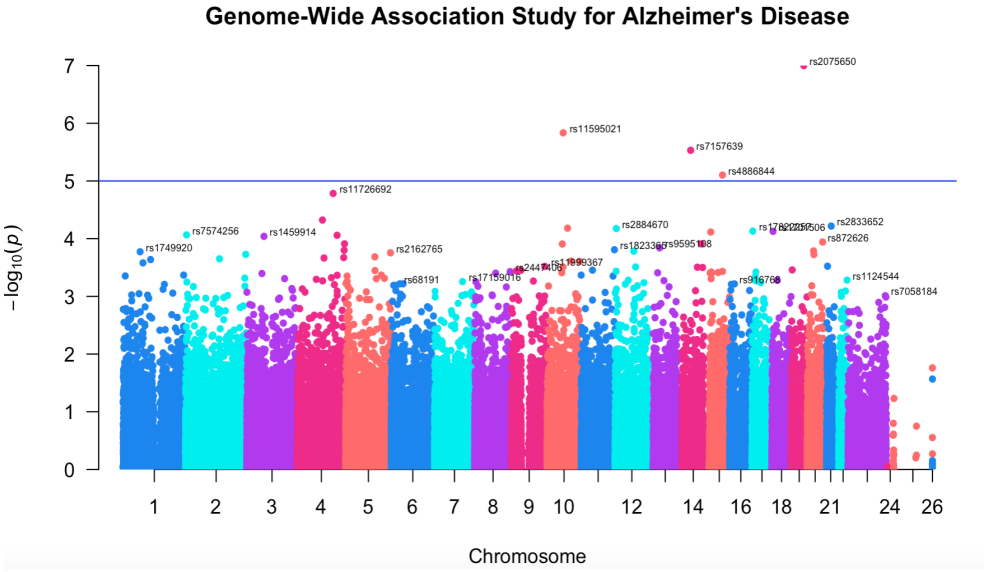
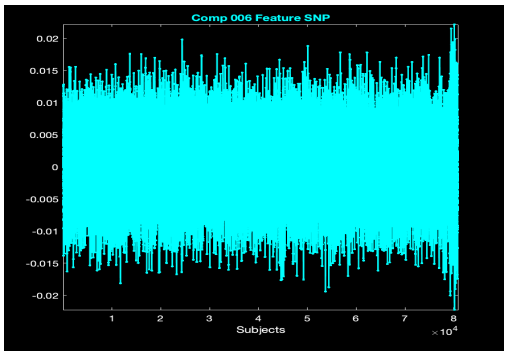


Table 4.8: Reference SNP (rs) Report on 4 significant SNPs from the MCI-AD GWAS (https://www.ncbi.nlm.nih.gov/snp/).

SNP Name	Position	Alleles	Gene	Publications
rs2075650	chr19:44892362	A>G	TOMM40	132
rs11595021	chr10:62903510	C>A	LOC107984012	0
rs7157639	chr14:52921443	C>G / C>T	FERMT2/LOC105370500	0
rs4886844	chr15:74016599	T>C	PML	1

Figure 4.26: SNP component output for the first 8 subjects.



4.3 Multi-modal Decomposition

The analyses so far have compared algorithms using only single modalities, one at a time with the various ICA algorithms under consideration. The cluster compactness has been explored and interpreted as information about the reliability and replicability of these methods. The ultimate goal of this study is to analyze these same algorithms under a multi-source scenario in order to understand the benefits that one gains from including multiple modalities in the analysis. Another desire is to explore the impact of using algorithms that are constrained vs unconstrained and linear vs nonlinear. In this section, I present the results of using the traditional methods of FastICA and Infomax with the newly implemented rICA and sfICA on mCCA+jICA and p-tIVA+mCCA.

The ADNI data have been collected from ADNI1-3 for the purpose of modeling the multi-modal scenario. I consider two two-way combinations, sMRI+SNP and FDG-PET+sMRI, and one three-way combination, FDG-PET+sMRI+SNP. For both the two- and three-way data fusion, the correlation will be examined among and within the modalities for the ICs and compared across methods. For three-way data fusion, group comparisons for disease statuses and ADNI protocols will be made both in terms of the mixing compo-

nents ICs within the modalities and then between modalities by comparing ICs of the same index. The results of the t-tests for each of these methods are be presented in tables. If components of the same index show group differences in more than one modality, then it is considered to be a joint group-discriminative IC. However, if it shows only significance within the modality, then it is a modal-specific group-discriminative IC that is telling of the underlying biology within that one modality.

4.3.1 Two-way Imaging Genetics

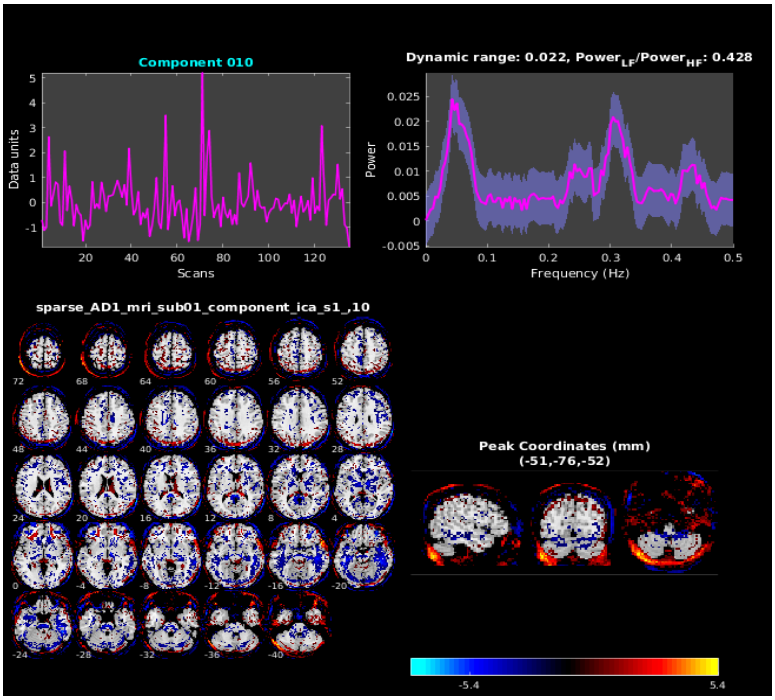
In this section, two-way MMNG analyses are considered. After mCCA is run on the extracted PCs and whitening is performed, then jICA is applied. The benefit of mCCA is that it takes into consideration the covarying relationships among and within modalities before concatenation. Each modality runs the same number of principal components, otherwise we would have unequally weighted correlation vectors corresponding to the different modalities. This is chosen by a maximal variance accounted for in the individual data sources, thereby utilizing the first 29 PCs per modality. The linear and nonlinear optimization results are explored using ICASSO and disease output.

Two-way Fusion With & Without SNP Data

First, pairwise MMNG is considered in order to assess the validity of results between two neuroimaging modalities (FDG+sMRI) and the sMRI modality with genetic information (sMRI+SNP). All four algorithms are tested using ICASSO with 30 iterations, and the components are sorted in descending order by the cluster compactness index, I_q. For jICA, recall each modality must have the same number of components. When tested alone, MDL selected the optimal number of components to be 136 for SNP, 25 for FDG and 20 for sMRI. Thus, for the pairwise data fusion, sMRI+SNP ran 20 components. However, when running FDG+sMRI, I saw significant declines in the cluster similarities around about 15 ICS, so I reduced this to 15 instead of 20. In order to compare the statistical consistency of using the two combinations of modalities, I provide results based on the first 15 extracted ICs of each study.

Under the MMNG analysis using sMRI+SNP data, the interpretation of the reconstructed features is now somewhat different for the mCCA+jICA framework. Indeed, the cross-correlations among and within modalities are considered in the mCCA step of the model. Since these features are the inputs to the fusion step, jICA, the extracted features are scaled and fused vectors representative of the complex relationships between the sMRI and SNP modalities,

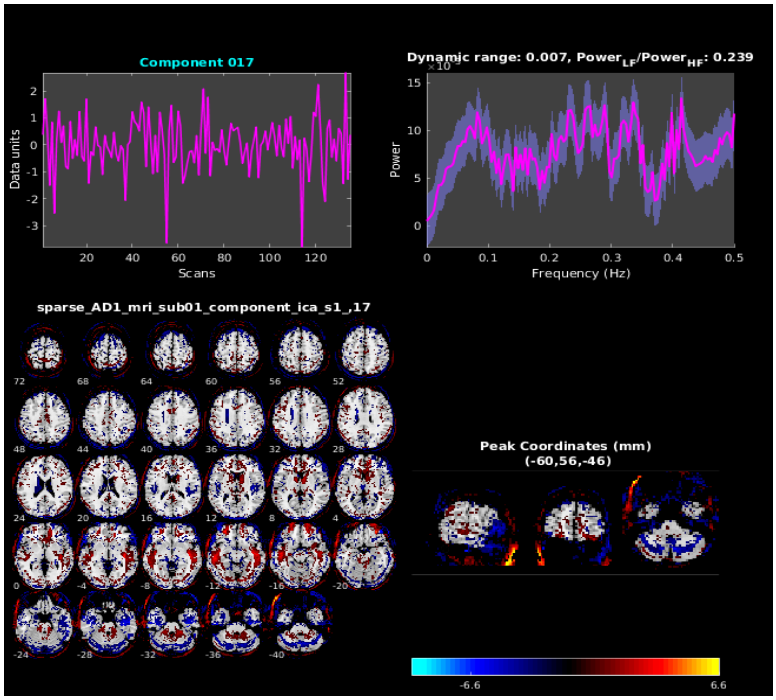
Figure 4.27: Component 10 of sMRI using sfICA for sMRI+SNP combination.



while also controlling for the intra-modal correlation. Similar to the concept of regression, the ICs specific to the sMRI modality are averaged over the correlation with the SNP modality, or holding this constant. Figure 4.27 shows the output of the sMRI components for component 10. Notice that with the sMRI data, we have a scale from blue to yellow on the interval of (-5.4, 5.4). It is well known that loss of grey matter volume in the cortex occurs in MRI and AD patients (M. W. Weiner, Veitch, Aisen, Beckett, Cairns, Green, et al., 2017). This shows up as a negative value in the color bar. Therefore, the areas in blue are locating the particular regions of the brain where AD/MCI patients experience the most prominent atrophy. Consequently, the red/yellow shows where the CN group has higher volume than the MCI/AD patients, or similarly the MCI patients having greater cortical volume or thickness than AD patients.

Two-way fusion of mCCA+jICA is then performed on two imaging modalities, FDG and sMRI. In the previous section, both modalities were run separately using the same technique of SBM. In this section, we may compare the results that are obtained when combining this information. AD progression is characterized both by reduction in gray matter as well as reduction in FDG

Figure 4.28: Component 17 of sMRI using sfICA for sMRI+FDG combination.



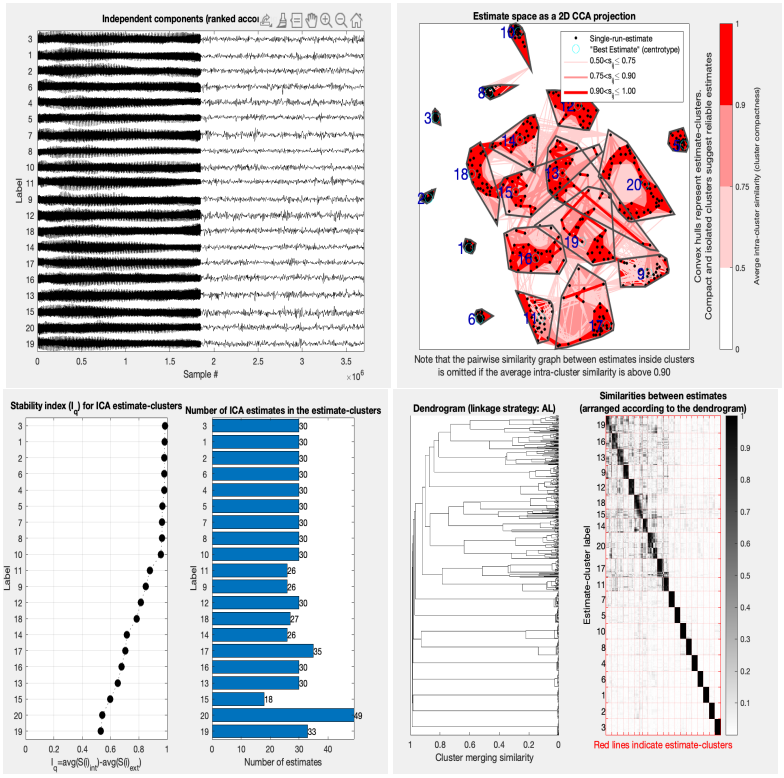
uptake. Thus, it is expected that there is a positive correlation among these quantities. However, it is important to point out the non-intuitive interpretation of FDG+sMRI fusion. Since sMRI and FDG uptake are both expected to decline in those undergoing the progression of AD, it is less obvious whether these reductions occur in similar locations spatially. That is, the extracted features elucidate the magnitude of the correlation within overlapping brain regions. Thus, two-way fusion of two volumetric modalities is going to reveal the structural pathways within the brain. In other words, each feature represents the maximal spatial correlation between brain atrophy and glucose metabolism.

Figure 4.28 shows the sMRI results when data fusion is run between sMRI and FDG on the same subject as before. The components that are overlaid on the brain are representative of the differences in disease status observed in gray matter structure, adjusted for the impact (or weight within the loading vector) of FDG within that spatial region. The scale ranges from (-6.6, 6.6) where the right end of the scale, or the red color, that we predominantly see in this output showcases locations of increased volume in gray matter for CN patients. Notice this range is wider than it was for the sMRI+SNP modality. This means that we

can observe a greater impact of AD in the structural regions of the brain when adjusting for the impact of FDG. The blue regions are then areas that apply to AD/MCI showing less volume.

Replicability of Two-way AD Studies

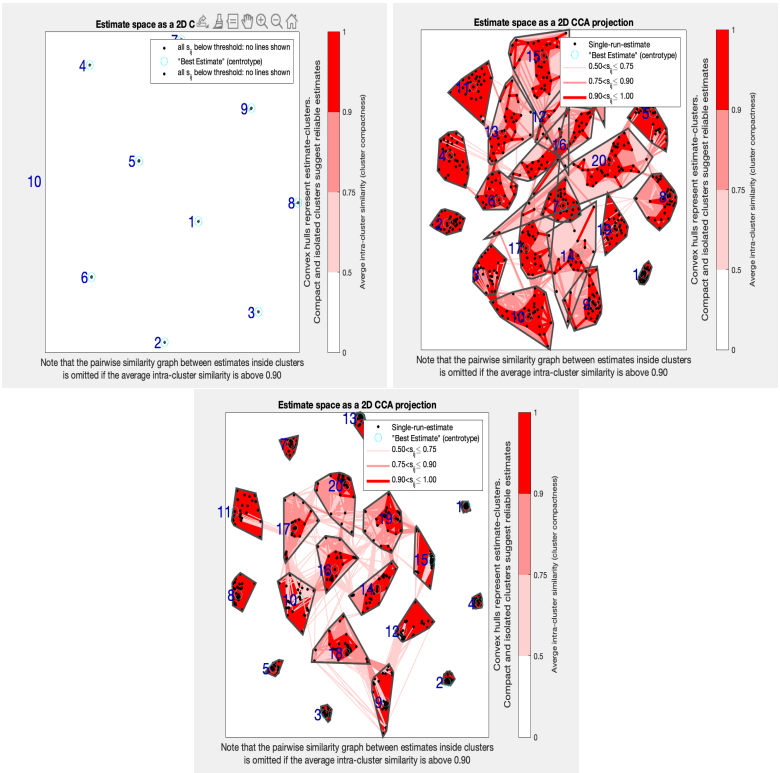
Figure 4.29: ICASSO results for MRI+SNP fusion usig sfICA.



Next, I discuss how the ICASSO plots change when fusing two modalities. Recall how we can visualize the extracted sources, the behavior of the inter- and intra-cluster similarity of the clustered ICs over bootstrapped iterations of the data, and even a dendogram of the hclust procedure and similarity matrix code by strength of similarties. This is shown for the MRI+SNP fusion using sfICA in Figure (4.29). The first plot (top left) displays the source signals that are extracted, and how 50% weight is given to each of the two modalities. The first half is representative of the SNP data and the second half shows the signals of the sMRI components. These “rows" on the y-axis actually represent the columns of the ICs. Then, the similarity dendogram and matrix visualization show the complexity of the relationships of the sMRI data but the high similarity. When crossing the SNP cluster similarity with the sMRI cluster similarity (bottom left corner of the 4th plot), one can see much lower similarity. This is a

good thing, because we want high intra-cluster similarity and low inter-cluster similarity. Thus, sfICA is able to parse the two modalities effectively. In addition, the stability index is very high for the first 10 components and then drops considerably. This means that the mutual information gained from fusing the sMRI ICs with the SNP ICs reaches its peak performance around component 10, well below the suggestion of MDL. The plots representative of the CCA projections onto the 20 ICs are shown in Figure 4.30, from FastICA, Infomax, and rICA, respectively, with FastICA outperforming the other optimization procedures. ICASSO summary measures are shared in Table 4.9 and will be discussed with the FDG+sMRI fusion.

Figure 4.30: ICASSO results for MRI+SNP fusion.



Furthermore, extracted ICs are concatenated across modalities so that each IC contains signals related to FDG and signals related to sMRI. Figure (4.31) shows this data fusion for FDG+sMRI, including the similarities among the clusters and the behavior of the stability index, I_q, for Infomax. Note that this does not appear to perform as well for the two neuroimaging modalities as it did for the combination of imaging and genetic data. One may notice how the source signals themselves also appear differently. While the first two ICs show a high cluster tightness, this declines rapidly after the first 5 ICs. The den-

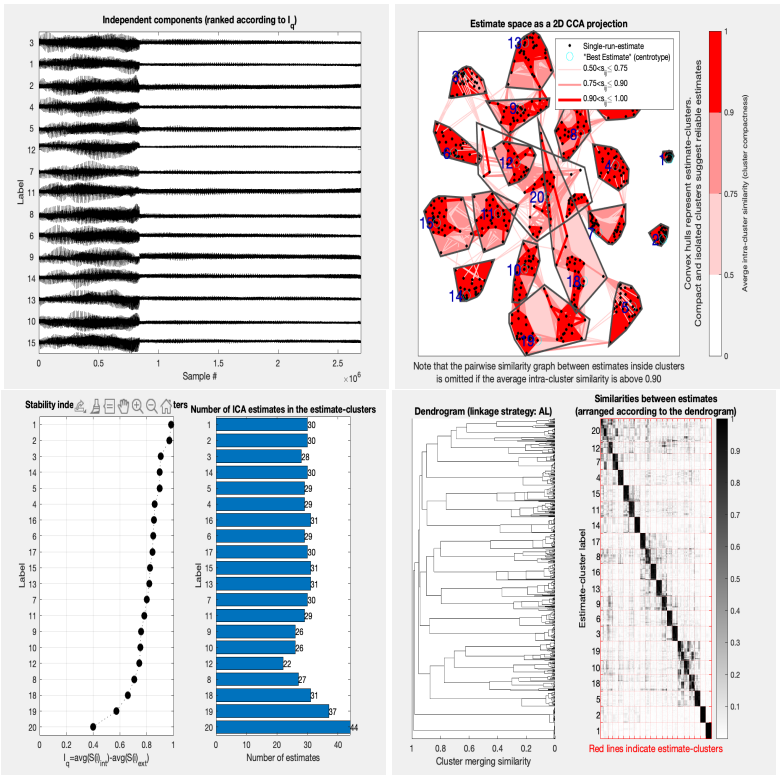
drogram, similarity matrix and CCA projection further confirm that there is a greater amount of inter-cluster compactness, meaning there is less separability among the ICs. Fortunately, we see improvement in the FastICA (top two left plots) and rICA algorithms (middle two plots). However, sfICA has a bit more trouble determining the clusters. Since the dots in the plot represent a component, these sharp, geometrical shapes we see emerging mean that there are ICs within the cluster that have a greater distance from the rest of the components that have been placed in this cluster. In other words, we begin to see some components that act as potential outliers after multiple runs of the algorithm.

Next, I present the results in Table 4.9 and compare the statistical reliability among the algorithms and between the two different modality combinations. Within the first two-way data fusion set, we see that FastICA has the highest maximum but Infomax shows the highest mean. However, sfICA has a lower variability among the components and lower correlation among the unmixing matrix, denoted $r\_$. For the neuroimaging combination, we see a higher maximum in rICA but a higher mean in FastICA. The lowest variability and correlation now occurs in rICA. Thus, both methods show improvement in the spread of the components from run to run within the sMRI+SNP data. The complex and diverse data fusion of imaging and genetics data provide "long-run" benefits of cluster stability, even if the overall maximum is higher with other algorithms. Within the neuroimaging data, FastICA overall appears to be the best in terms of the mean and standard deviation. This says that the strength of FastICA on neuroimaging data is that the stability is maintained beyond the first couple of components. However, rICA provides the maximum and the lowest correlation, meaning it performs best in separating the impact of the two diverse imaging modalities.

Table 4.9: Summarization of ICASSO results for mCCA+jICA two-way fusion.

Data	Algorithm/Measure	Max	Mean	Sd	$r\_$
sMRI+SNP	**FastICA**	**0.9861**	0.9001	0.1019	0.0379
	Infomax	0.9822	**0.9252**	0.0689	0.0518
	rICA	0.9477	0.8090	0.0892	0.0474
	sfICA	0.9837	0.8534	**0.0685**	**0.0343**
FDG+sMRI	**FastICA**	0.9825	**0.9659**	**0.0101**	0.07
	Infomax	0.8568	0.5970	0.1622	0.1266
	rICA	**0.9863**	0.9215	**0.0541**	**0.0532**
	sfICA	0.9462	0.8077	0.0638	0.0615

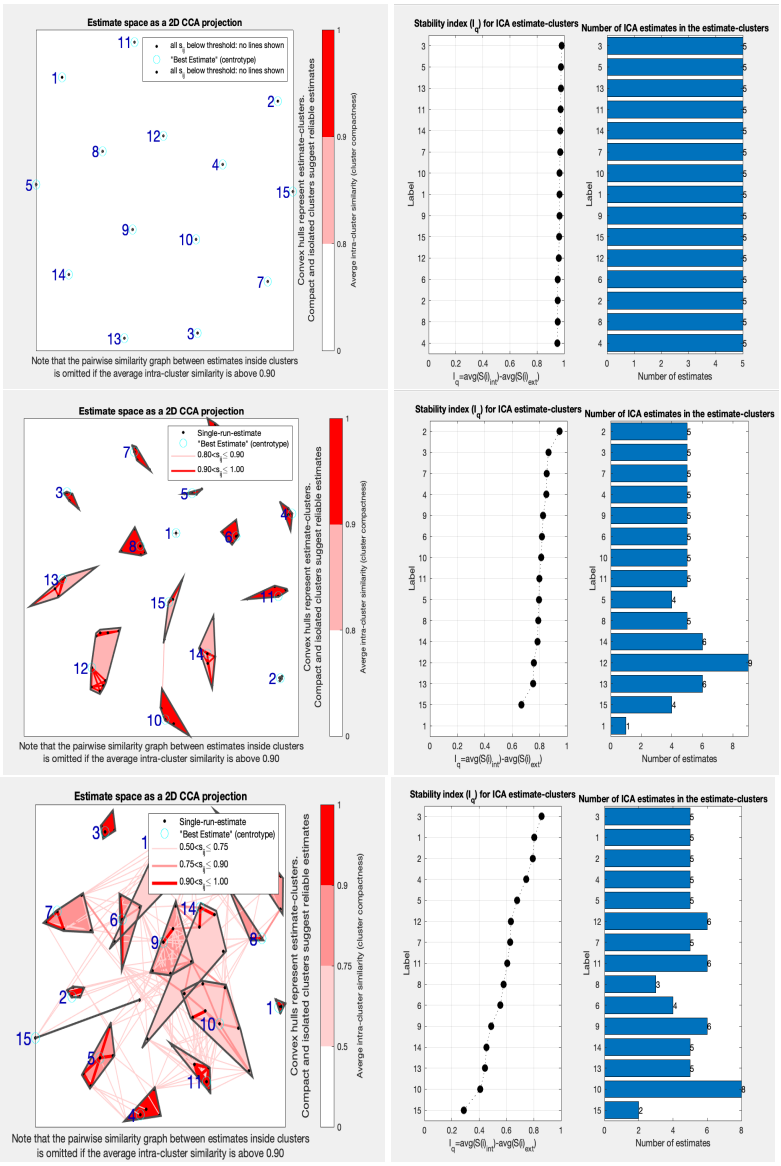
Figure 4.31: ICASSO results for FDG+sMRI fusion.



4.3.2 Three-way Imaging Genetics Comparisons

There is a key element of multi-modal data fusion that lends to an interesting interpretation concerning the comparison of disease states. Due to the fact that AD has a progressive pathology, certain modalities may experience more changes in some stages versus the other stages. For example, it is believed that the uptake in glucose diminishes even prior to structural atrophying. Thus, when including FDG and sMRI modalities in data fusion, a comparison of the components between MCI and AD stages may not show a strong correlation at the MCI stage but may very well show a strong correlation at the AD stage. However, this is also under the assumption that this correlation is explanatory of an overlapping spatial location. Cross-correlations within different regions of the brain may be considered when splitting the IC features and running correlation analyses on each "top" part of the component with all the other "bottom"s (referring to the literal source matrices). The same may be said for any other neuroimaging modality. However, this conceptual issue is not present with genetic data, because our genetic material does not change throughout hu-

Figure 4.32: ICASSO results for FDG+sMRI fusion.



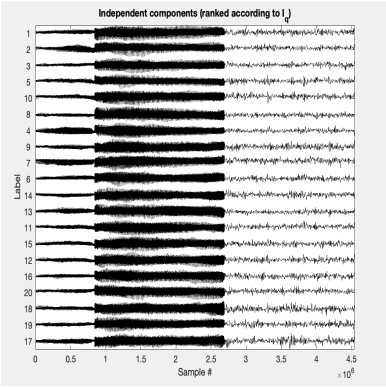
man lifetime. Instead, genetic information provides static, or an unchanging, indication of AD. This means that another benefit of including genetic data is to determine patients who are at risk of AD, even prior to the anatomical or functional changes in the brain. Furthermore, it allows one to observe the interactive effects of a static biological measurement with dynamic neuroimaging. In the last section, I combined the discussion of statistical reliability with anatomical interpretations in the context of group comparisons among compo-

nents and within two-way MMNG decomposition. Both mCCA+jICA and p-tIVA+mCCA frameworks are now considered for all four algorithms with the combination of both neuroimaging modalities with genetic data. Then I discuss group comparisons of the AD categories and the implications of all approaches on the consistency of the IC estimates.

Group Comparisons Using 3-way mCCA+jICA

First, three-way MMNG is considered on ADNI1 data, FDG+SNP+sMRI using only the rICA and sfICA optimization techniques. This is to test the newly written algorithms into the Matlab Toolbox and to also see whether the anatomical and disease status results are in line with past results using the more traditional algorithms. Anatomical plots and component graphs showcase the ability of a three-way multi-modal fusion to enhance the combination of neuroimaging and genetic modalities compared with the previous two-way discoveries.

Figure 4.33: Recovered 3-way source signals.



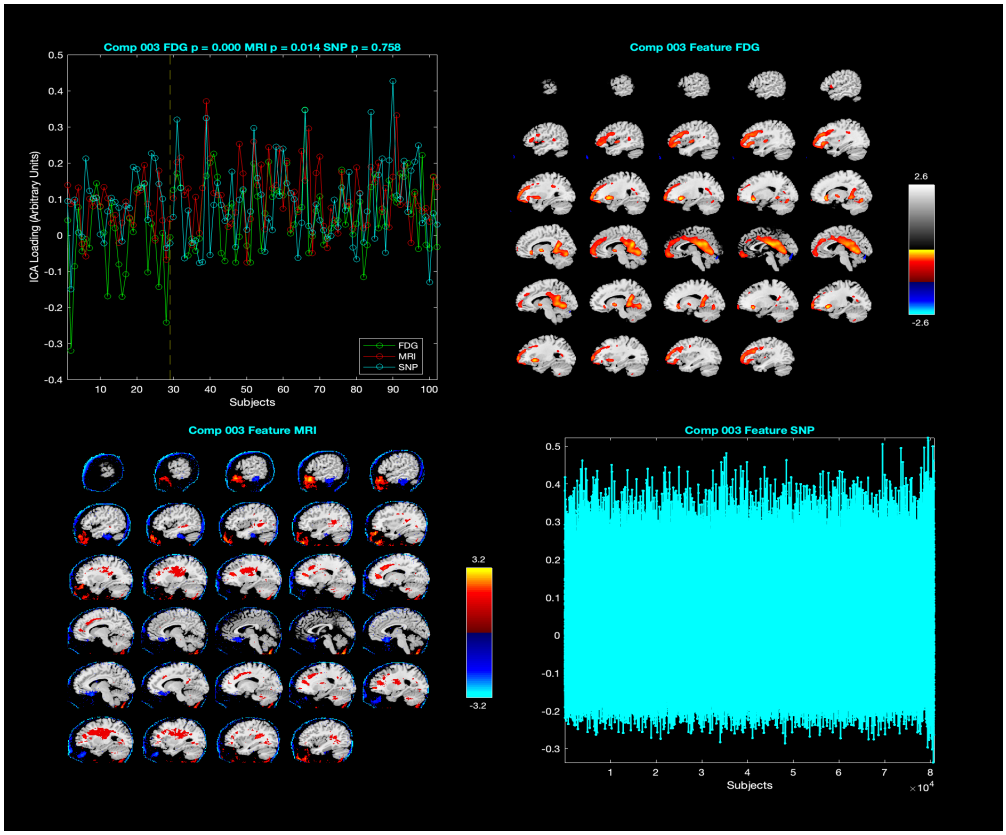
Using ADNI1 data, I consider FDG-PET, sMRI, and SNP data on the same 135 subjects. This is run two ways, on my laptop using a subset of the data to check how long this would take to run and to test the new algorithms and then in a batch file submitted through a job in the scratch computing cluster provided by the GACRC of UGA. Thirty repetitions of the algorithms through ICASSO are performed for FastICA, Infomax, rICA and sfICA algorithms. This computational load was certainly greater than the two-way fusion (also submitted to the cluster). Running all four algorithms with three modalities and 30 ICASSO runs each took about a day, whereas the two-way components took roughly half of that time. I should point out the diversity of the signals we now observe within the recovered sources. Figure 4.33 shows the signals in order of FDG, SNP and then sMRI. The representation of FDG is smaller than sMRI because of the dimensionality. This means that the number of elements within

the components are selected based on the size of the modality. While some may be concerned this places more weight on the higher dimensions, keep in mind that the data-driven technique of decomposition assigns weights to each vector by considering all the other points in that data. So, even though the FDG dimension is smaller, the weights themselves may carry more literal weight in the interpretation of the cross-modality information in the reconstructed features.

Group comparisons may be made using both statistical tests on the components and visualizations both obtained through the FusionICA Toolbox (FIT) in Matlab. Figures 4.34-4.36 and Tables 4.10 and 4.11 present the group comparison results for the first 6 components. As mentioned, the components are concatenated so that one component contains material for each modality included in the analysis. One may then extract the elements from that component that pertain to the modality and perform group comparison tests such as a two-sample t-test. Remember that this is possible because the weights within each vector are feature/component-specfic but are also subject-specific. Therefore, a t-test of a sliced component by modality tests the mean differences of the reconstructed weights between modalities. Similarly, each modality contains disease-specific information nested within the biological source. This means that the elements of the ICs may be further split up so that disease states may be compared within and across modalities. The immediate output is presented, which shows a comparison of the disease states within modalities adjusted for the weights, or influence, of the other modalities.

Figure 4.34 shows the results for component 3, sliced by disease status and modality. The top left plot shows the recovered source signals, or ICA loading vectors, on the same plot. At the top of this plot, in the title, are the p-values obtained from a group comparison of AD and MCI disease groups. There were 29 patients with AD and 77 with MCI. This is an interesting result because the IC results show significant differences between AD and MCI patients both in the structural atrophy and the metabolism of glucose. Additionally, the difference in the loading vectors for the SNPs at this component is not significant. So while differences between the prodromal and AD diagnoses are significant, the genetic material is similar. This confirms what we know about the genetic risk being similar between patients with MCI and AD, since MCI is a progressive stage prior to full diagnosis of AD. The other three plots of Figure 4.34 reveal the source signals of the modalities and are displayed with the corresponding location in the brain for the neuroimaging modalities. Within the FDG plot, the red/orange signifies a value closer to 0 on the interval of (2.6,-2.6), which translates to regions or structural pathways of the brain in which AD way observe a greater reduction of FDG uptake than MCI. That is, MCI has more

Figure 4.34: T-test of 3-way mCCA+jICA for component 3 and corresponding sMRI and FDG-PET anatomical plots.



uptake in the brain than AD, confirming this to be a statistical significant drop in the metabolism of glucose between the stages of MCI and AD. Equivalently, the sections of the brain in red for plot of MRI show locations where brain volume is greater for MCI than AD.

One may further compare this component by observing the beta weights corresponding to the p-values. That is, the magnitude of the weights reveals the relative significance as a comparison within modalities, while the direction informs the difference in the mean is attributed to greater values for AD or MCI. Table 4.10 provides the beta weights of the first 6 components for each modality. The weights are fit to the mean of the loading vector within the corresponding modality, so that the biological significance may be interpreted in the units of the data source. For component 3, the FDG beta weight has a value of -0.1996 and sMRI has a weight of -82.0840. Because these are negative

values, we may interpret this to mean there is both a loss of FDG and sMRI. If there is a component that is negative for FDG and positive for sMRI, this would indicate a spatial location in the brain in which FDG diminishes but the structural atrophy has not occurred yet. Notice, there is not a difference in sign for either of these modalities.

Table 4.10: Beta weights of FDG+sMRI+SNP for mCCA+jICA with rICA.

Component	**FDG**	**sMRI**	**SNP**
1	0.0066	61.0110	0.1368
2	-1.354	-14.9408	-0.0472
3	-0.1996	-82.0839	-0.2228
4	-0.0851	-43.4290	-0.0690
5	0.1117	1.4838	0.0100
6	0.0826	4.0685	0.0275

The beta weights for component 3 are presented in Table 4.10 and the p-values for the top 6 components comparing MCI-AD, CN-MCI, CN-AD are provided in Table 4.11. Significant values are in bold using significance level 10%. From this table, we may confirm the p-values that we saw in Figure 4.34. Also within the MCI-AD comparison, we see that component 2 is significant in FDG but not in sMRI. Similarly, component 4 is significant for sMRI but not for FDG. This means that locations in the brain have been identified where a reduction in uptake occurred but not structural atrophy and vice versa. This shows that the spatial location and rate of decline in both neuroimaging modalities are not equivalent. The same is true for components 5 and 6 of the imaging data in the CN-AD category, as displayed in Figure 4.35. When FDG is not significant, we see more accumulation of the blue magnitude, whereas when FDG is significant, we predominantly see red. Thus, the magnitude closer to 0 shows a greater reduction in the FDG uptake for AD than for CN subjects. Within the CN-MCI comparison, the only significance we see is for the SNP data. However, surprisingly, there is not a significant SNP difference in the jump from CN to AD (at least for this sample). Another strategy for group comparison is to rather slice the components by disease status only and test the combined effect of all three modalities in terms of disease status. Figure 4.36 provides plots of the source signals for IC1 of CN-AD, IC1 of AD-MCI and IC5 in terms of MCI-CN. Notice that IC1 shows significant differences for CN-AD and AD-MCI comparisons, but component 5 does not show a significant difference between MCI and CN.

Figure 4.35: Comparison of components 5 and 6 for 3-way fusion and corresponding FDG brain images.

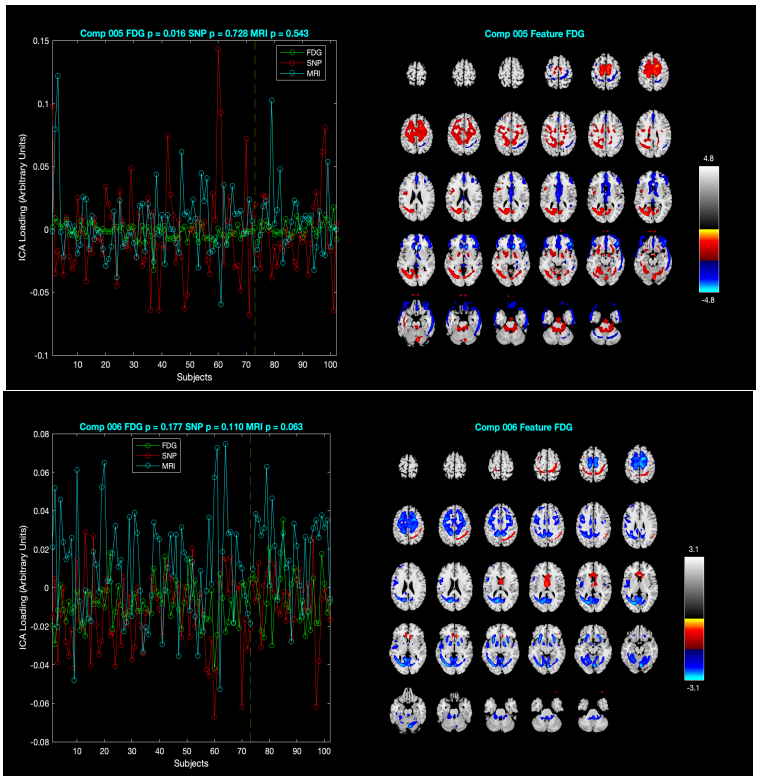


Figure 4.36: T-tests of disease categories for components 1 and 5 over all modalities for 3-way mCCA+jICA using rICA; from the top left, top right and bottom, t-tests are given for CN-AD, AD-MCI and MCI-CN.

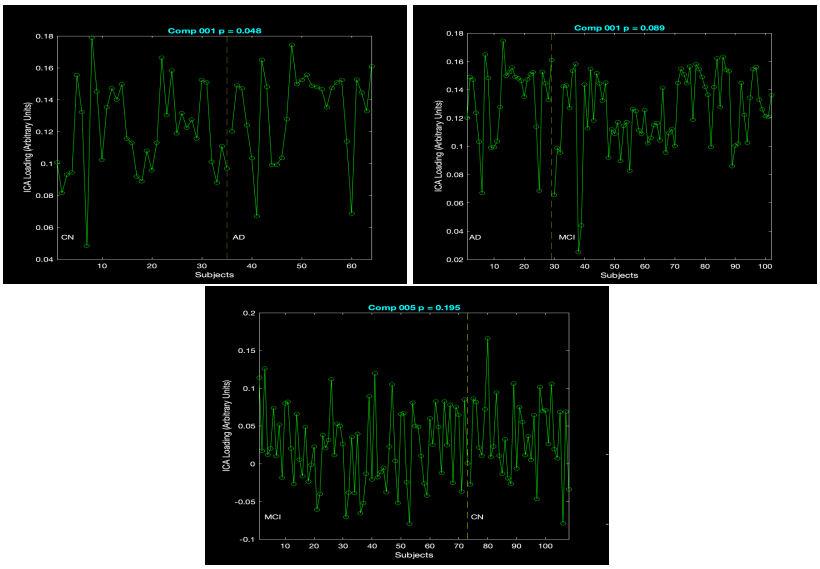


Table 4.11: mCCA+jICA T-test of CN-AD, MCI-AD, CN-MCI for jICA

MCI-AD	1	2	3	4	5	6
FDG	0.2517	**0.0172**	**0.0002**	0.4086	0.1997	0.7906
MRI	0.6427	0.7396	**0.0137**	**0.0227**	0.94214606	0.2198
SNP	0.7041	0.8799	0.75805801	0.5108	0.6509	0.7957
CN-MCI	1	2	3	4	5	6
FDG	0.6724	0.9001	0.4217	0.9432	0.9835	0.3948
MRI	0.9841	0.3525	0.5721	0.1852	0.9977	0.8432
SNP	**0.0740**	0.5566	**0.0720**	0.2146	0.9317	0.8319
CN-AD	1	2	3	4	5	6
FDG	0.9538	0.6840	0.2623	0.7530	**0.0159**	0.1767
MRI	0.8539	0.9933	0.7908	0.8449	0.5431	**0.0633**
SNP	0.2347	0.5695	0.5561	0.9697	0.7283	0.1200

Reliability of 3-way ICA Frameworks

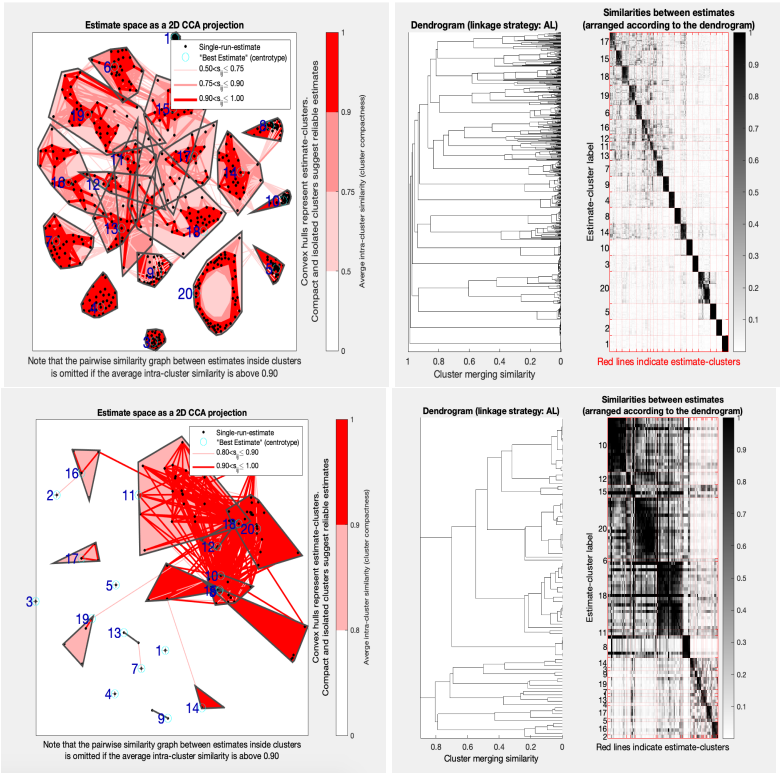
In light of these group comparisons, it is important to come back to the initial goal of this research. Multi-modal data fusion with application to imaging genetics combines neuroimaging and genetic material with the purpose of advancing knowledge of neuropsychiatric illnesses. However, data fusion itself is just one step in making the clinical translation of this information. The purpose of this research is to assess the validity of multi-way decomposition analyses using novel and traditional algorithms to see what we may learn about AD. Although multi-modal decomposition is only available in more recent literature, the optimization techniques applied are more often than not from traditional methods designed for lower-dimensional data that are linear and/or constrained. Thus, I want to accomplish two things: 1) determine the statistical reliability and therefore replicability of the IC estimates for traditional methods (FastICA and Infomax) and compare this with the new methods I have developed (rICA and sfICA applied in a multi-modal scenario); 2) compare the most common multi-modal ICA technique (mCCA+jICA) with p-tIVA while adding mCCA as a post-processing step; 3) to assess how this varies for the different modalities in a specific application to AD using the ADNI data. The goal of ICA is to extract much lower dimensional sources from complex high-dimensional data and to ultimately feed this into other analyses, including group comparisons, longitudinal analyses, or even classification and deep learning algorithms. Therefore, in the remaining part of this chapter, I will focus on the consistency of the IC estimates for the three-way multi-modal scenario with FDG+sMRI+SNP data as input, compared under the frameworks of mCCA+jICA and p-tIVA+mCCA

for all four algorithms, taking into account constrained vs unconstrained and linear vs nonlinear assumptions.

In order to assess this statistical reliability, I begin where I left off with the output for the mCCA+jICA framework under ICASSO repetitions. I do believe that it can become redundant to present all plots for all algorithms. For this reason, I present two of the four algorithms for each comparison. However, I summarize the results of all algorithms in Table 4.12. First, I begin by discussing Figure 4.37, which gives the 2D CCA projections of the clustered components over 30 runs and the associated dendrograms and similarity matrices of fusion using FDG+sMRI+SNP. The two algorithms under consideration in this plot are FastICA and rICA. The cluster plot is somewhat difficult to interpret. There are several clusters that appear to perform well in rICA (bottom left), but we see the more geometrical shapes appearing and some clusters with the pink color, symbolizing $I_q = (0.5, 0.75)$. In addition, there are overlapping hulls in the FastICA plot (top left), but we see a darker red, that indicates a high intra-cluster similarity. The story comes together when viewing the dendrogram and similarity plots on the right. Based on these plots, one can see that the similarity graph has lower levels (almost white) on the off-diagonal, corresponding to the inter-cluster similarity scores being low (high cluster separability). The dark areas we do see are still close to the diagonal, corresponding to ICs within the same cluster (at least for the most part). On the other hand, the dendrogram and similarity graph in the bottom right (for rICA) some very high inter-cluster similarities for the first two modalities, FDG and sMRI with the SNP data performing better. Notice that the heights of the dendrogram also correspond to the similarity measures. From these graphs we may speculate that FastICA retains statistical consistency better than rICA for repetitions of these data and modalities.

There is an explanation for the results in this plot. The original idea of p-tIVA is to allow each modality to be run under a single ICA using one of the four algorithms. Then, the extracted features are run through mCCA to retain components that express the correlations among and within the modalities. After this step, the fused canonical variants are then run under the ICASSO technique. The expected advantage of this multi-modal ICA framework is that it allows each modality to have its own mixing and unmixing matrices. On the other hand, mCCA+jICA concatenates these vectors assuming that the covarying information is similar among modalities. Thus, testing mCCA+jICA and p-tIVA+mCCA is assessing whether the order in which the correlation analysis is applied - either to the normalized data, the former, or the ICA source vectors, the latter - has an impact on the statistical reliability of the final ex-

Figure 4.37: FastICA vs rICA for 3-way mCCA+jICA

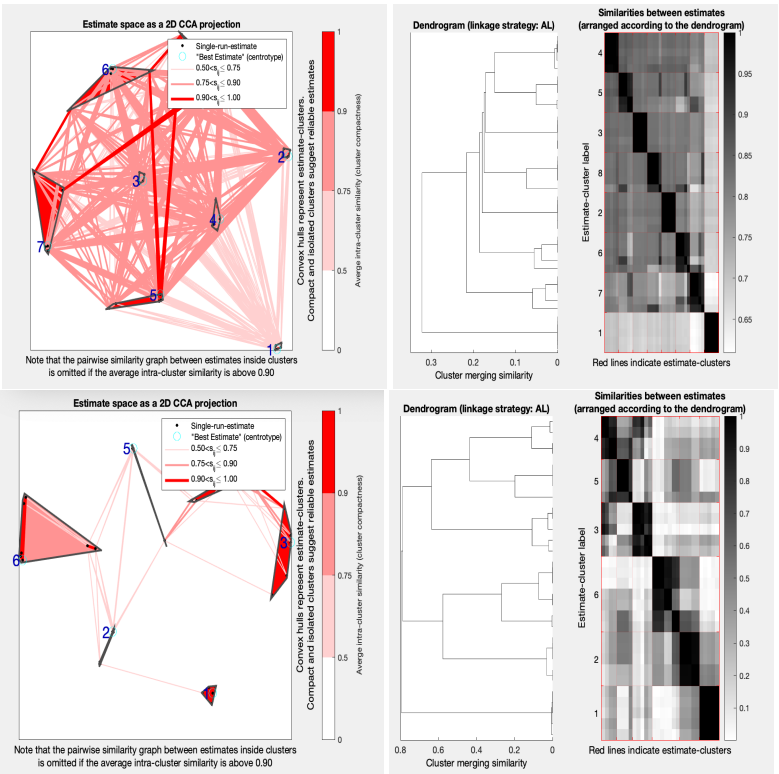


tracted features. However, recall that mCCA does not perform as well when the correlation relationships are too similar. Given the pre-processing, PCA and whitening, that occurs for the linear algorithms (FastICA and rICA), one would not expect mCCA to perform well on the linear algorithms simply because the data have already gone through three steps of data reduction and are therefore less dissimilar than the normalized data as inputs into mCCA, as the previous model does.

Another comparison that may be made between the two ICA frameworks is also due to the order of the mCCA step and the algorithm pre-processing that occurs. When performing p-tIVA, tIVA is applied to each modality individually, meaning that the number of components to withdraw are based on MDL. Then feeding these results into mCCA further reduces the dimension. However, when the mCCA step is run first on the entire data, this is feeding a fully concatenated data, with a higher dimension, into jICA. The next plots are a result of running p-tIVA+mCCA on FDG+SNP+sMRI data. I let the algorithm guide the number of components extracted to see how this would be

determined for the nonlinear methods. Since Infomax and sfICA are both nonlinear methods, PCA is not run prior to feeding the data into the ICA model. The Infomax algorithm reduces the dimension by maximization of the entropy of a neural network with non-linear outputs, and sfICA uses a sparse filtering approach that applies an L_2 norm penalty to the cost function. Since this level of sparsity has not been included in multi-modal analyses before, it is unknown how this will perform (other than my simulation studies). Based on Figure 4.38, we can see that 8 vectors were selected by Infomax and 6 were selected with sfICA. Infomax has a very low performance. However, literature supports the fact that gradient descent does not bode well for high-dimensional data Sun et al., 2018; Verghese et al., 2021. On the other hand, sfICA offers some improvement for a nonlinear and unconstrained optimization approach for MMNG. Several of the clusters show the darker red color and simultaneously reveal cluster tightness. Within the similarity graph and reflective dendrogram, we can clearly see that sfICA offers this improvement. Although we still see high similarities in the off-diagonal, it is not to the same degree of Infomax. These results are discussed in more detail and presented in Table 4.12.

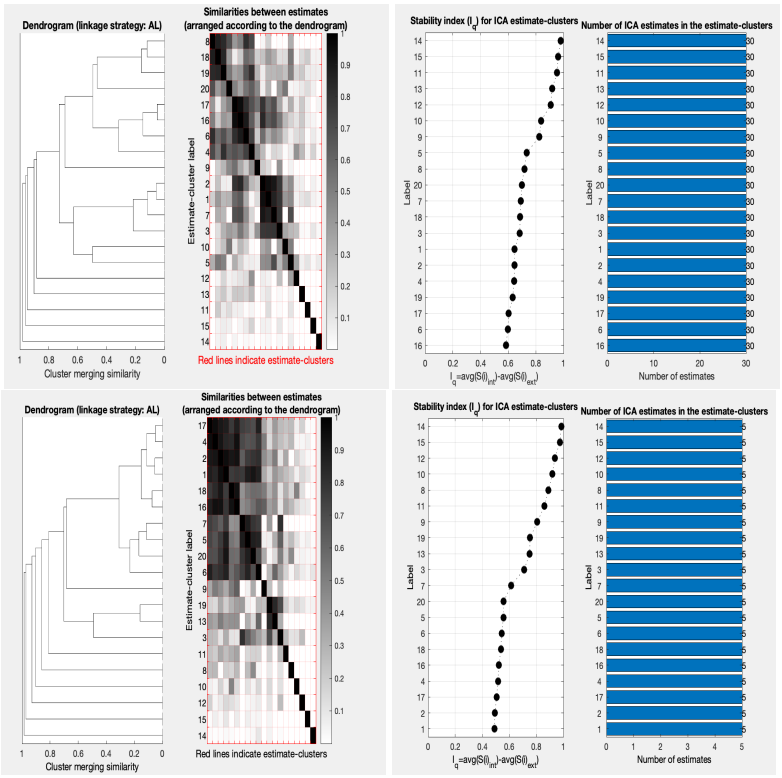
Figure 4.38: Infomax v sfICA for 3-way tIVA.



Before providing a full summarization of the stability results, I would like to showcase the results of FastICA and rICA as comparisons for p-tIVA+mCCA. Instead of allowing the algorithm to guide the number of features extracted in the process, I impose 20 on each to compare the behavior of the algorithms, going beyond the suggested 6 or 8 features recommended for the nonlinear algorithms discussed. Figure 4.39 shows the output of the dendrogram and similarty matrices on the left and a graph of the stability index, I_q on the right. The outcomes are quite similar for both of the linear algorithms overall. However, one area of uniqueness between the two linear approaches is that within the two neuroimaging modalities (the two rectangles within the similarity graphs in the top left), we can see higher intra-cluster similarity for FDG but lower intra-cluster similarity for sMRI and vice versa. On the other hand, the performance of SNPs appears very similarly, where there appears to be lower intra- and inter-cluster similarity. Consequently, the heights of the dendrograms between the two algorithms are close, indicating a similar magnitude of similarity averaged over that cluster with the modality. For the graphs of the stability index, the first four ICs show high cluster compactness, above 0.90, among the bootstrapping and initialization iterations, but this quickly drops below 0.80 by the 8th IC. From there it drops for both at about the same rate, with a slightly steeper decline by the 20th IC for rICA. We can compare these in more detail in the next table.

Table 4.12 provides the results of all 3-way MMNG analyses, for all four algorithms, under the two multi-way frameworks. For mCCA+jICA, each modality must have the same number of components, so I chose the lowest recommendation among the three modalities, which was 20. The FastICA and rICA plots were discussed, and this table confirms that FastICA preserves a higher statistical consistency overall. However, the results of rICA are quite close to that of FastICA, and the maximum is the highest. Infomax did not perform well, and sfICA has high variability. As mentioned, I allow the p-tIVA+mCCA algorithm to drive the number of components. Although Infomax and sfICA both have a high maximum, the means are much lower. However, the correlation within the unmixing matrices are still relatively low. Although Figure 4.39 shows that 20 ICs were extracted for FastICA and rICA, the summarizations in the table are based on the recommendation, which was 10 and 9, respectively. When this is done, rICA maintains a higher mean and lower standard deviation than FastICA. That being said, the two linear algorithms are highly comparable. When comparing between the mCCA+jICA and p-tIVA+mCCA frameworks, one can see that the linear algorithms perform better overall. While nonlinear Infomax does not perform well, sfICA offers some improvement for

Figure 4.39: FastICA vs rICA for 3-way tIVA.



both methods, just not at the level of the linear algorithms. In addition, we see higher variability within within the nonlinear algorithms. With the exception of Infomax under mCCA+jICA, almost all algorithms contain between approximately between 0.04 and 0.06. That is, overall the unmixing matrices are able to parse out the complexity. For both frameworks, FastICA and rICA perform very similar, meaning that rICA has consistently shown to measure up to the FastICA results.

If we want to compare the unconstrained vs the constrained algorithms, then we are comparing the traditional algorithms with novel rICA and sfICA. While rICA is clearly comparable to the outcome of FastICA, it doesn't necessarily offer an improvement, at least in the context of this data. In terms of speed, it was also about the same. On the other hand, sfICA was able to show improvement over Infomax, meaning that it may offer another option over Infomax when a nonlinear approach is necessary. It is speculative whether these outcomes could also change drastically with the introduction of different modalities.

Table 4.12: Summarization of ICASSO results for three-way data fusion.

Model	Data	Algorithm	Max	Mean	Sd	$r\_$
mCCA+jICA	FDG+sMRI+SNP	FastICA	0.9837	**0.8885**	**0.0533**	**0.0443**
		Infomax	0.7896	0.6618	0.0793	0.1249
		rICA	**0.9967**	0.8501	0.0622	0.0497
		sfICA	0.8401	0.7294	0.0955	0.0628
p-tIVA+mCCA	FDG+sMRI+SNP	FastICA	**0.9920**	0.9354	0.056	0.0575
		Infomax	0.9632	0.6035	0.1488	0.0522
		rICA	0.9833	**0.9484**	**0.0444**	0.0431
		sfICA	0.9590	0.7896	0.0810	0.0597

Chapter 5

Conclusion

The ways in which data are collected and analyzed in this day and age are experiencing a revolution based on the sheer volume of data and the increasing complexity of the measurements obtained from advanced technology. Data scientists and statisticians are working to develop a variety of methods that will be applicable to many of the diverse applications now available. Multi-modal data fusion is challenging but imperative to make sense of the massive influx of novel data that many disciplines are currently experiencing. The hope is to motivate researchers across disciplines to invite multi-modal data fusion into their own field. Another goal is to encourage researchers to include diverse data sets outside of their study area to discover how multi-disciplinary interactions can improve the understanding of many scientific processes. More specifically, I aim to elucidate how combinations of this information helps us to learn more about neuropsychiatric illnesses. The ADNI data has opened the door to tremendous discoveries of Alzheimer's Disease pathology. However, there is still a lot of work to be done to understand how to take the data from multiple sources, be it imaging, genetic or clinical, and combine the data in a meaningful way that will have a direct clinical impact.

Pursuing data driven applications will diminish the restrictions found in model-based approaches and makes it possible to work with multiple large and complex data sources. Two mainstream decomposition data fusion strategies, joint ICA and parallel IVA, were discussed in light of their theoretical underpinnings and challenges with connection to constrained and linear optimization techniques that may be limiting. Component analysis provides a more concise and concrete analysis by effectively reducing the dimension of the data, while preserving interpretation, thereby promoting both global and local generalizability. As discovered, each algorithm I used was useful at one point, but no one method is the "best". This saying goes for data fusion as well. The trade-

offs under these two frameworks must be considered by each researcher. The algorithm chosen to fuse the data must be based on the factors of data size, missing values, the future goals of inferential analysis for the data as well as the computational skills of the researcher. That being said, we must make an evaluation of the models explored, discuss remaining challenges to be considered, and highlight the wealth of future work that awaits in MMNG studies.

5.0.1 Model Evaluation

The results that were carried out this dissertation may be discussed in terms of the ICA frameworks, the use of unconstrained versus constrained or linear vs nonlinear algorithms and connected to anatomical or disease-specific interpretations. I will assess these methods based on the criteria stated before: statistical reliability (which I now include with algorithmic stability), interpretability and generalizability. In this work, I divided the analysis into two multi-modal ICA frameworks, mCCA+jICA and p-tIVA+mCCA, where the point of data concatenation, the steps of pre-processing before running the algorithms, and the order of the correlation among the data sources are considered. In addition, I looked at four different optimization techniques, two which are traditional and constrained, FastICA and Infomax, and two which are new techniques that had not been applied to the multi-modal scenario prior to this research, unconstrained rICA and sfICA. While FastICA and rICA make linear assumptions, Infomax and sfICA make nonlinear assumptions. These algorithms were compared by assessing the statistical reliability of the algorithms using various modalities represented by simulated data and real-life ADNI data for one-, two-, and three-way ICA models. Through repetitions and bootstrapping, we asses the inter- and intra-cluster reliability as well as the overall cluster compactness index score, I_q. This was viewed in terms of maximum, average and correlation, and the correlation among the fused modalities were assessed.

Overall, the novel methods using rICA and sICA scaled well to most scenarios, often showing similar results as the traditional algorithms and sometimes improving these findings. With the simulated data, we saw very low stability among all of the algorithms, although there was slight improvement with an increased sample size. While the stability of simulated SNP data was the lowest of the simulated sources, both rICA and sfICA showed increased reliability as the sample size of the subjects increase. In fact, as the data size and complexity increased, all four algorithms performed better. The simulations of the rsfMRI data using SimTB were suspect, as they performed much better over all. I believe this is because SimTB is derived in a very similar manner to the structure of the ICA model. In all cases, the ICASSO results made it possible to both visu-

alize and quantify the consistency of the new methods employed. The SimTB simulations did permit visualizations of the source components which matched well with the original ground truth spatial maps.

For the real ADNI data, single modalities of sMRI, FDG-PET and rsfMRI were considered. A GWAS showed the very little information gained from the SNP data. Group ICA using SBM was applied to sMRI and FDG-PET separately, and spatial ICA with the post-processing technique of dFNC was applied to the rsfMRI data. Analyzing the real data allowed us to examine the extracted sources as spatial images overlaying slices of the brain in axial, coronal and sagittal views. The intensity of the loading vectors highlighted areas in the brain with known relevance to AD pathology. In all of these cases, we saw tremendous improvements in the ability of all of the algorithms to retain statistical reliability. Interestingly, rICA performed the best with sMRI data and sfICA performed the best with FDG-PET data. However, Infomax maintained the highest consistency for rsfMRI data.

Then two-way modalities using sMRI+FDG and sMRI+SNP were examined. The two-way models were considered using mCCA+jICA with all four algorithms. Anatomical maps of sMRI were compared for both models to see if more information is learned when paired with SNP or FDG-PET. Overall, the intensity ranges were wider when paired with FDG uptake. Interestingly, FastICA and rICA showed close performance to each other, in terms of the mean and maximum I_q values, for FDG+sMRI. However, FastICA and Infomax performed similarly in sMRI+SNP, outperforming rICA and sfICA. Note that sfICA showed the lowest variability when SNPs were included with sMRI. Infomax did not perform well on FDG+sMRI, and both rICA and sfICA showed marked improvement over the constrained, nonlinear approach Infomax.

Finally, three-way MMNG decomposition analyses were carried out with sMRI+FDG+SNP using both mCCA+jICA and p-tIVA+mCCA for the traditional and new algorithms. Again, FastICA and rICA were close in performance in both multi-modal scenarios. The linear approaches performed better for data fusion with the imaging modalities, while sparsity seemed to improve the analyses with SNP data. The sfICA model outperformed the other nonlinear model, Infomax, under both 3-way scenarios. Under the three-way fusion structure, I incorporated group comparisons and component analyses through visualizations a t-tests of the disease statuses. Statistically significant findings were recovered from the analyses, providing evidence that MMNG decomposition analyses are interpretable and confirm previous findings for disease progression.

The results of consistency seem to vary quite a bit depending on the modalities being used, the number of modalities, or even the dimension sizes of the data. This could also highlight the property of ICA of obtaining stochastic outputs. Ultimately, what we can take away from this is that multi-modal fusion using decomposition can actually retain a high consistency, depending on the data being used and the algorithm. Thus, one may conclude that the strategies employed match the data modalities considered and may be used in future downstream analyses.

5.0.2 Remaining Challenges

The first huge hurdle for me was downloading, organizing and even understanding the data. This is still a challenge for every imaging genetics problem, because each study may contain a different illness or data source. Although I provided the outline of a pipeline, describing each step that I had to take in performing multi-modal data fusion using ADNI data, this only touched on the surface of what is needed to make this pipeline a more efficient process. This requires improved organization and more detail to afford reproducability. Most of the trouble I had in the latter part of my research was trying to organize and make sense of the enormous output that I had. Although this comes with experience, I believe there are standard steps that may be taken. If I could go back and do one thing differently, then I would be more organized and much earlier on. In part, this comes from not really knowing what you are getting yourself into at first with such a large dimension of data and when applying such complex methods. The other part of this involves taking ample time to understand the data meaning, the data format and the format and organization well before attempting analyses. Therefore, going forward I would like to develop more concrete steps that I can follow, and to share these steps, in order to be more efficient and accurate in my findings.

The next challenge for me was in the development of the computation programs or functions I had to write in order to make the analyses possible. In addition, I had to access open-source code from Matlab and R. In almost every computational source I attempted to use, software, data, or other open resources, I ran into issues with errors in the code, incompatibility with newer versions of the packages, or especially instructions on programs that were not well-explained and left the user with only a vague idea of how to proceed. In my future research, I aim to keep close tabs on the processes I take, as well as the analyses I implement, including the code I write. Open science is an ultimate necessity for expanding projects in the area of imaging genetics. This

work should be a call to other researchers to prioritize open science in order to achieve statistical reliability and study replicability.

5.1 Future Work

In addition to sharing ideas for future work that stem from this research, several points of caution should be taken under consideration. Often when analyzing big data, the complexity of the hetereogeneity may actually be overestimated. This results in more complicated optimization procedures than may be needed, which in turn may result in the overfitting of data. Future research should explore this balance between modeling the complicated interwoven relationships among the modalities and making such an elaborate model that perhaps interpretation is lost or results are far from reality. With DL methods, the data are only as good as at the input stage and the quality of the labels attached to the data. In addition, the success of the model largely depends on the end goals. Decomposition uses higher order statistics to reconstruct the data into more malleable parts, ready for downstream analysis. However, the computational programs available for component analysis are in dire need of some work. The MatLab toolbox engineered is a wonderful resource. However, other algorithms should be built for R and perhaps other statistical software. The development of multi-modal data fusion methodology is well underway, but much work remains to be done. Let this be your call, to your discipline, to your research team - embrace multi-modal data fusion. Discover the impact that a multi-source data fusion will have on the scientific community as a whole.

5.1.1 Expand MMNG to Downstream Analyses

First, I begin by taking us back to the initial aim of this research. MMNG using decomposition has the ability to maintain a replicable data-driven method with unbounded possibilities for application. I set out to study multi-modal data fusion through ICA with the full idea that well-parsed features representative of the fused data sources would provide the endless possibility of downstream analyses. However, in my research, I discovered that the lack of global minima puts into question the consistency of IC estimates. Thus, my focus shifted a bit. It was pertinent to determine whether the extractd ICs could be justifiable in a statistical sense. Although there were some issues when increasing the dimensionality with multiple sources, overall many of these algorithms scaled well. That being said, the next step is to assess the performance of ICA in analyses, such as linear mixed models, longitudinal analyses, or categorical tests involving

the disease statuses. Furthermore, it must still be clearly assessed whether the fused data through decomposition will provide the ability to build highly predictive accuracy models. In addition, ICA models need to be compared to the deep learning algorithms that are generally harder to understand, implement and interpret. Thus, I plan to extend these methods to classification and test the predictive accuracy.

Perhaps the "elephant in the room" of imaging genetics problems is how the findings will actually make a clinical translation and impact the way that neuropsychiatric illnesses are studied, understood, or even accurately predicted. The ultimate goal is to provide treatment for the disease, or at the very least therapeutic efforts to slow the mental decline and improve the quality of life of AD patients. In order to do this, we must take a multi-disciplinary approach through and through by connecting with scientists and being open to the true underlying biological meaning. What is the proper end point of an analyses? I believe it is much more than a high classification accuracy. In addition, the goal of imaging genetics is to understand the biological impact enough to be able to include the more measurable genetic and imaging information into the diagnosis decision. Thus, much more discovery needs to take place if we are going to have true biological surrogates with the mental decline.

5.1.2 Computation of Multi-Modal Data Fusion

Due to the high demand of algorithms which allow for data fusion of multiple diverse data sets, it is vital for researchers to understand the first principles of how to proceed with such an analysis. The computational steps for performing decomposition and deep learning are discussed in the context of the imaging genetics problem. However, the aim is to provide an outline for a multi-modal analysis not constrained to any one discipline or application. Various statistical programming tools are available for computation, but this document focuses on the use of R and MatLab. SAS and Python software are also recommended for researchers, depending on his or her computing expertise. To date, these methods have been extended to a toolbox in Matlab for up to a three-way multi-modal analysis. For an m-way analysis, software will need to be further developed. The analytical results of the imaging genetics study are left out of this explanation, for the crux of this document is to compare the methods based on the conceptual impact as discussed in the next section. A thorough imaging genetics application with discussion is beyond the scope of the paper.

ICA decomposition methods and DL neural networks may be run in the **MatLab Fusion ICA Toolbox (FIT)** developed by the TReNDS center in Atlanta, GA under the leadership of Professor Vince Calhoun from Georgia

State University. The Toolbox directly accepts NiFTI or image files for neuroimaging and text or ASCII files for the SNP-level genetic data. If different file formats are desirable, particularly if imaging files are not available or not applicable to the research, then the hard-coded functions are still made available to allow for diverse data sets. The latest update of the Toolbox (**FITv2.0e**) may be downloaded directly from GitHub (https://github.com/trendscenter/fit). By using these functions (or the Toolbox itself) the researcher has the option to select up to three modalities and up to three subgroups. An online help manual is available on the TReNDS website (https://trendscenter.org/software/fit/) along with example data and clear instructions to obtain practice results. The researcher will have access to the following methods: jICA, mCCA, mCCA+jICA, paraICA, transposed IVA (tIVA) Adali et al., 2015, parallel group ICA + ICA (PGICA) Qi et al., 2019 and Deep Fusion (using a modified form of the neural network) Plis et al., 2018. In addition to making multiple multimodal data fusion techniques available, the graphical interface and follow-up results for between group comparison is helpful for summarization of the results.

Multi-modal decomposition analysis is also possible in R. However, there currently does not exist a package containing an explicit function for the derivation of two-way to M-way ICA. Nonetheless, two ICA packages allow unimodal component analysis, **ica** and **fastICA**. The **ica** package consists of three algorithm choices, Infomax, FastICA, and JADE. The latter algorithm has yet been mentioned because it is less common, see Calhoun et al., 2009; Hyvärinen and Oja, 1999; Sui et al., 2009. Recall that the difference between the two methods mCCA+jICA and paraICA is the point of the data fusion and the correlation analysis. The R package **PMA**, contains a function, *MultiCCA*, that can perform mCCA on the multiple sources prior to using the ICA algorithm. For paraICA, the correlation analysis comes after separate ICA is is run. However, recall that the mixing matrices are derived so that they are adjusted based on the degree of correlation with the other modalities. The mixing matrices outputted from the ICA model may be used in a optimization procedure such that the maximal correlation is obtained for each. Thus, it is possible to build upon these functions in R so that it may be implemented for an m-way analysis.

The process of computing a multi-modal analysis using DL algorithms is not as straightforward as the decomposition methods. The software Python is known as the data scientist's first choice for running these end-to-end techniques. Several modules have been created to combine multiple data sets, such as **skfusion** (https://github.com/mims-harvard/scikit-fusion), **data-fusion-**

sm (https://pypi.org/project/data-fusion-sm/), **sensorfusion** (https://github.com/datascopeanalytics/sensor_fusion). For data storage and integrating large data files, Google Cloud offers some user-friendly options (https://cloud.google.com/data-fusion/docs/concepts/overview). Implementing neural networks in Python may be done utilizing the DL platform **PyTorch** (https://pytorch.org). This is recommended to the researcher when handling one or more big data sets. However, the data in the application above, although large, does not necessarily need to integrate these heavier computational procedures.

5.1.3 Longitudinal Application

MMNG data fusion is in the early stages of methodological development, and has yet to be analyzed for longitudinal applications. A few multivariate models applied to the ADNI study have been used to analyze the change in phenotypic response over time. One way to model the longitudinal change in a continuous response is to use a LMM with covariates for time and interaction of time and diagnosis. Another way is to take the difference in, say, brain volume per ROI, for the 6-month, 12-month, and so on, time points, and let the response become the change in volume. Both of these methods were applied in Chapter 2. A strong regularization may be included or a sparse, low-rank regression in order to adjust for the noisy phenotype-genotype relationships. However, these methods are not easily applied in an n-way data fusion and it is of concern whether the complex heterogeneous relationships are accounted for in the restrictive covariance structures.

If one were interested in studying how the one year change in brain volume and the one-year change in FDG uptake in addition to gene expressions interact in patients, it is possible to take the differences and analyze them as features in the MMNG data fusion process. One may even include multiple changes in response over time and include these as the features. The advantage of n-way data fusion is that the modalities may represent separate points of time for the neuroimaging modalities, while still studying the interaction of the genetic data. The model proposed for the n-way component analysis will also be extended to an analysis of the phenotype-genotype relationships over time in this way. Another potential longitudinal model is to use the extracted vectors from the n-way components as covariates in a LMM model, analyzed with the main effect of time and the interaction of time and diagnosis. This latter method limits the scope of inference considerably because selection bias is possibly induced when hand-picking the response variable.

5.1.4 Expanding to Potential Modalities

The most significant contribution of this research is the budding formulation of a multi-modal pipeline that may generalize to many different research problems. The goal is to expand multi-modal analysis to many more problems in medicine, engineering, language formation, remote sensing and more. But first, I would like to explore how this may apply to novel AD research. A recent paper revealed that the diagnosis of AD may be as simple as a blood test that measures the presence of β-amyloid from p-tau isoforms, which have been proven to show a strong predictive behavior of AD. In light of the classification issues mentioned in this work, having a stronger surrogate indicator could redefine the diagnosis criteria for AD in the very near future, catching the disease earlier than ever before, (Barthélemy et al., 2020). What if the p-tau levels could be included as a modality, or what if we used this to develop more accurate classification labels in order to improve the diagnostic accuracy? As drug trials for AD are now being approved, it is imperative to classify true surrogate biomarkers that elucidate the progression of AD. Future ADNI modalities may incorporate these recently celebrated biomarkers by applying the MMNG decomposition strategies to the new biomarkers. Furthermore, incorporating this information with cognitive measures would help doctors pinpoint how biological changes help us understand dementia in new ways. This biological classification would further support clinical trials for new drugs and forever alter the future of scientific discovery of AD pathology.

Bibliography

Adali, T., Levin-Schwartz, Y., & Calhoun, V. D. (2015). Multimodal Data Fusion Using Source Separation: Two Effective Models Based on ICA and IVA and Their Properties. *Proceedings of the IEEE*, *103*(9), 1478–1493. https://doi.org/10.1109/JPROC.2015.2461624

Allen, E. A., Erhardt, E. B., Wei, Y., Eichele, T., & Calhoun, V. D. (2012). Capturing inter-subject variability with group independent component analysis of fMRI data: A simulation study. *NeuroImage*, *59*(4), 4141–4159. https://doi.org/10.1016/j.neuroimage.2011.10.010

Alzheimer, T., Initiative, D. N., Go, A., Mri, T., Clinic, M., & Qc, T. (2018). Alzheimer ' s Disease Neuro Imaging MRI Overview. (May), 2017–2018.

Bell, A. J., & Sejnowski, T. J. (1989). An information-maximisation approach to blind separation and blind deconvolution. *1034*(February 1995), 1004–1034.

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database. *Nature Genetics*, *39*(1), 17–23. https://doi.org/10.1038/ng1934

Bigos, K. L., & Weinberger, D. R. (2010). Imaging genetics-days of future past. *NeuroImage*, *53*(3), 804–809. https://doi.org/10.1016/j.neuroimage.2010.01.035

Boettiger, C. (2015). An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79.

Bruno, A., & David, J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images.

Butler, P. M., Chiong, W., Perry, D. C., Miller, Z. A., Gennatas, E. D., Brown, J. A., Pasquini, L., Karydas, A., Dokuru, D., Coppola, G., Sturm, V. E., Boxer, A. L., Gorno-Tempini, M. L., Rosen, H. J., Kramer, J. H., Miller, B. L., & Seeley, W. W. (2019). Dopamine receptor D 4 (DRD 4) polymorphisms with reduced functional potency intensify atrophy in

syndrome-specific sites of frontotemporal dementia. *NeuroImage: Clinical*, *23*(April 2018), 101822. https://doi.org/10.1016/j.nicl.2019.101822

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional mri data using independent component analysis (Human Brain Mapping (2001) 14 (140-151)). *Human Brain Mapping*, *16*(2), 131. https://doi.org/10.1002/hbm.10044

Calhoun, V. D., & Adali, T. (2006). Unmixing fMRI with independent component analysis. *IEEE Engineering in Medicine and Biology Magazine*, *25*(2), 79–90. https://doi.org/10.1109/MEMB.2006.1607672

Calhoun, V. D., Liu, J., & Adali, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, *45*(1 Suppl). https://doi.org/10.1016/j.neuroimage.2008.10.057

Calhoun, V. D., & Sui, J. (2016). Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(3), 230–244. https://doi.org/10.1016/j.bpsc.2015.12.005

Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal processing letters*, *4*(4), 112–114.

Chen, J., Calhoun, V., & Pearlson, G. (2012). Multifaceted Genomic Risk for Brain Function in Schizophrenia. *Bone*, *23*(1), 1–7. https://doi.org/10.1038/jid.2014.371

Chen, J., Calhoun, V. D., Ulloa, A. E., & Liu, J. (2014). Parallel ICA with multiple references: A semi-blind multivariate approach. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, 6659–6662. https://doi.org/10.1109/EMBC.2014.6945155

Chung, J., Wang, X., Maruyama, T., Zhang, X., Sherva, R., Takeyama, H., Lunetta, K. L., Farrer, L. A., Jun, G. R., Genomics, I., & Bioscience, M. (2018). Genome-Wide Association Study of Alzheimer Disease Endophenotypes at Prediagnosis Stages. *14*(5), 623–633. https://doi.org/10.1016/j.jalz.2017.11.006.Genome-Wide

Crosson, B., Ford, A., McGregor, K. M., Meinzer, M., Cheshkov, S., Xiufeng, L., Walker-Batson, D., & Briggs, R. W. (2010). Functional imaging and related techniques: An introduction for rehabilitation researchers. *Journal of Rehabilitation Research and Development*, *47*(2), 7–33. https://doi.org/10.1682/jrrd.2010.02.0017

Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., & He, Y. (2012). Discriminative analysis of early Alzheimer's disease using multi-modal imaging

and multi-level characterization with multi-classifier (M3). *NeuroImage*, *59*(3), 2187–2195. https://doi.org/10.1016/j.neuroimage.2011.10.003

Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J., & Benediktsson, J. A. (2015). Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing. *Proceedings of the IEEE*, *103*(9), 1585–1601. https://doi.org/10.1109/JPROC.2015.2462751

Damaraju, E., Allen, E. A., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., Mueller, B. A., Pearlson, G. D., Potkin, S. G., Preda, A., Turner, J. A., Vaidya, J. G., Van Erp, T. G., & Calhoun, V. D. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage: Clinical*, *5*(July), 298–308. https://doi.org/10.1016/j.nicl.2014.07.003

Davis, D. H., Creavin, S. T., Noel-Storr, A., Quinn, T. J., Smailagic, N., Hyde, C., Brayne, C., Mcshane, R., & Cullum, S. (2013). Neuropsychological tests for the diagnosis of Alzheimer's disease dementia and other dementias: A generic protocol for cross-sectional and delayed-verification studies. *Cochrane Database of Systematic Reviews*, *2013*(3). https://doi.org/10.1002/14651858.CD010460

Dimitromanolakis, A., Xu, J., Krol, A., & Briollais, L. (2019). sim1000G: A user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, *20*(1), 1–9. https://doi.org/10.1186/s12859-019-2611-1

Duan, K., Silva, R. F., Calhoun, V. D., & Liu, J. (2020). aNy-way Independent Component Analysis, 3–7.

Erhardt, E. B., Allen, E. A., Wei, Y., Eichele, T., & Calhoun, V. D. (2012). SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability. *NeuroImage*, *59*(4), 4160–4167. https://doi.org/10.1016/j.neuroimage.2011.11.088

et al. F.S. Collins, E.S. Lander, J. Rogers. (2004). International Human Genome Sequencing Consortium,Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

et al Saykin, A. J. (2010). Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans Andrew. *6*(3), 265–273. https://doi.org/10.1016/j.jalz.2010.03.013. Alzheimer

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Gao, S., Calhoun, V. D., & Sui, J. (2020). Multi-modal component subspace-similarity-based multi-kernel SVM for schizophrenia classification. (March). https://doi.org/10.1117/12.2550339

Garibaldi, M., & Zarzoso, V. (2013). Exploiting intracardiac and surface recording modalities for atrial signal extraction in atrial fibrillation. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 6015–6018. https://doi.org/10.1109/EMBC.2013.6610923

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Fiske, A., & Pedersen, N. L. (2006). Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*, *63*(2), 168–174. https://doi.org/10.1001/archpsyc.63.2.168

Ge, T., Nichols, T. E., Ghosh, D., Mormino, E. C., Smoller, J. W., & Sabuncu, M. R. (2015). A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application. *NeuroImage*, *109*, 505–514. https://doi.org/10.1016/j.neuroimage.2015.01.029

Grasby, K. L., Jahanshad, N., Painter, J. N., Colodro-Conde, L., Bralten, J., Hibar, D. P., Lind, P. A., Pizzagalli, F., Ching, C. R. K., McMahon, M. A. B., Shatokhina, N., Zsembik, L. C. P., Agartz, I., Alhusaini, S., Almeida, M. A. A., Alnæs, D., Amlien, I. K., Andersson, M., Ard, T., … Medland, S. E. (2018). The genetic architecture of the human cerebral cortex. *bioRxiv*. https://doi.org/10.1126/science.aay6690

Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., & Nathoo, F. S. (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics (Oxford, England)*, *33*(16), 2513–2522. https://doi.org/10.1093/bioinformatics/btx215

Groves, A. R., Beckmann, C. F., Smith, S. M., & Woolrich, M. W. (2011). Linked independent component analysis for multimodal data fusion. *NeuroImage*, *54*(3), 2198–2217. https://doi.org/10.1016/j.neuroimage.2010.09.073

Grünwald, P. D. (2019). The Minimum Description Length Principle. *The Minimum Description Length Principle*, (January 2007). https://doi.org/10.7551/mitpress/4643.001.0001

Guillén, J. L. (2017). Study of Reconstruction ICA for Feature Extraction in Images and Signals.

Gupta, C. N., Turner, J. A., & Calhoun, V. D. (2019). Source-based morphometry: a decade of covarying structural brain patterns. *Brain Structure*

and Function, *224*(9), 3031–3044. https://doi.org/10.1007/s00429-019-01969-8

Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S. L., Saykin, A. J., Yao, X., & Shen, L. (2020a). Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease. *Medical Image Analysis*, *60*, 101625. https://doi.org/10.1016/j.media.2019.101625

Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S. L., Saykin, A. J., Yao, X., & Shen, L. (2020b). Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease. *Medical Image Analysis*, *60*. https://doi.org/10.1016/j.media.2019.101625

Hariri, A. R., & Weinberger, D. R. (2003). Imaging genomics. *British Medical Bulletin*, *65*(1), 259–270. https://doi.org/10.1093/bmb/65.1.259

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M. J., Potkin, S. G., Jr, C. R. J., Weiner, M. W., Toga, A. W., & Paul, M. (2012). Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *56*(4), 1875–1891. https://doi.org/10.1016/j.neuroimage.2011.03.077.Voxelwise

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., Aribisala, B. S., Armstrong, N. J., Bernard, M., Bohlken, M. M., Boks, M. P., Bralten, J., Brown, A. A., Mallar Chakravarty, M., Chen, Q., … Medland, S. E. (2015). Common genetic variants influence human subcortical brain structures. *Nature*, *520*(7546), 224–229. https://doi.org/10.1038/nature14101

Himberg, J., Hyvärinen, A., & Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage*, *22*(3), 1214–1222. https://doi.org/10.1016/j.neuroimage.2004.03.027

Himberg, J., & Hyvxrinen, A. (2003). ICASSO : SOFTWARE FOR INVESTIGATING T H E RELIABILITY O F ICA ESTIMATES BY C LUST ER I N G A N D VIS U A L I Z AT I.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6). https://doi.org/10.1371/journal.pgen.1000529

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., & Zhu, H. (2015). FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage*, *118*, 613–627. https://doi.org/10.1016/j.neuroimage.2015.05.043

Hyv, A. (1999). Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *10*(3), 626–634.

Hyvärinen, A., & Oja, E. (1999). Independent component analysis: A tutorial. *Notes for International Joint Conference on Neural Networks (IJCNN'99), Washington DC*, *1*, 1–30. https://doi.org/10.1093/mnras/stv2744

Iraji, A. (2021). Multi-Spatial Scale Dynamic Interactions between Functional Sources Reveal Sex-Specific Changes in Schizophrenia, 1–48.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., … Weiner, M. W. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. https://doi.org/10.1002/jmri.21049

Kim, S., & Lee, C. Y. (2020). A consistent approach to the genotype encoding problem in a genome-wide association study of continuous phenotypes. *PLoS ONE*, *15*(7 July), 1–24. https://doi.org/10.1371/journal.pone.0236139

Kueper, J. K., Speechley, M., & Montero-Odasso, M. (2018). The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review. *Journal of Alzheimer's Disease*, *63*(2), 423–444. https://doi.org/10.3233/JAD-170991

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., van der Lee, S. J., Amlie-Wolf, A., Bellenguez, C., Frizatti, A., Chouraki, V., Martin, E. R., Sleegers, K., Badarinarayan, N., Jakobsdottir, J., Hamilton-Nelson, K. L., Moreno-Grau, S., … Pericak-Vance, M. A. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics*, *51*(3), 414–430. https://doi.org/10.1038/s41588-019-0358-2

Kwan, R. K., Evans, A. C., & Pike, B. (1999). MRI simulation-based evaluation of image-processing and classification methods. *IEEE Transactions on Medical Imaging*, *18*(11), 1085–1097. https://doi.org/10.1109/42.816072

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*, *103*(9), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thornton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., … Seshadri, S. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, *45*(12), 1452–1458. https://doi.org/10.1038/ng.2802

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T. A., McHugh, C., … Weir, B. S. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602. https://doi.org/10.1002/gepi.20516

Lazar, N. (2008). *The statistical analysis of functional mri data*. Springer Science & Business Media.

Le, Q. V., Karpenko, A., Ngiam, J., & Ng, A. Y. (2011). ICA with reconstruction cost for efficient overcomplete feature learning. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, 1–9.

Lee, D., Kim, J., Moon, W. J., & Ye, J. C. (2019). Collagan: Collaborative gan for missing image data imputation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 2482–2491. https://doi.org/10.1109/CVPR.2019.00259

Lee, J.-h., Lee, T.-w., Jolesz, F. A., & Yoo, S.-s. (2008). Independent vector analysis (IVA): Multivariate approach for fMRI group study. *40*, 86–109. https://doi.org/10.1016/j.neuroimage.2007.11.019

Lehne, B., Lewis, C. M., & Schlitt, T. (2011). From SNPs to genes: Disease association at the gene level. *PLoS ONE*, *6*(6). https://doi.org/10.1371/journal.pone.0020133

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. *Nature Methods*, *14*(7), 641–642. https://doi.org/10.1038/nmeth.4346

Li, B., Tang, H., Chen, S., He, N., & Yan, F. (2017). [P1-096]: Multi-Model Cognitive Training for Mild Cognitive Impairment Patients: Clinical and Functional Neuroimaging Outcomes From a Pilot Study. *Alzheimer's*

Dementia, *13*(7S$_{P}art_{5}$), P276–P277. https://doi.org/10.1016/j.jalz.2017.06.163

Li, Y. O., Adali, T., & Calhoun, V. D. (2007). Estimating the number of independent components for functional magnetic resonance imaging data. *Human Brain Mapping*, *28*(11), 1251–1266. https://doi.org/10.1002/hbm.20359

Li, Y. O., Adali, T., Wang, W., & Calhoun, V. D. (2009). Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, *57*(10), 3918–3929. https://doi.org/10.1109/TSP.2009.2021636

Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, *23*(4), 439–464. https://doi.org/10.1214/09-STS282

Liu, J., Li, T., Xie, P., Du, S., Teng, F., & Yang, X. (2020). Urban big data fusion based on deep learning: An overview. *Information Fusion*, *53*(February 2019), 123–133. https://doi.org/10.1016/j.inffus.2019.06.016

Liu, J., Demirci, O., & Calhoun, V. D. (2008). A parallel independent component analysis approach to investigate genomic influence on brain function. *IEEE Signal Processing Letters*, *15*, 413–416. https://doi.org/10.1109/LSP.2008.922513

Liu, J., Ghassemi, M. M., Michael, A. M., Boutte, D., Wells, W., Perrone-Bizzozero, N., Macciardi, F., Mathalon, D. H., Ford, J. M., Potkin, S. G., Turner, J. A., & Calhoun, V. D. (2012). An ICA with reference approach in identification of genetic variation and associated brain networks. *Frontiers in Human Neuroscience*, *6*(FEBRUARY 2012), 1–10. https://doi.org/10.3389/fnhum.2012.00021

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., & Calhoun, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Human Brain Mapping*, *30*(1), 241–255. https://doi.org/10.1002/hbm.20508

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., & Fulham, M. J. (2015). Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer's Disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132–1140. https://doi.org/10.1109/TBME.2014.2372011

Lu, Z. H., Khondker, Z., Ibrahim, J. G., Wang, Y., & Zhu, H. (2017). Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *NeuroImage*, *149*, 305–322. https://doi.org/10.1016/j.neuroimage.2017.01.052

Ma, S., & Dai, Y. (2011). Principal component analysis based Methods in bioinformatics studies. *Briefings in Bioinformatics*, *12*(6), 714–722. https://doi.org/10.1093/bib/bbq090

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Marcus, C., Mena, E., & Subramaniam, R. M. (2014). Brain PET in the diagnosis of Alzheimer's disease. *Clinical Nuclear Medicine*, *39*(10), e413–e426. https://doi.org/10.1097/RLU.0000000000000547

Marenco, S., & Radulescu, E. (2010). Imaging genetics of structural brain connectivity and neural integrity markers. *NeuroImage*, *53*(3), 848–856. https://doi.org/10.1016/j.neuroimage.2009.11.030

Meda, S. A., Narayanan, B., Liu, J., Perrone-Bizzozero, N. I., Stevens, M. C., Calhoun, V. D., Glahn, D. C., Shen, L., Risacher, S. L., Saykin, A. J., & Pearlson, G. D. (2012). A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. *NeuroImage*, *60*(3), 1608–1621. https://doi.org/10.1016/j.neuroimage.2011.12.076

Meng, T., Jing, X., Yan, Z., & Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information Fusion*, *57*(2), 115–129. https://doi.org/10.1016/j.inffus.2019.12.001

Messer, H., Zinevich, A., & Alpert, P. (2006). Environmental monitoring by wireless communication networks. *Science*, *312*(5774), 713. https://doi.org/10.1126/science.1120034

Meyer-Lindenberg, A. (2012). The future of fMRI and genetics research. *NeuroImage*, *62*(2), 1286–1292. https://doi.org/10.1016/j.neuroimage.2011.10.063

Meyer-Lindenberg, A., Nicodemus, K. K., Egan, M. F., Callicott, J. H., Mattay, V., & Weinberger, D. R. (2008). False positives in imaging genetics. *NeuroImage*, *40*(2), 655–661. https://doi.org/10.1016/j.neuroimage.2007.11.058

Meyer-Lindenberg, A., & Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, *7*(10), 818–827. https://doi.org/10.1038/nrn1993

Michael, A. M., Anderson, M., Miller, R. L., Adali, T., & Calhoun, V. D. (2014). Preserving subject variability in group fMRI analysis: Performance evaluation of GICA vs. IVA. *Frontiers in Systems Neuroscience*, *8*(JUNE), 1–18. https://doi.org/10.3389/fnsys.2014.00106

Miettinen, J., Nordhausen, K., & Taskinen, S. (2017). Blind source separation based on joint diagonalization in r: The packages jade and bssasymp. *Journal of Statistical Software*, *76*.

Miler, R., & Calhoun, V. D. (2020). Hybrid dictionary learning-ICA approaches built on novel instantaneous dynamic connectivity metric provide new multiscale Insights into dynamic brain connectivity. (March), 66. https://doi.org/10.1117/12.2549368

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2008). *The Alzheimer's Disease Neuroimaging Initiative* (tech. rep.).

Nathoo, F. S., Kong, L., & Zhu, H. (2019). A review of statistical methods in imaging genetics. *Canadian Journal of Statistics*, *47*(1), 108–131. https://doi.org/10.1002/cjs.11487

Ngiam, J., Chen, Z., Bhaskar, S., Koh, P., & Ng, A. (2011). Sparse filtering. *Advances in neural information processing systems*, *24*, 1125–1133.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Nordhausen, K., Ollila, E., & Oja, H. (2011). On the performance indices of ICA and blind source separation. *IEEE Workshop on Signal Processing Advances in Wireless Communications, SPAWC*, 486–490. https://doi.org/10.1109/SPAWC.2011.5990458

Olshausen, B. A., & Fieldt, D. J. (1997). Strategy Employed by V1 ? *37*(23), 3311–3325.

Petersen, R. C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., & Fratiglioni, L. (2014). Mild cognitive impairment: A concept in evolution. *Journal of Internal Medicine*, *275*(3), 214–228. https://doi.org/10.1111/joim.12190

Petersen, R. C., & Jack, C. R. (2009). Imaging and biomarkers in early alzheimer's disease and mild cognitive impairment. *Clinical Pharmacology and Therapeutics*, *86*(4), 438–441. https://doi.org/10.1038/clpt.2009.166

Plinge, A., & Fink, G. A. (2014). Geometry calibration of distributed microphone arrays exploiting audio-visual correspondences. *European Signal Processing Conference*, (September), 116–120. https://doi.org/10.13140/2.1.4970.4328

Plis, S. M., Chekroud, A., Hjelm, D., Damaraju, E., Lee, J., Bustillo, J. R., Cho, K., Pearlson, G. D., Vince, D., Haven, N., Korea, S., & Haven, N. (2018). Reading the (functional) writing on the (structural) wall: multimodal fusion of brain structural and function via a deep neural network based

translation approach reveals novel impairments in schizophrenia, 734–747. https://doi.org/10.1016/j.neuroimage.2018.07.047.Reading

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Qi, S., Sui, J., Chen, J., Liu, J., Jiang, R., Silva, R., Iraji, A., Damaraju, E., Salman, M., Lin, D., Fu, Z., Zhi, D., Turner, J. A., Bustillo, J., Ford, J. M., Mathalon, D. H., Voyvodic, J., McEwen, S., Preda, A., … Calhoun, V. D. (2019). Parallel group ICA+ICA: Joint estimation of linked functional network variability and structural covariation with application to schizophrenia. *Human Brain Mapping*, *40*(13), 3795–3809. https://doi.org/10.1002/hbm.24632

Rad, B. B., Bhatti, H. J., & Ahmadi, M. (2017). An introduction to docker and analysis of its performance. *International Journal of Computer Science and Network Security (IJCSNS)*, *17*(3), 228.

Rafati, J., & Marica, R. F. (2020). Quasi-newton optimization methods for deep learning applications. *Deep learning applications* (pp. 9–38). Springer.

Rivet, B., Wang, W., Naqvi, S. M., & Chambers, J. A. (2014). Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, *31*(3), 125–134. https://doi.org/10.1109/MSP.2013.2296173

Saykin, A. J., Shen, L., Yao, X., Kim, S., Nho, K., Risacher, S. L., Ramanan, V. K., Foroud, T. M., Faber, K. M., Sarwar, N., Munsie, L. M., Hu, X., Soares, H. D., Potkin, S. G., Thompson, P. M., Kauwe, J. S., Kaddurah-Daouk, R., Green, R. C., Toga, A. W., & Weiner, M. W. (2015a). Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. *Alzheimer's and Dementia*, *11*(7), 792–814. https://doi.org/10.1016/j.jalz.2015.05.009

Saykin, A. J., Shen, L., Yao, X., Kim, S., Nho, K., Risacher, S. L., Ramanan, V. K., Foroud, T. M., Faber, K. M., Sarwar, N., Munsie, L. M., Hu, X., Soares, H. D., Potkin, S. G., Thompson, P. M., Kauwe, J. S., Kaddurah-Daouk, R., Green, R. C., Toga, A. W., & Weiner, M. W. (2015b). Ge-

netic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. https://doi.org/10.1016/j.jalz.2015.05.009

Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., Blennow, K., Soares, H., Simon, A., Lewczuk, P., Dean, R., Siemers, E., Potter, W., Lee, V. M., & Trojanowski, J. Q. (2009). Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology*, *65*(4), 403–413. https://doi.org/10.1002/ana.21610

Shivappa, S. T., Trivedi, M. M., & Rao, B. D. (2010). Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, *98*(10), 1692–1715. https://doi.org/10.1109/JPROC.2010.2057231

Silver, M., Montana, G., & Nichols, T. E. (2011). False positives in neuroimaging genetics using voxel-based morphometry data. *NeuroImage*, *54*(2), 992–1000. https://doi.org/10.1016/j.neuroimage.2010.08.049

Soheili-Nezhad, S., Beckmann, C., & Sprooten, E. (2021). Independent genomic sources of brain structure and function. *bioRxiv*. https://doi.org/10.1101/2021.01.06.425535

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., Corneveaux, J. J., DeChairo, B. M., Potkin, S. G., Weiner, M. W., & Thompson, P. M. (2010). Voxelwise genome-wide association study (vGWAS). *NeuroImage*, *53*(3), 1160–1174. https://doi.org/10.1016/j.neuroimage.2010.02.032

Stein, J. L., Hua, X., Morra, J. H., Lee, S., Hibar, D. P., Ho, A. J., Leow, A. D., Toga, A. W., Sul, J. H., Kang, H. M., Eskin, E., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., … Thompson, P. M. (2010). Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. *NeuroImage*, *51*(2), 542–554. https://doi.org/10.1016/j.neuroimage.2010.02.068

Stingo, F. C., Guindani, M., Vannucci, M., & Calhoun, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association*, *108*(503), 876–891. https://doi.org/10.1080/01621459.2013.804409

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the

Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, *12*(3), 1–10. https://doi.org/10.1371/journal.pmed.1001779

Sui, J., Adali, T., Li, Y.-O., Yang, H., & Calhoun, V. D. (2010). A review of multivariate methods in brain imaging data fusion. *Medical Imaging 2010: Biomedical Applications in Molecular, Structural, and Functional Imaging*, *7626*(March 2010), 76260D. https://doi.org/10.1117/12.843922

Sui, J., Adali, T., Pearlson, G. D., Clark, V. P., & Calhoun, V. D. (2009). A method for accurate group difference detection by constraining the mixing coefficients in an ICA framework. *Human Brain Mapping*, *30*(9), 2953–2970. https://doi.org/10.1002/hbm.20721

Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. A., Yu, Q., Clark, V. P., Castro, E., White, T., Mueller, B. A., Ho, B. C., Andreasen, N. C., & Calhoun, V. D. (2013). Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia. *NeuroImage*, *66*, 119–132. https://doi.org/10.1016/j.neuroimage.2012.10.051

Sui, J., Pearlson, G. D., Du, Y., Yu, Q., Jones, T. R., Chen, J., Jiang, T., Bustillo, J., & Calhoun, V. D. (2015). In Search of Multimodal Neuroimaging Biomarkers of Cognitive Deficits in Schizophrenia. *Biological Psychiatry*, *78*(11), 794–804. https://doi.org/10.1016/j.biopsych.2015.02.017

Sun, Z., Qiao, Y., Lelieveldt, B. P., & Staring, M. (2018). Integrating spatial-anatomical regularization and structure sparsity into SVM: Improving interpretation of Alzheimer's disease classification. *NeuroImage*, *178*(April 2018), 445–460. https://doi.org/10.1016/j.neuroimage.2018.05.051

Szefer, E., Lu, D., Nathoo, F., Beg, M. F., & Graham, J. (2017). Multivariate association between single-nucleotide polymorphisms in Alzgene linkage regions and structural changes in the brain: Discovery, refinement and validation. *Statistical Applications in Genetics and Molecular Biology*, *16*(5-6), 367–386. https://doi.org/10.1515/sagmb-2016-0077

Thomaz, C. E., Kitani, E. C., & Gillies, D. F. (2006). A maximum uncertainty LDA-based approach for limited sample size problems — with application to face recognition. *Journal of the Brazilian Computer Society*, *12*(2), 7–18. https://doi.org/10.1007/BF03192391

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, *36*(1), 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

Van Essen, D., Smith, S., Barch, D., & Consortium, H. (2013). The WU-Minn Human Connectome Project: An Overview David. *Bone*, *23*(1), 1–7. https://doi.org/10.1038/jid.2014.371

Vergara, V. M., Ulloa, A., Calhoun, V. D., Boutte, D., Chen, J., & Liu, J. (2014). A three-way parallel ICA approach to analyze links among genetics, brain structure and brain function. *NeuroImage*, *98*, 386–394. https://doi.org/10.1016/j.neuroimage.2014.04.060

Verghese, S. L., Liao, I. Y., Maul, T. H., & Chong, S. Y. (2021). An Empirical Study of Several Information Theoretic Based Feature Extraction Methods for Classifying High Dimensional Low Sample Size Data. *IEEE Access*, *9*, 69157–69172. https://doi.org/10.1109/ACCESS.2021.3077958

Vielzeuf, V., Lechervy, A., Pateux, S., & Jurie, F. (2019). Multilevel Sensor Fusion with Deep Learning. *IEEE Sensors Letters*, *3*(1), 1–12. https://doi.org/10.1109/LSENS.2018.2878908

Vincent, P. (2010). A Connection Between Score Matching and Denoising Autoencoders, 1–13.

Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., Restaino, R., & Wald, L. (2015). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(5), 2565–2586. https://doi.org/10.1109/TGRS.2014.2361734

Vogel, J. (2020). Data-driven network models to characterize the distribution and spread of tau in the Alzheimer ' s disease brain. (April).

Vounou, M., Nichols, T. E., & Montana, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage*, *53*(3), 1147–1159. https://doi.org/10.1016/j.neuroimage.2010.07.002

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., & Shen, L. (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics*, *28*(2), 229–237. https://doi.org/10.1093/bioinformatics/btr649

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., Donohue, M. C., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L., Thompson, P. M., Toga, A. W., & Trojanowski, J. Q. (2015). Impact of the Alzheimer's Disease Neuroimaging Initiative, 2004 to 2014. *Alzheimer's and Dementia*, *11*(7), 865–884. https://doi.org/10.1016/j.jalz.2015.04.005

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Salazar, J., Saykin, A. J., Shaw, L. M., Toga, A. W., & Trojanowski, J. Q. (2017). The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement. https://doi.org/10.1016/j.jalz.2016.10.006

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Nigel, J., Green, R. C., Harvey, D., Jr, C. R. J., Jagust, W., John, C., Petersen, R. C., Salazar, J., Saykin, A. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., Francisco, S., Francisco, S., Francisco, S., … Sciences, I. (2017). The Alzheimer's Disease Neuroimaging Initiative 3: continued innovation for clinical trial improvement. *13*(5), 561–571. https://doi.org/10.1016/j.jalz.2016.10.006.The

Weiner, M., & ADNI. (2013). The ADNI initiative: review of paper published since its inception. *Alzheimer Dementia*, *9*(5), e111–e194. https://doi.org/10.1016/j.jalz.2013.05.1769.The

Welvaert, M., & Rosseel, Y. (2014). A review of fMRI simulation studies. *PLoS ONE*, *9*(7). https://doi.org/10.1371/journal.pone.0101953

Wimley, W. C. (2017). ęŏ HHS PUBLIC ACCESS. *Physiology behavior*, *176*(10), 139–148.

Wolf, R. C., Rashidi, M., Fritze, S., Kubera, K. M., Northoff, G., Sambataro, F., Calhoun, V. D., Geiger, L. S., Tost, H., & Hirjak, D. (2020). A Neural Signature of Parkinsonism in Patients With Schizophrenia Spectrum Disorders: A Multimodal MRI Study Using Parallel ICA. *Schizophrenia Bulletin*, 1–10. https://doi.org/10.1093/schbul/sbaa007

Wyman, B. T., Harvey, D. J., Crawford, K., Bernstein, M. A., Carmichael, O., Cole, P. E., Crane, P. K., Decarli, C., Fox, N. C., Gunter, J. L., Hill, D., Killiany, R. J., Pachai, C., Schwarz, A. J., Schuff, N., Senjem, M. L., Suhy, J., Thompson, P. M., Weiner, M., & Jack, C. R. (2013). Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's and Dementia*, *9*(3), 332–337. https://doi.org/10.1016/j.jalz.2012.06.004

Xanthis, C. G., & Aletras, A. H. (2019). CoreMRI: A high-performance, publicly available MR simulation platform on the cloud. *PLoS ONE*, *14*(5), 1–26. https://doi.org/10.1371/journal.pone.0216594

Xu, L., Pearlson, G., & Calhoun, V. D. (2009). Joint source based morphometry identifies linked gray and white matter group differences. *NeuroImage*, *44*(3), 777–789. https://doi.org/10.1016/j.neuroimage.2008.09.051

Xue, H., Jiang, W., Miao, C., Yuan, Y., Ma, F., Ma, X., Wang, Y., Yao, S., Xu, W., Zhang, A., & Su, L. (2019). DeepFusion, 151–160. https://doi.org/10.1145/3323679.3326513

Yadav, S. P., & Yadav, S. (2020). Image fusion using hybrid methods in multimodality medical images. *Medical and Biological Engineering and Computing*, *58*(4), 669–687. https://doi.org/10.1007/s11517-020-02136-6

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., & Narayan, V. A. (2012). Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurology*, *12*. https://doi.org/10.1186/1471-2377-12-46

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, *55*(3), 856–867. https://doi.org/10.1016/j.neuroimage.2011.01.008

Zhang, J., Qin, G., & Liu, Y. (2012). Speech separation in the vehicle environment based on fastica algorithm. *Journal of Multimedia*, *7*(1).

Zhang, R. Y., Wei, X. X., & Kay, K. (2020). Understanding multivariate brain activity: Evaluating the effect of voxelwise noise correlations on population codes in functional magnetic resonance imaging. *PLoS Computational Biology*, *16*(8 August), 1–29. https://doi.org/10.1371/journal.pcbi.1008153

Zhang, W., Zhang, Y., Zhai, J., Zhao, D., Xu, L., Zhou, J., Li, Z., & Yang, S. (2018). Multi-source data fusion using deep learning for smart refrigerators. *Computers in Industry*, *95*, 15–21. https://doi.org/10.1016/j.compind.2017.09.001

Zhang, Y., Simon-Vermot, L., Araque Caballero, M. T., Gesierich, B., Taylor, A. N., Duering, M., Dichgans, M., & Ewers, M. (2016). Enhanced resting-state functional connectivity between core memory-task activation peaks is associated with memory impairment in MCI. *Neurobiology of Aging*, *45*, 43–49. https://doi.org/10.1016/j.neurobiolaging.2016.04.018

Zheng, X., Shi, J., Li, Y., Liu, X., & Zhang, Q. (2016). Multi-modality stacked deep polynomial network based feature learning for Alzheimer's disease diagnosis. *Proceedings - International Symposium on Biomedical Imaging*, *2016-June*, 851–854. https://doi.org/10.1109/ISBI.2016.7493399

Zhou, T., Liu, M., Thung, K. H., & Shen, D. (2019). Latent Representation Learning for Alzheimer's Disease Diagnosis with Incomplete Multi-Modality Neuroimaging and Genetic Data. *IEEE Transactions on Med-*

ical Imaging, *38*(10), 2411–2422. https://doi.org/10.1109/TMI.2019.2913158

Zhou, T., Thung, K. H., & Zhu, X. (2017). Feature Learning and Fusion of Multimodality Neuroimaging and Genetic Data for Multi-status Dementia Diagnosis. *10541*(November), 106–113. https://doi.org/10.1007/978-3-319-67389-9

Zhu, W. (2016). GWAS of Secondary Imaging Phenotypes from the ADNI. *Physiology behavior*, *176*(1), 139–148. https://doi.org/10.1016/j.physbeh.2017.03.040