STATISTICAL METHODS FOR FUNCTIONAL MAGNETIC RESONANCE IMAGING DATA

by

Amanda F. Mejia

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2016

© Amanda F. Mejia 2016

All rights reserved

Advisor: Martin Lindquist, Ph.D.

Co-Advisor: Brian Caffo, Ph.D.

Thesis Readers: James J. Pekar, Ph.D. and Brion Maher, Ph.D.

Acknowledgments

First and foremost, to Martin Lindquist and Brian Caffo: I could not have asked for a better pair of advisors. Martin, thank you for your generosity and wisdom and all the opportunities you've created for me. I am lucky to have had the chance to learn from one of the foremost experts in fMRI. Brian, thank you for always having an open door and for being a wonderful mentor. I have learned a tremendous amount about how to think about statistics from you, both in and out of the classroom.

To Elizabeth Sweeney, Taki Shinohara, Mary Beth Nebel, and John Muschelli: when I started my thesis research, I knew nothing about brain imaging. You helped me climb the learning curve, and I thank you for being available and patient. You are all amazing collaborators, and I hope to continue working together in the future.

To Jim Pekar, Stewart Mostofsky, and Daniel Reich: thank you for welcoming me into your labs and for helping me learn about radiology and neurology. I still have much to learn, but working with you has given me an understanding of the science that is essential to developing sensible and impactful statistical methods.

To Scott Zeger: thank you for your enthusiastic and unwavering support during

my job search. Your help meant the world and was truly instrumental. I also feel fortunate to have worked alongside you teaching the Master's methods series last year. You showed me how to connect with students and create an interactive, dynamic class environment.

To all the faculty at Johns Hopkins Biostatistics who have cheered me on along the way, including Karen Bandeen-Roche, Dan Scharfstein, Ciprian Crainiceanu, Constantine Frangakis, Ani Eloyan, Vadim Zipunnikov, Betsy Ogburn, and Ingo Ruczinski: thank you for believing in me and for all the support you provide to the students in the department.

To all my fellow students and friends at Hopkins: you guys made this an incredible experience. My time in the Ph.D. program at Hopkins has been some of the best years of my life. From lunches and tea time to pirate boating and happy hours, it's been an awesome ride. I'll miss seeing you every day, but I know we'll keep in touch.

Dedication

This thesis is dedicated to my family. To Jorge, who challenged me to go for a PhD and who has been there every step along the way. Thank you for making me laugh and keeping life fun. To my mom, who has always been my solid rock and who never settled for less than my very best. Thank you for being a shining example of strength and character to me and everyone around you ... and for the many king cakes! To my dad, who was my first (and maybe best) math teacher and remains my biggest cheerleader. Thank you for guiding me to where I am today. To Andrew, the yin to my yang and the smartest, most talented person I know. Thank you for sharing your wisdom and talent and for all the fun times (past and future!) in New Orleans. And to Aida, Jorge, and Michelle: thank you for welcoming me into your family and for all the beautiful moments in Colombia.

Contents

Acknowledgments **iii**

List of Tables **x**

List of Figures **xi**

1 Introduction **1**

2 Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators **6**

2.1 Introduction 8

2.2 Methods 11

2.2.1 Shrinkage Methods 12

2.2.1.1 Shrinkage Model 12

2.2.1.2 Variance Component Estimation 15

2.2.2 Subject-level Parcellations 21

2.2.3 Performance of shrinkage methods 22

2.2.3.1 Reliability of functional connectivity measures 22
2.2.3.2 Reliability of parcellations 22
2.2.4 Data 23
2.2.4.1 Simulated Data 23
2.2.4.2 Real fMRI Data 28
2.3 Results 31
2.3.1 Simulation Results 31
2.3.2 Real fMRI Dataset Results 39
2.4 Discussion 48

3 PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data 56

3.1 Introduction 57
3.2 Methods 62
3.2.1 Dimension Reduction 62
3.2.2 Principal components leverage 64
3.2.2.1 Motivation and definition 64
3.2.2.2 Thresholding rule 66
3.2.3 Principal components robust distance 67
3.3 Simulation Study 70
3.3.1 Construction of baseline scans 70
3.3.2 Artifact-free images 71

3.3.3 Images with artifacts 72
3.4 Experimental Data Results 75
3.4.1 fMRI Dataset 76
3.4.2 Identification of subject-level brain networks through GICA and dual regression 79
3.4.3 Measuring reproducibility of subject-level brain networks 83
3.4.4 Effect of outlier removal on reproducibility 84
3.5 Discussion 86

4 Scan Length, Shrinkage and Reliability of Resting-State Functional Connectivity in the Human Connectome Project 90

4.1 Introduction 90
4.2 Methods 97
4.2.1 Data and connectivity estimation 97
4.2.1.1 Human Connectome Project Data 97
4.2.1.2 Connectivity matrix estimation 99
4.2.2 Reliability of connectivity estimates 107
4.3 Results 110
4.4 Discussion 116

5 Discussion 129

Bibliography **132**

List of Tables

2.1 Simulation parameters varied (one at a time) and the values they are varied over. The default value is the value at which each parameter is fixed while the other parameters are varied. 26

3.1 Varying parameter and range of values for each artifact. At each iteration, the artifact intensity is generated from a uniform distribution with the specified range. 73

3.2 For each dataset, the total number of subjects (N), number of subjects that passed quality inspection (N^+), number of subjects used to perform GICA (N_{ICA}), and the number of signal GICA networks identified (Q^+). 77

List of Figures

2.1 Simulated signal image consisting of four 25-voxel clusters for the group and two subjects. 24

2.2 Data setup to perform shrinkage and evaluate the performance using both test-retest data (a) and single-scan data (b). Resting-state fMRI time series for 5 voxels are shown. 27

2.3 The degree of shrinkage under the default simulation parameters of $I = 20$, $T = 200$, $\rho = 0.05$, and $\sigma_X^2 = 0.02$, using either single session data (left) or test-retest data (right). 31

2.4 The MSE of correlation estimates (a), and Dice coefficient of similarity of parcellations with the true parcellations (b) under the default simulation parameters of $I = 20$, $T = 200$, $\rho = 0.05$, and $\sigma_X^2 = 0.02$, using either single session data (left) or test-retest data (right). The percent decrease in the median MSE of the shrinkage estimates, compared with the MSE of the raw estimates, is reported below each boxplot in (a); the percent increase in the Dice coefficient of the shrinkage-based parcellations, compared with the Dice coefficients of the raw parcellations, is reported below each boxplot in (b). 32

2.5 Dice similarity (with each subject's true parcellation) within two different regions: the region of the image where all subjects share the same parcellation (a), and the region of the image where subjects' parcellations differ (b). Results were computed under the default simulation parameters of $I = 20$, $T = 200$, $\rho = 0.05$, $\sigma_X^2 = 0.02$, using either single session or test-retest data. 36

2.6 Effect of each simulation parameter on the degree of shrinkage, MSE of the correlation estimates, and Dice similarity of the subject-level parcellations. Each point shows the median value over all subjects and iterations. Error bands show ± 2 standard errors around the median (may not be visible due to narrow width). 37

2.7 Degree of shrinkage (percent weighting of the group mean, averaged over all voxel-pairs) by noise variance estimation method and type of dataset (single session or test-retest) used to perform shrinkage. For the "common" and "global" noise variance methods, all subjects have the same shrinkage parameter at each voxel-pair, so the boxplot shows a single value. [12pt] 40
2.8 MSE of raw and shrinkage correlation estimates, by noise variance estimation method and the type of dataset (single session or test-retest) used to perform shrinkage. Each dotted line shows the MSE for a single subject's raw and shrinkage estimates, and the boxplots show the distributions over all subjects. The percent decrease in the median MSE of each shrinkage estimate (compared to the median MSE of the raw estimate) is reported below each boxplot. 40
2.9 Dice coefficients of similarity (with test set parcellations) of parcellations produced using raw and shrinkage correlation estimates. Results are shown by noise variance estimation method and the type of dataset (single session or test-retest) used to perform shrinkage. Each dotted line shows the Dice coefficients for a single subject, and the boxplots show the distributions over subjects. The percent increase in the median Dice coefficient of each shrinkage parcellation (compared to the median Dice coefficient of the raw parcellation) is reported below each boxplot. 41
2.10 Subject-level parcellations of the motor cortex from 3 example subjects. 44
2.11 Group-level parcellations from sessions 1 and 2. 45
2.12 The relationship between scan length and noise variance. Results are shown using untranformed correlations (C, shown in red) and Fisher-transformed correlations (Z, shown in teal) to compute the noise variance. Panel (a) shows diminishing session-to-session variance as scan length increases. Panel (b) shows how the adjustment factor $\theta(t) = \sigma_U^2(t)/\sigma_U^2(t/2)$ changes as t increases. The fitted line relating $\theta(t)$ to $\log(t)$ is shown in black. 47

3.1 Examples of each artifact type. Figure (a) shows a normal volume on the left and a volume with a spike artifact on the right. Figure (b) shows the image mask before and after rotation. The spike, rotation and ghosting artifacts are generated from the maximum artifact intensity as described in Table 3.1; the banding artifact is generated randomly as described in Table 3.1. [10pt] 74
3.2 Sensitivity and specificity of each method in the presence of artifacts by simulation setup. Each line represents a different scan for each artifact type and shows the mean across 1000 iterations with a 95% bootstrap confidence interval. 75

3.3 For a single subject, the motion parameters, leverage function, and robust distance function. Leverage and distance outliers are indicated in red. Below the plot, the volumes corresponding to the spikes at time points 60, 90, 135 and 150 (shaded on the plot) are shown. Three of the spikes are leverage and distance outliers, while the spike at time point 90 is only a leverage outlier. Obvious banding artifacts are seen at time points 60 and 150, a moderate banding artifact is seen at time point 135, and no visible artifact is apparent at time point 90. While the artifact at time point 150 would be detected using motion measures, the other spikes would likely go undetected using only motion. 80

3.4 Distribution of the percent of volumes in each scan identified as outliers using each outlier removal method. The leverage method tends to identify no or very few outliers more often than the distance method. 81

3.5 Estimates and 95% confidence intervals for the model coefficients. The left-hand plot displays the coefficients for each outlier removal method (α_m); both methods result in statistically significant improvement to reproducibility of subject-level brain networks identified through ICA, and removal of leverage outliers results in greater improvement than removal of distance outliers. The right-hand plot displays the fixed effects for each dataset (γ_k) and illustrates the heterogeneity in reproducibility of ICA results across ABIDE datasets. [10pt] 85

3.6 Estimates and 95% confidence intervals for the model coefficients after stratifying by quality inspection results. The left-hand plot shows that while both groups of subjects benefit from outlier removal, subjects who failed quality inspection (QC= 0) tend to benefit more than those who passed quality inspection (QC= 1). The right-hand plot shows that subjects who failed quality inspection tend to have lower baseline reproducibility than those who passed quality inspection. 87

4.1 Illustration of the data for a single subject. For five regions, the full time series, consisting of 4800 volumes, is shown. Each time series consists of four sessions, occurring over two visits. In the GICA provided in the HCP data release, the sessions were reordered so that both visits are concatenated in the LR/RL order. 99

4.2 Illustration of data setup for intersession and end-point reliability analysis for a single subject i. For intersession reliability analysis, we are interested in how similar the connectivity estimates $W_{i1}^{(\ell)}$, $\tilde{W}_{i1}^{(\ell)}$ and $\tilde{W}_{i1}^{*(\ell)}$ are to the full visit 2 raw estimate $W_{i2}^{(L)}$, $L = 2400$, as ℓ varies from 300 to 2400. For end-point reliability analysis, we are interested in how close the connectivity estimates $W_{i1}^{(\ell)}$, $\tilde{W}_{i1}^{(\ell)}$ and $\tilde{W}_{i1}^{*(\ell)}$ are to the full visit 1 raw estimate $W_{i1}^{(L)}$, as ℓ varies from 300 to 2400. 100

4.3 We summarize reliability at three different resolutions: omnibus, seed-level and edge-level. *Omnibus reliability* is computed as the median edge-level reliability across all unique pairs of regions, resulting in a single scalar summary measure of reliability. *Seed-level reliability* is computed as the median edge-level reliability within each seed, resulting in a single map of reliability. *Edge-level reliability* is computed as the median reliability across subjects at each edge, resulting in a map of reliability for each seed region. [10pt] 109
4.4 Three selected seed regions selected from model order 300, lying respectively in the visual cortex (IC 16), the somatomotor cortex (IC 21), and the DMN (IC 30). The visual seed is located in the bilateral lingual gyrus; the somatomotor seed is located in the left dorsolateral pre- and post-central gyri; the DMN seed is located in the posterior cingulate cortex (PCC). 110
4.5 Comparison of omnibus intersession and end-point reliability of raw connectivity estimates by scan length at each model order (25, 50, 100, 200, 300). Smaller values of absolute percent error signify greater reliability. For each model order and at every scan length, end-point reliability is a poor proxy for intersession reliability, as end-point error greatly underestimates the true intersession error. Furthermore, this bias increases sharply as the scan length increases to $L = 2400$. This is because the estimate produced from $\ell < L$ volumes becomes more similar to the reference produced from L volumes, eventually reaching equality at $\ell = L$. 119
4.6 **Top panel:** Degree of shrinkage versus scan length for each model order (25, 50, 100, 200, 300) (median over all connections). For oracle shrinkage, the degree of shrinkage tends to decrease as scan length increases. However, for single-session shrinkage, the degree of shrinkage exhibits an initial decrease followed by an increase. This is likely artifact of the unique HCP acquisition protocol, in which the phase encoding method changes midway through each session, leading to an inflation of the within-subject variance estimated from a single session. **Bottom panel:** Comparison of omnibus intersession reliability of raw and shrinkage connectivity estimates by scan length for each model order. Smaller values of absolute percent error signify greater reliability. Both single-session and oracle shrinkage estimates exhibit greater intersession reliabilty than raw estimates across all model orders and scan lengths. 120
4.7 Seed-level reliability. 121
4.8 Edge-level reliability, visual seed. 122
4.9 Edge-level reliability, somatomotor seed. 123
4.10 Edge-level reliability, DMN seed. 124

4.11 **Seed-level variance components and degree of shrinkage.** For each quantity, the median value over all connections with a given seed is displayed. As scan length increases, between-subject variance stays relatively constant, while within-subject variance and hence the degree of shrinkage decreases. 125
4.12 Edge-level variance components and degree of shrinkage, visual seed. . 126
4.13 Edge-level variance components and degree of shrinkage, somatomotor seed. 127
4.14 Edge-level variance components and degree of shrinkage, DMN seed. . 128

Chapter 1

Introduction

Understanding how the brain functions is one of the most important goals in science and medicine today. For example, the BRAIN (Brain Research through Advancing Innovative Neurotechnologies) Initiative,[1] which was started in 2013 by President Obama, has pledged hundreds of millions of dollars to the development of technologies to enhance understanding of brain function. Additionally, a number of governmental agencies, laboratories, foundations and private companies have recently announced major programs devoted to the study of the brain. Such initiatives hold great promise, as neurological diseases and disorders, such as autism, Alzheimer's and sports-related brain injury, present major challenges to the health and welfare of affected individuals and their families, and in many cases effective treatments remain elusive. Recent advances in various neurotechnologies have now made it possible to study brain function from a multitude of different angles, in order to understand the neurological changes

associated with such disorders. These technologies, along with the growing interest in the study of brain function, may make it possible to identify ways to treat, or even prevent or cure, many neurological and psychiatric disorders.

One of the technologies being used to study brain function noninvasively in humans is functional magnetic resonance imaging (fMRI).[2] fMRI measures the blood oxygen-level dependent (BOLD) response to neuronal activity at moderate temporal and spatial resolution. In a typical fMRI run, a series of brain volumes each consisting of roughly $100,000$ volumetric elements ("voxels") are collected every 2 seconds. While neuronal signals occur on the order of milliseconds, the haemodynamic response occurs over several seconds, peaking around 6 seconds following the neuronal signal.[3] Therefore, a temporal resolution of 2 seconds per volume allows for the detection of neuronal activation in specific areas of the brain. By contrast, MRI techniques used to study structural brain features, such as T_1-weighted images, takes several minutes to acquire.[4] The rapid acquisition time of fMRI data is made possible through reductions in spatial resolution and signal-to-noise ratio; therefore, noise levels in fMRI data are relatively high.[2] Furthermore, fMRI data is subject to many sources of systematic noise ("artifacts") due to subject motion, scanner instabilities, and processing errors.[2,5]

fMRI data therefore presents a triad of statistical and computational challenges. First, there are many sources of variability due to noise and artifacts, changes in the cognitive state of the subject over time, and inter-individual differences. Second,

fMRI data exhibits a complex spatial and temporal dependence structure. Third, a single fMRI run can consist of 10 million to 1 billion data points, depending on the run length, temporal resolution and voxel size, and modern fMRI datasets consist of hundreds of subjects.[4,6–9] Therefore, fMRI data is large, and its analysis presents computational challenges, particularly for complex methods or those that require combining information across subjects.

Historically, the most popular use of fMRI data has been to study the location and intensity of brain activation in response to a particular task or stimulus. Task fMRI has shed great insight into the function of different regions of the brain and the association between brain activity with certain disorders, traits and behaviors. For example, Sharer et al. (2015)[10] found that brain regions involved in visuomotor sequence learning activated less in children with autism compared with typically developing children during a reaction time task; such findings help shed light on the brain mechanisms related to impairments in skill development observed in subjects with autism.

However, in recent years there has been increasing interest in functional connectivity, which is defined as simultaneous neuronal activation between anatomically distinct regions of the brain[11] and is typically measured using "resting-state fMRI" (rsfMRI), in which the subject is not asked to perform an explicit task.[12] Research on resting-state functional connectivity (rsFC) has historically focused on analyzing rsfMRI data at the group level in order to understand the organization of typical hu-

man brains. For example, the use of group independent components analysis (ICA) to identify resting-state networks, defined as large regions displaying coordinated patterns of BOLD activation, was popularized by Beckmann et al. (2005).[13] Other common measures of functional connectivity are correlation between the fMRI time series of different voxels or regions,[14] and parcellations or clusterings of voxels into functionally similar regions.[15]

Recently, there is growing interest in studying the organization and connectivity of the brains of individual subjects.[16,17] However, high levels of noise and artifacts present in fMRI data, combined with relatively short rsfMRI runs (5-10 minutes typically), make it difficult to accurately estimate connectivity at the individual level.[18] Furthermore, short- and long-term changes in the cognitive state of the subject present an additional source of variation that results in low reproducibility of subject-level rsFC measures.[19]

In this thesis, I address these issues and make several contributions to the analysis of rsFC at the subject level. In Chapter 2, I develop shrinkage estimates of pairwise connectivity between different voxels or regions. Specifically, I propose a method of estimating within-subject variance using only a single fMRI run from each subject. I apply the proposed methods to subject-level estimates of voxel-wise connectivity within the motor cortex and demonstrate the benefits of shrinkage on the reliability of resulting parcellations of the motor cortex.

In Chapter 3, I propose a novel outlier removal method for high-dimensional data.

I apply this method to identify artifacts in rsfMRI data, and I demonstrate the benefits of artifact removal on the reliability of subject-level resting-state networks estimated using ICA.

In Chapter 4, I return to the theme of shrinkage and propose a new method to estimate the within-subject variance of connectivity based on separating two sources of within-subject variance: variance of true connectivity over time and sampling variance. I apply the proposed methods to subject-level estimates of pairwise connectivity between 300 ICA-identified regions using the Human Connectome Project (HCP), a large, state-of-the-art fMRI dataset. I demonstrate that shrinkage results in improved reliability of subject-level connectivity estimates, even when 30 minutes of high-quality, high temporal resolution rsfMRI data is collected for each subject.

Chapter 2

Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators

A recent interest in resting state functional magnetic resonance imaging (rsfMRI) lies in subdividing the human brain into anatomically and functionally distinct regions of interest. For example, brain parcellation is often a necessary step for defining the network nodes used in connectivity studies. While inference has traditionally been performed on group-level data, there is a growing interest in parcellating single subject data. However, this is difficult due to the inherent low signal-to-noise ratio of

rsfMRI data, combined with typically short scan lengths. A large number of brain parcellation approaches employ clustering, which begins with a measure of similarity or distance between voxels. The goal of this work is to improve the reproducibility of single-subject parcellation using shrinkage-based estimators of such measures, allowing the noisy subject-specific estimator to "borrow strength" in a principled manner from a larger population of subjects. We present several empirical Bayes shrinkage estimators and outline methods for shrinkage when multiple scans are not available for each subject. We perform shrinkage on raw inter-voxel correlation estimates and use both raw and shrinkage estimates to produce parcellations by performing clustering on the voxels. While we employ a standard spectral clustering approach, our proposed method is agnostic to the choice of clustering method and can be used as a pre-processing step for any clustering algorithm. Using two datasets – a simulated dataset where the true parcellation is known and varies across subjects and a test-retest dataset consisting of two 7-minute resting-state fMRI scans from 20 subjects – we show that subject-level parcellations produced from shrinkage correlation estimates have higher validity and reliability than those produced from raw correlation estimates. Validity of parcellations is assessed through simulated data and is shown to improve by up to 30% overall due to shrinkage, and by up to 20% within regions where true subject-level parcellations differ. Application to test-retest data shows that shrinkage increases the reproducibility of parcellations of the motor cortex by up to 30%.

2.1 Introduction

There has been a long-standing interest in subdividing the human brain into anatomically and functionally distinct regions. Previously these subdivisions, or parcellations, were based primarily on mapping anatomical features from post-mortem brains.[20] More recently, the use of resting-state functional magnetic resonance imaging (rsfMRI) has provided the means for performing parcellation on living subjects using functional information.[21–23]

There are several potential reasons for the increased interest in functional parcellation of the brain. First, it provides an atlas that can be used to more accurately compare inter-subject fMRI time series by incorporating functional and anatomical features into inter-subject registration approaches.[15] Second, it allows for dimension reduction in fMRI analysis by grouping together functionally similar voxels, which not only reduces computational burdens, but also alleviates the problem of multiple comparisons and overly conservative family-wise error rate (FWER) corrections.[24] Third, the identification of functionally homogeneous regions of interest (ROIs) is necessary for defining meaningful brain network nodes.[25]

Many methods have been used to functionally parcellate the brain. These include, among others, fuzzy C-means,[26] independent components analysis,[27–29] expectation maximization,[30] hierarchical clustering,[31–33] spectral clustering,[22,34] and K-means clustering.[35] The goal of clustering is to group together items that are similar to each other and separate items that are dissimilar from each other. As such, all clus-

tering methods for brain parcellation employ a measure of distance or similarity that is used to classify voxels into coherent clusters. There are many such metrics available, including euclidian distance, correlation, and eta-squared,[16,34] and the choice of metric will have a direct effect on the result of clustering. Moreover, these metrics are subject to error whenever the underlying data are measured with noise, and the degree of noise may have a strong effect on clustering results. While the noise levels of rsfMRI data may be sufficiently low when the data from many subjects is combined or averaged to form a group-level parcellation, the noise in a single subject's data is substantially higher. The primary approaches to overcome this limitation have been collecting greater amounts of rsfMRI data on a single subject (30-60 minutes rather than the standard 5-10 minutes)[16,17,33] and utilizing constrained clustering algorithms.[22,33,36] For example, [33] proposed a subject-level clustering method in which a set of stable seeds is grown into an initial parcellation that is further clustered using a hierarchical approach that enforces spatial contiguity. However, in this paper and others, reliability is assessed on a single subject with a large amount of scan time. In contrast, most rsfMRI data are collected on much shorter intervals, making validation and replication of such approaches hard for practical purposes. Thus, the generalizability of such methods to more than one subject and their reliability on scans of shorter length are still open questions. Moreover, accurate assessment of reliability and validity of constrained clustering methods is difficult as the ground truth is unknown, and constraints artificially inflate reliability metrics by reducing

the degrees of freedom of the problem.[33]

Our proposal is to directly improve the reliability of distance metrics by using shrinkage estimators. Advantages of this approach are that the amount of scan time required to produce reliable subject-level parcellations is greatly reduced; resulting data can be used with standard, efficiently implemented clustering algorithms; and clustering results are a product of only the data itself, and not of external constraints.

The goal of this work is to investigate whether shrinkage-based methods can improve the reproducibility of subject-level parcellations generated using rsfMRI data. Shrinkage methods allow noisy subject-level estimators to "borrow strength" in a principled manner from a larger population of subjects. In the statistics literature, shrinkage estimators[37,38] have been shown to improve the mean squared error (MSE) of many traditional estimators by shrinking the estimators towards some fixed constant value, such as the population mean. Shrinkage is implicit in Bayesian inference, penalized likelihood inference and multi-level models[39] and is directly related to the empirical Bayes estimators commonly used in neuroimaging.[40–42] Recently, [43] applied shrinkage in the context of rsfMRI seed-based connectivity analysis and showed a nearly 30% average improvement in intra-subject reliability of correlation estimates, with improvement of over 50% in several subjects.

This paper extends the work of [43] and offers a number of methodological contributions. First, we develop shrinkage estimators for the full voxel-by-voxel distance matrix, which is required for clustering. Second, we propose methods for constructing

shrinkage estimators in the practical case where only a single scan is available for some or all subjects. Third, we explore the utility of shrinkage estimators where the degree of shrinkage performed is subject-dependent to account for differences in intra-subject variability. Finally, we perform clustering on shrinkage estimates and demonstrate improved test-retest reliability of the resulting subject-level parcellations.

To illustrate the feasibility of our proposed approach, we focus on one simple unsupervised learning technique, namely normalized spectral clustering. We generate simulated data, where the true parcellation is known and is allowed to vary across subjects. In addition, we apply the method to real test-retest resting-state fMRI data from 20 subjects and show that we can increase the reliability of single-subject parcellations of the motor cortex by up to 30%.

2.2 Methods

In this section, we discuss shrinkage methods and illustrate how they can be used for single-subject parcellation. We begin by detailing our shrinkage model and methods for estimating the relevant parameters. We perform shrinkage on measures of functional connectivity and obtain parcellations from these measures using two sets of data, which are described below. The first is a simulation, for which the ground truth parcellation is known. The second is a test-retest dataset of resting-state fMRI scans, for which we use part of each subject's data as a proxy for the unknown ground truth

parcellation.

2.2.1 Shrinkage Methods

2.2.1.1 Shrinkage Model

The quantity of interest for each subject i, $i = 1, \ldots, I$, is the true $V \times V$ functional connectivity matrix C_i, where V is the number of voxels in a pre-defined region of interest (ROI) of the brain, which we wish to parcellate. Our goal is to use information from the other $I - 1$ subjects to provide stable estimates of the subject-specific value C_i; the idea is to find compromise estimators between the unbiased but highly variable raw subject-specific connectivity estimators and the biased, but much smoother, mean connectivity estimators.

More precisely, let v and v' be two distinct voxels in a particular ROI, let $X_i(v, v')$ be the true quantity of interest for subject i (e.g. pairwise correlation), and let $W_{ij}(v, v')$ be the observed value of $X_i(v, v')$ obtained from session j. The classical measurement error model[44] is

$$W_{ij}(v, v') = X_i(v, v') + U_{ij}(v, v'),$$

where $U_{ij}(v, v')$ is subject-level measurement error for subject i during session j at voxel-pair (v, v'). We assume that $X_i(v, v')$ and $U_{ij}(v, v')$ are independent for all i and j. We further assume that the $X_i(v, v')$, $i = 1, \ldots, I$, are independently drawn

from a population distribution with between-subject variance $\sigma_X^2(v, v')$, i.e.

$$X_i(v, v') \sim N\left\{\mu_X(v, v'), \sigma_X^2(v, v')\right\}.$$

Finally, we assume that for each subject i, the $U_{ij}(v, v')$ are independently and identically distributed for all j and

$$U_{ij}(v, v') \sim N\left\{0, \sigma_{U,i}^2(v, v')\right\}.$$

Returning our attention to the quantity of interest, the shrinkage estimator of $X_i(v, v')$ using session j is given by

$$\tilde{W}_{ij}(v, v') = \lambda_i(v, v') * \bar{W}_j(v, v') + \{1 - \lambda_i(v, v')\} * W_{ij}(v, v'),$$

where $\bar{W}_j(v, v') = \frac{1}{I}\sum_{i=1}^{I} W_{ij}(v, v')$, and the shrinkage parameter $\lambda_i(v, v')$ represents the relationship between within-subject variance $\sigma_{U,i}^2(v, v')$ and between-subject variance $\sigma_X^2(v, v')$:

$$\lambda_i(v, v') = \frac{\sigma_{U,i}^2(v, v')}{\sigma_X^2(v, v') + \sigma_{U,i}^2(v, v')}.$$

Here $\lambda_i(v, v')$ ranges between 0 and 1 and represents the relative weight given to the group mean $\bar{W}_j(v, v')$ compared to the raw subject-level estimate $W_{ij}(v, v')$. As the within-subject variance $\sigma_{U,i}^2(v, v')$ increases, the subject-level information is less reliable, the shrinkage parameter increases, and the shrinkage estimate is more weighted

towards the group mean. As between-subject variance $\sigma^2_X(v, v')$ increases, the group mean becomes less representative of the true subject-level values, so shrinkage is less beneficial, $\lambda_i(v, v')$ decreases, and the shrinkage estimate is more weighted towards the subject-level observation. $\lambda_i(v, v')$ is estimated directly from the data and is designed to achieve the optimal balance between the raw subject-level estimate and the group mean.

We employ the Fisher-transformed correlation coefficient as our measure of functional connectivity, which fulfills the model assumptions of Normality and independence of $X_i(v, v')$ and $U_{ij}(v, v')$.[43] Given a correlation estimate r, the Fisher-transformed estimate $z(r)$ is given by the transformation $r \rightarrow \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$, and is approximately Normally distributed with variance $(T-3)^{-1}$, where T is the number of time points in the scan. By contrast, the sampling variance of an untransformed correlation coefficient decreases as the true correlation increases, which violates the signal-noise independence assumption. After shrinkage is performed, we then apply the inverse transformation $z \rightarrow \frac{exp(2z)-1}{exp(2z)+1}$ to obtain an estimate of correlation for the purposes of parcellation. However, for completeness we also evaluate the benefits of applying shrinkage directly to the untransformed correlation estimates.

Note that the within-subject variance $\sigma^2_{U,i}(v, v')$ is allowed to vary across subjects i. This allowance stems from the observation that within-subject variance comes from multiple sources, including sampling variability and session-to-session variability. Sampling variability reflects the error of an estimate (e.g. correlation estimate)

around the value it is estimating and is directly related to the number of time points used to compute the estimate. For example, as described above, a Fisher-transformed correlation estimate has asymptotic sampling variance $\frac{1}{T-3}$. It follows that as the number of time points increases to infinity, this source of variability will decrease to zero. While sampling variability may be roughly equal across subjects with equal scan lengths, session-to-session variability may vary across subjects. This type of variability reflects differences in a subject's true functional connectivity across multiple scanning sessions due to variations in brain behavior. Moreover, session-to-session variability may dominate sampling variability. In our sample, for example, we find that within-subject variance of the Fisher-transformed correlation matrices tends to be around five times larger than the theoretical sampling variance. We therefore allow $\sigma^2_{U,i}(v, v')$ to differ across subjects. However, as there are drawbacks to estimating $\sigma^2_{U,i}(v, v')$ completely separately for each subject, we propose several other methods of estimating within-subject ("noise") variance. In total, we propose four methods, which are discussed in detail below.

2.2.1.2 Variance Component Estimation

Henceforth, we will use the terms *within-subject* and *noise* variance interchangeably, and we will use the terms *between-subject* and *signal* variance interchangeably.

Noise Variance Estimation

To estimate the noise variance, it is ideal to have access to multiple scanning sessions

for each subject ("test-retest data"). However, in many cases only a single scan is available for each subject. For these situations, we propose two approaches. First, create a *pseudo*-test-retest dataset by dividing each subject's single scan into two sub-scans, each containing half of the original time points. Second, estimate a global measure of within-subject variance for all voxel-pairs and subjects using an external test-retest dataset, a subset of subjects for which multiple sessions are available, or through psuedo-test-retest data combined with extrapolation. This will be discussed in more detail below.

We now describe four noise variance estimators, which we denote common (C), individual (I), scaled (S), and global (G). The common estimator assumes that $\sigma^2_{U,i}(v,v') \equiv \sigma^2_U(v,v')$ is the same across all subjects, while allowing variation across voxel-pairs. The individual and scaled estimators allow $\sigma^2_{U,i}(v,v')$ to vary across subjects and voxel-pairs. The individual estimator estimates $\sigma^2_{U,i}(v,v')$ separately for each subject, while the scaled estimator starts with the common noise variance estimator and adjusts it by a subject-specific factor to produce a different noise variance estimate for each subject. The global estimator assumes that $\sigma^2_{U,i}(v,v') = \sigma^2_U$ is the same across all subjects and voxel-pairs. This estimator is primarily intended for the case when limited or no test-retest data is available for the dataset of interest. All four noise variance estimators can be computed using true test-retest data or pseudo-test-retest data created from a single scan.

Common Noise Variance Letting $D_i(v,v') = W_{i2}(v,v') - W_{i1}(v,v')$, the common

noise variance can be estimated as[43,44]

$$\hat{\sigma}_{U,i}^{2(C)}(v,v') \equiv \hat{\sigma}_{U}^{2(C)}(v,v') := \frac{1}{2} Var_i\left\{D_i(v,v')\right\} = \frac{1}{2(I-1)} \sum_{i=1}^{I} \left\{D_i(v,v') - \bar{D}(v,v')\right\}^2,$$

where $\bar{D}(v,v') = \frac{1}{I}\sum_{i=1}^{I} D_i(v,v')$. To see this, note that

$$\begin{aligned} Var_i\left\{D_i(v,v')\right\} &= Var_i\left\{W_{i2}(v,v') - W_{i1}(v,v')\right\} \\ &= 2Var_i\left\{U_{ij}(v,v')\right\}, \end{aligned}$$

so $Var_i\left\{U_{ij}(v,v')\right\} = \frac{1}{2} Var_i\left\{D_i(v,v')\right\}$.

Individual Noise Variance Given two estimates $W_{ij}(v,v')$, $j = 1,2$, of the term $X_i(v,v')$, the individual noise variance can be estimated as follows:

$$\begin{aligned} \hat{\sigma}_{U,i}^{2(I)}(v,v') &:= \frac{1}{J-1} \sum_{j=1}^{J} \left\{U_{ij}(v,v') - \bar{U}_i(v,v')\right\}^2 \\ &= \left\{U_{i1}(v,v') - \bar{U}_i(v,v')\right\}^2 + \left\{U_{i2}(v,v') - \bar{U}_i(v,v')\right\}^2 \\ &= \frac{1}{2}\left\{U_{i2}(v,v') - U_{i1}(v,v')\right\}^2 \\ &= \frac{1}{2}\left\{\left[W_{i2}(v,v') - X_i(v,v')\right] - \left[W_{i1}(v,v') - X_i(v,v')\right]\right\}^2 \\ &= \frac{1}{2}\left\{W_{i2}(v,v') - W_{i1}(v,v')\right\}^2 \end{aligned}$$

Scaled Noise Variance Given the common noise variance estimate $\hat{\sigma}_{U}^{2(C)}(v,v')$, we use

a subject-specific scaling factor γ_i to obtain the scaled noise variance estimate

$$\hat{\sigma}^{2(S)}_{U,i}(v,v') = \gamma_i \times \hat{\sigma}^{2(C)}_{U}(v,v').$$

The scaling factor γ_i is equal to the test-retest MSE of subject i relative to the average test-retest MSE over all subjects:

$$\gamma_i = \bar{D^2}_i / \bar{\bar{D}^2},$$

where $\bar{D^2}_i = \frac{2}{V(V-1)} \sum_{v>v'} D_i^2(v,v')$ and $\bar{\bar{D}} = \frac{1}{I} \sum_{i=1}^{I} \bar{D^2}_i$. To see that this provides a reasonable estimate of an individual subject's noise variance, notice that

$$\hat{\sigma}^{2(I)}_{U,i}(v,v') = \frac{1}{2} D_i^2(v,v')$$

and

$$\hat{\sigma}^{2(C)}_{U}(v,v') \approx \frac{1}{I} \sum_{i=1}^{I} \hat{\sigma}^{2(I)}_{U,i}(v,v') = \frac{1}{2I} \sum_{i=1}^{I} D_i^2(v,v') = \frac{1}{2} \bar{D^2}(v,v').$$

The approximate equality above follows from the fact that the two terms have the same expected value (Appendix). Therefore,

$$\frac{\hat{\sigma}^{2(I)}_{U,i}(v,v')}{\hat{\sigma}^{2(C)}_{U}(v,v')} \approx \frac{D_i^2(v,v')}{\bar{D^2}(v,v')},$$

so across all voxel-pairs (v,v'), the ratio of the the individual noise variance to the

common noise variance is approximately equal to γ_i. The benefit of the scaled noise variance, compared with the individual noise variance, is that it is based upon a more stable estimate of the noise variance, $\hat{\sigma}_U^{2(C)}(v, v')$ and requires the estimation of much fewer terms.

Global Noise Variance We estimate the global noise variance as the mean value of the common noise variance over all $V(V-1)/2$ unique voxel-pairs:

$$\hat{\sigma}_{U,i}^{2(G)}(v, v') \equiv \hat{\sigma}_U^{2(G)} := \frac{2}{V(V-1)} \sum_{v>v'} \hat{\sigma}_U^{2(C)}(v, v').$$

If pseudo-test-retest data is used to compute $\hat{\sigma}_U^{2(G)}$, the noise variance will be overestimated due to shorter scan length and should be adjusted. Let $\sigma_U^2(t)$ be the expected global noise variance for a scan of length t, and let $\hat{\sigma}_U^2(t)$ be an estimate of $\sigma_U^2(t)$. When we use pseudo-test-retest data by splitting a scan of length T, we obtain an estimate of $\sigma_U^2(\frac{T}{2})$. To obtain an estimate of $\sigma_U^2(T)$, let

$$\theta(T) = \frac{\sigma_U^2(T)}{\sigma_U^2(\frac{T}{2})}.$$

Then $\sigma_U^2(T)$ can be estimated as

$$\hat{\sigma}_U^2(T) = \hat{\theta}(T) \times \hat{\sigma}_U^2(\tfrac{T}{2}). \tag{2.1}$$

The adjustment factor $\theta(T)$ can be estimated if multiple scans are available for a

subset of subjects. However, we also provide an estimate of $\theta(t)$ as follows. Using the test-retest fMRI dataset described below, for $t = \{1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 7\}$ minutes, we resample scans of length t within each scanning session and estimate $\sigma_U^2(t)$ for each resampled dataset. We then compute the average over all resampled datasets to obtain $\hat{\sigma}_U^2(t)$. We compute $\hat{\theta}(t) = \hat{\sigma}_U^2(t)/\hat{\sigma}_U^2(\frac{t}{2})$ for $t = \{2, 3, 4, 5, 6, 7\}$. This gives a curve estimating the relationship between T and $\theta(T)$ for $T = \{2, 3, 4, 5, 6, 7\}$. To extrapolate to other scan lengths, we fit a regression curve relating log scan length to $\theta(t)$:

$$\theta(t) = \beta_0 + \beta_1 \times \log(t). \tag{2.2}$$

Using our coefficient estimates of β_0 and β_1, one can estimate the appropriate scaling factor $\theta(T)$ for scans of length T by simply plugging in their values into equation 2.2. One can then use this scaling factor to adjust the global noise variance estimate obtained from pseudo-test-retest data using equation 2.1 and hence obtain an appropriate noise variance estimate.

Signal Variance Estimation

The between-subject or *signal* variance $\sigma_X^2(v, v')$ is equal to the difference between the total variance and noise variance. While noise variance may vary across subjects, signal variance is a population parameter. Therefore, even if we choose to estimate the noise variance individually, we use the common or global noise variance estimator to obtain an estimate of the signal variance.

The total variance $\sigma^2_W(v, v')$ at voxel-pair (v, v') is estimated as[44]

$$\hat{\sigma}^2_W(v, v') := \frac{1}{J}\sum_{j=1}^{J} \hat{\sigma}^2_{W_j}(v, v') = \frac{1}{J(I-1)}\sum_{j=1}^{J}\sum_{i=1}^{I} \left\{W_{ij}(v, v') - \bar{W}_j(v, v')\right\}^2.$$

We can then simply estimate the signal variance as

$$\hat{\sigma}^2_X(v, v') = \hat{\sigma}^2_W(v, v') - \hat{\sigma}^2_U(v, v'),$$

Shrinkage Parameter Estimation

We obtain four estimators for the shrinkage parameter $\lambda_i(v, v')$ corresponding to the four noise variance estimators:

$$\lambda_i^{(M)}(v, v') = \frac{\sigma_{U,i}^{2(M)}(v, v')}{\sigma^2_X(v, v') + \sigma_{U,i}^{2(M)}(v, v')},$$

where $M \in \{I, S, C, G\}$.

2.2.2 Subject-level Parcellations

Shrinkage estimates of correlation were obtained by applying the inverse-Fisher transformation to the shrinkage estimate of the Fisher-transformed correlation. That is, we first Fisher-transform the raw estimate, perform shrinkage, then apply the inverse Fisher transformation. We then generated subject-level parcellations by performing spectral clustering as described by [45], using the raw and shrinkage correlation esti-

mates as a metric of similarity. We chose to look for five clusters based on previously published findings (using the same test-retest resting state data set) that this was the optimal number of functional partitions for the precentral gyrus in terms of test-retest reliability.[34]

2.2.3 Performance of shrinkage methods

2.2.3.1 Reliability of functional connectivity measures

We define reliability of a functional connectivity measure (e.g. correlation) as the MSE between the estimated measure $\hat{C}_i$ and the truth C_i. We assess the performance of a shrinkage estimate as the percent decrease in MSE of the shrinkage estimate relative to the MSE of the raw estimate.

2.2.3.2 Reliability of parcellations

We define reliability of a parcellation estimate as the Dice similarity coefficient compared with the true parcellation. Let $\hat{A}_i$ be the adjacency matrix obtained from clustering, where $\hat{A}_i(v, v') = 1$ if v and v' are assigned to the same parcel for subject i and 0 otherwise. Let A_i be the adjacency matrix corresponding to the true parcellation of subject i. Dice's coefficient of similarity between the estimated and true parcellations is defined as

$$S(A_i, \hat{A}_i) = \frac{2|A_i \cap \hat{A}_i|}{|A_i| + |\hat{A}_i|}.$$

We assess the performance of a parcellation obtained from a shrinkage estimate of functional connectivity as the percent increase in Dice coefficient relative to the parcellation obtained from the corresponding raw estimate.

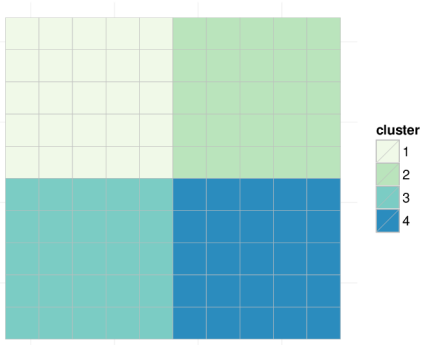
For the simulation described below, the true connectivity matrix and parcellation are known quantities, so we can compute exactly the MSE and Dice coefficient of the raw and shrinkage estimates. For our fMRI dataset, however, the true connectivity matrix and parcellation are unknown. We get around this by reserving part of each subject's data as a proxy for the truth, which we call the test set. We compute raw and shrinkage estimates for the remaining data and compare both estimates to the raw estimate from the test set.

2.2.4 Data

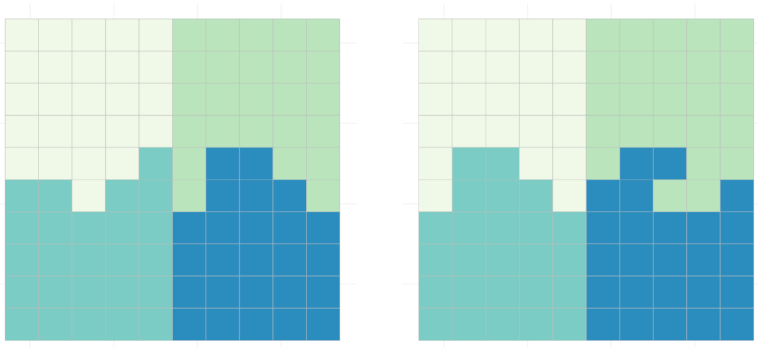
2.2.4.1 Simulated Data

We simulated a 10-by-10 voxel parcellation consisting of four clusters, each cluster corresponding to one quadrant of the image at the group level (Figure 2.1a). Each subject-level parcellation was generated by randomly permuting cluster labels along the borders of clusters 1 and 3 and clusters 2 and 4 (Figure 2.1b).

The true correlation matrices within each cluster were assumed to follow an *exchangeable* structure, meaning that each pair of voxels within cluster k of subject i at session j has the same pairwise correlation ρ_{ijk}. We further assumed that each



(a) Group parcellation



(b) Subject parcellations

Figure 2.1: Simulated signal image consisting of four 25-voxel clusters for the group and two subjects.

subject i has a fixed within-cluster correlation value $\rho_{ijk} \equiv \rho_i$ across all sessions j and clusters k. Let ρ represent the population average within-cluster correlation and $z(\rho)$ represent the Fisher-transformation of ρ. Random variation among subjects $i = 1, \ldots, I$ was introduced by adding Gaussian noise to $z(\rho)$, then applying the inverse Fisher-transformation, $z^{-1}(\cdot)$:

$$\rho_i = z^{-1}\big(z(\rho) + u_i\big),\ u_i \sim N(0, \sigma_X^2)$$

As negative within-cluster correlations do not make sense under an exchangeable correlation structure, any negative correlations generated through this process were resampled until all ρ_i were positive. Between-cluster correlations were assumed to equal zero. The true correlation matrix C_i^* for each subject i was therefore constructed as a block diagonal matrix with the four diagonal blocks corresponding to the within-cluster correlation matrices for subject i, and the off-diagonal blocks set to zero.

For each subject $i = 1, \ldots, I$ and session $j = 1, 2$, a time series of length T was generated for each voxel in the cluster. Each time point was drawn from a multivariate Normal distribution with mean zero and covariance matrix $\Sigma_i \equiv C_i$. As correlations are agnostic to within-voxel variance and temporal correlation, these were not considered. The observed voxel time series were combined to form a 3D image (2D x time) for each subject i and session j. Observed correlation matrices $\hat{C}_{ij}$ were computed from the 3D images.

Parameter	Values	Default Value
Number of subjects (I)	$(10, 20, 30, 100)$	20
Length of time series (T)	$(100, 200, 300, 1000)$	200
Within-cluster correlation (ρ)	$(0.01, 0.05, 0.1)$	0.05
Between-subject variance (σ_X^2)	$(0.01, 0.02, 0.03, 0.04, 0.05)$	0.02

Table 2.1: Simulation parameters varied (one at a time) and the values they are varied over. The default value is the value at which each parameter is fixed while the other parameters are varied.

We varied the following simulation parameters: number of subjects (I), length of time series (T), population average within-cluster correlation (ρ), and between-subject or signal variance (σ_X^2). The parameter values tested are given in Table 2.1. Parameter values were changed one at a time, while all other parameters were fixed at a default value. The default value of each parameter is also shown in Table 2.1.

We simulated 1000 datasets for each of the 13 unique designs defined by the parameter values in Table 2.1. For each dataset, we computed the observed correlation matrix $\hat{C}_{i1}$ from each subject's first session. We performed shrinkage on these matrices using each noise variance estimation method $M \in \{I, S, C, G\}$, described in Section 2.2.1.2. For each method M, we computed the noise variance assuming that two sessions $j = 1, 2$ were available for each subject and again assuming that only one session $j = 1$ was available for each subject. Let $\tilde{C}_i^{M,\ell}$ be the shrinkage estimate obtained by shrinking estimate $\hat{C}_{i1}$ using noise variance estimation method M and assuming availability of $\ell = 1, 2$ scans for each subject. We performed clustering as described in Section 2.2.2 on the raw correlation matrices $\hat{C}_{i1}$, $i = 1, \ldots, I$ and each shrinkage correlation matrix $\tilde{C}_i^{M,\ell}$, $M \in \{I, S, C, G\}$, $\ell = 1, 2$ and $i = 1, \ldots, I$.

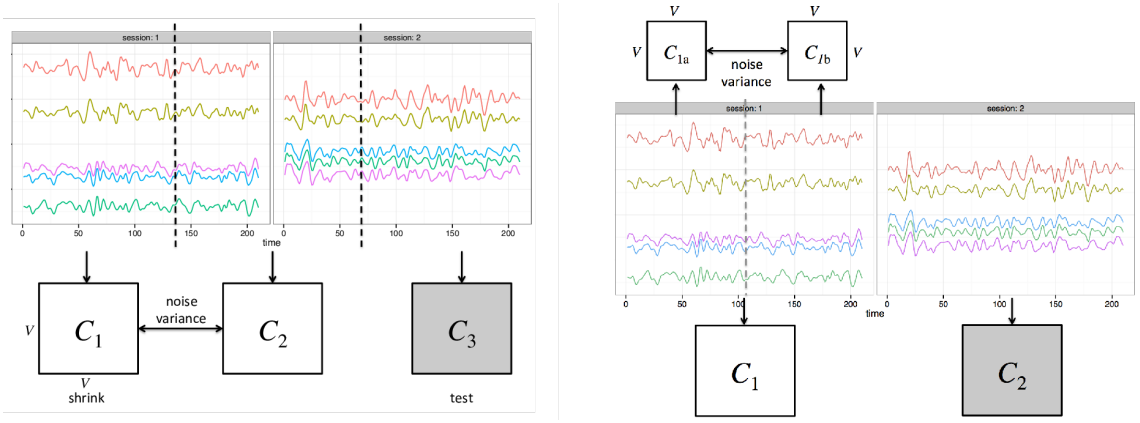


Figure 2.2: Data setup to perform shrinkage and evaluate the performance using both test-retest data (a) and single-scan data (b). Resting-state fMRI time series for 5 voxels are shown.

Analysis S1: Performance of shrinkage estimates and parcellations

Using the default design specified in Table 2.1, we computed the degree of shrinkage, performance of the raw and shrinkage correlation matrices, and performance of parcellations obtained using shrinkage estimates, as described in Section 2.2.3. The degree of shrinkage is defined as the average value of $\lambda_i(v, v')$ over all voxel-pairs.

Analysis S2: Sensitivity to simulation parameters

For each alternative design specified in Table 2.1, we computed the degree of shrinkage and performance of correlation estimates and parcellations to understand how each parameter affects the degree of shrinkage towards the group mean and the impact of the shrinkage procedure on the reliability of similarity metrics and parcellations.

2.2.4.2 Real fMRI Data

We use data from the publicly available Multi-Modal MRI Reproducibility Resource (`http://www.nitrc.org/projects/multimodal`). Image acquisition parameters are described in detail elsewhere.[46] In short, a high resolution T1-weighted MPRAGE and two 7-minute resting state scans were acquired from 21 healthy adult volunteers. Both resting state scans were acquired on the same day, and in between the two scans the subject exited the scanner.

The anatomical scan was registered to the first functional volume and normalized to Montreal Neuological Institute (MNI) space using SPM8's unified segmentation/normalization procedure. Resting state data were adjusted for slice time acquisition, and rigid body realignment estimates were calculated with respect to the first functional volume to account for participant motion. The non-linear spatial transformation estimated during the unified segmentation/normalization was then applied to the functional data along with the estimated rigid body realignment parameters and resulted in 2-mm isotropic voxels. Each resting state scan was then temporally detrended on a voxelwise basis. An aCompCor strategy was used to estimate spatially coherent noise components, as this method has been shown to effectively attenuate physiological noise[47] as well as motion artifacts.[48] The aCompCor noise components were regressed from the resting state data along with linearly detrended versions of the rigid body realignment parameters and their first derivatives (computed by backward differences). Functional data were then spatially smoothed (6-mm FWHM

Gaussian kernel) and temporally filtered using a .01-.1 Hz pass band. Data from one participant were excluded from analysis due to a misalignment of the first and second resting-state scans.

Our region of interest (ROI) for this experiment is the precentral gyrus (M1), a key component of the motor control network and a region whose gross functional organization has long been recognized.[49] The precentral gyrus ROI was selected from the "Type II Eve Atlas"[50] and contained $V = 7396$ voxels after being transformed to MNI space.

To evaluate performance of shrinkage using test-retest data to estimate the variance components, we split each subject's data into three parts as illustrated in Figure 2.2a. As a total of 420 images (14 minutes) were collected over the two sessions, each of the three parts consisted of 140 images (4 minutes and 40 seconds). For the middle third, each session was demeaned before concatenating the time series. The first two parts were used to compute the variance components. We performed shrinkage on the first part and reserved the third part as the test set.

To evaluate performance of shrinkage using only a single scan from each subject, we split the first session in half to create a pseudo-test-retest dataset, which we use to estimate the noise variance. We performed shrinkage on the first session and reserved the second session as the test set (Figure 2.2b). The psuedo-test-retest dataset was used to compute the common, individual and scaled noise variance estimates; the global noise variance was computed using both full sessions.

For both setups, we first computed the V-by-V observed correlation matrices $\hat{C}_{ij}$ for each subject i and session $j = 1, 2$ or part $j = 1, 2, 3$. The estimates produced using the first session or part, $\hat{C}_{i1}$, were treated as the "raw" estimates. We then applied the Fisher-transformation to obtain $\hat{Z}_{ij}$ for all i and j. We computed the variance components and shrinkage parameter $\lambda_i^{(M)}(v, v')$, $M \in \{I, S, C, G\}$, and performed shrinkage using $\lambda_i^{(M)}(v, v')$ on the $\hat{Z}_{i1}$. We then applied the inverse Fisher transform to obtain shrinkage estimates $\tilde{C}_{i1}$. For completeness, we also performed the same procedure directly on the $\hat{C}_{i1}$ without Fisher-transforming to obtain shrinkage estimates $\tilde{\tilde{C}}_{i1}$.

Analysis R1: Performance of shrinkage estimates

The performance of the shrinkage correlation estimates computed from session or part 1 was assessed as described in Section 2.2.3, using the raw estimates from the test set as a proxy for the unknown ground truth.

Analysis R2: Performance of parcellations

Subject-level parcellations were generated using both raw and shrinkage correlation estimates, as described in Section 2.2.2. We used the estimates obtained by applying shrinkage directly to the correlation estimates, as these were shown to have better performance than shrinkage estimates obtained through the Fisher-transformed correlation estimates. The performance of the parcellations generated using shrinkage estimates from session or part 1, relative to the parcellations generated using the corresponding raw estimates, was assessed as described in Section 2.2.3, using the

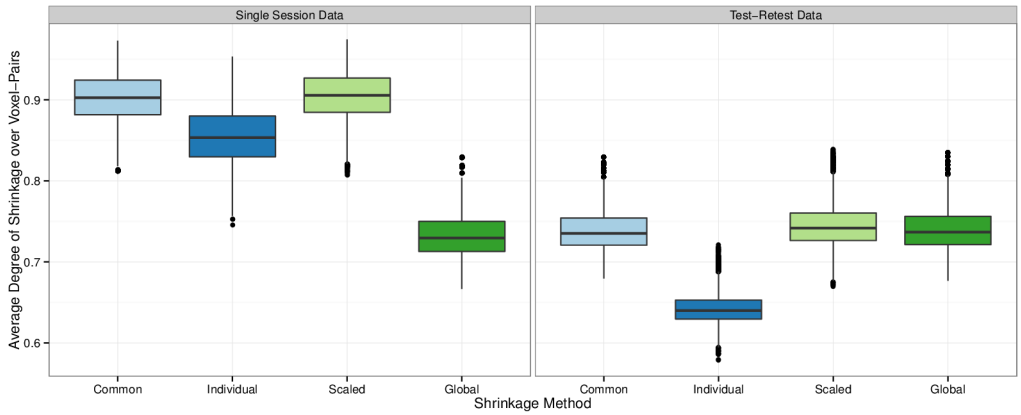


Figure 2.3: The degree of shrinkage under the default simulation parameters of $I = 20$, $T = 200$, $\rho = 0.05$, and $\sigma_X^2 = 0.02$, using either single session data (left) or test-retest data (right).

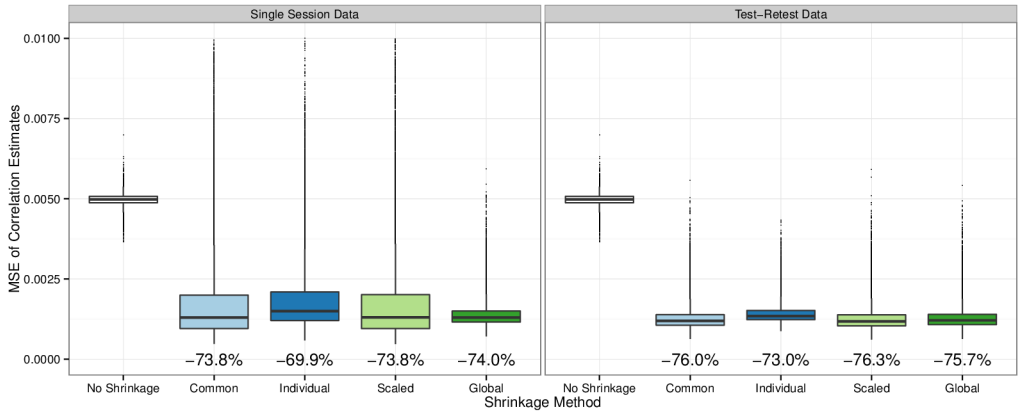
parcellations generated from the raw estimates from the test set as a proxy for the ground truth.

2.3 Results

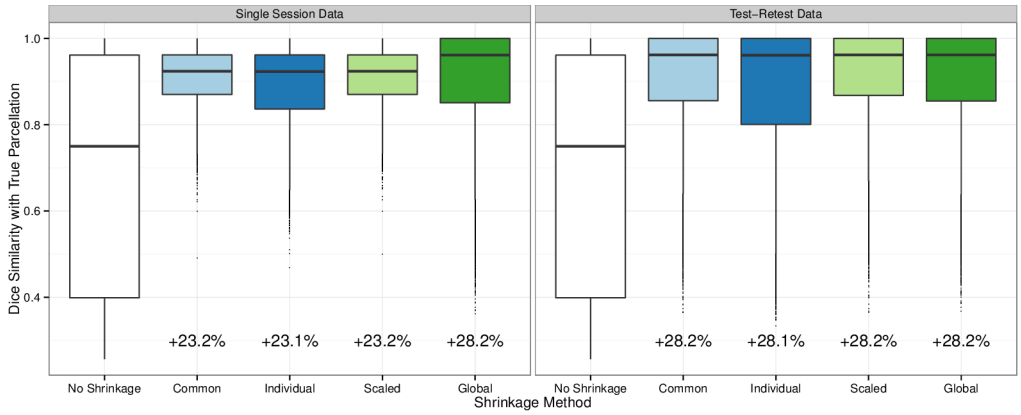
2.3.1 Simulation Results

Analysis S1: Performance of shrinkage estimates and parcellations

Figures 2.3-2.4 show the degree of shrinkage, MSE of raw and shrinkage correlation estimates, and Dice similarity of parcellations with the true parcellations, under the default simulation settings ($I = 20$, $T = 200$, $\rho = 0.05$, $\sigma_X^2 = 0.02$). Results are shown by shrinkage method and the type of dataset used, single session (left) or test-retest data (right), and each boxplot shows the distribution of values over all subjects and



(a) MSE of Correlation Estimates



(b) Dice Similarity of Parcellations

Figure 2.4: The MSE of correlation estimates (a), and Dice coefficient of similarity of parcellations with the true parcellations (b) under the default simulation parameters of $I = 20$, $T = 200$, $\rho = 0.05$, and $\sigma_X^2 = 0.02$, using either single session data (left) or test-retest data (right). The percent decrease in the median MSE of the shrinkage estimates, compared with the MSE of the raw estimates, is reported below each boxplot in (a); the percent increase in the Dice coefficient of the shrinkage-based parcellations, compared with the Dice coefficients of the raw parcellations, is reported below each boxplot in (b).

simulation iterations. In Figure 2.3, the degree of shrinkage for a given subject and simulation iteration was computed as the average value of the shrinkage parameter $\lambda_i(v, v')$ over all voxel-pairs (v, v'). The degree of shrinkage was sensitive to the noise variance method employed and the type of dataset (single session or test-retest) used to perform shrinkage. Using a single session to perform shrinkage, the median degree of shrinkage over all subjects and iterations was 90.3% with a common noise variance; 85.3% with individual noise variance; 90.6% with scaled noise variance; and 73.0% with a global noise variance. Using test-retest data to perform shrinkage, the median degree of shrinkage was 73.5% with a common noise variance; 64.0% with individual noise variance; 74.2% with scaled noise variance; and 73.7% with a global noise variance.

Figure 2.4a shows the MSE of the raw and shrinkage correlation estimates. The average improvement in MSE due to shrinkage was fairly uniform across shrinkage methods under the default simulation parameters. However, when a single session was used to perform shrinkage, there were more large outliers than when test-retest data was used, except when the global noise variance estimator was employed. The median MSE over all subjects and iterations of the raw correlation estimates was 0.00498. Using a single session to perform shrinkage, the median MSE of the shrinkage correlation estimates was 0.00130 (73.9% lower) with a common noise variance; 0.00150 (69.9% lower) with individual noise variance; 0.00131 (73.7% lower) with scaled noise variance; and 0.00130 (73.9% lower) with a global noise variance. Using test-retest

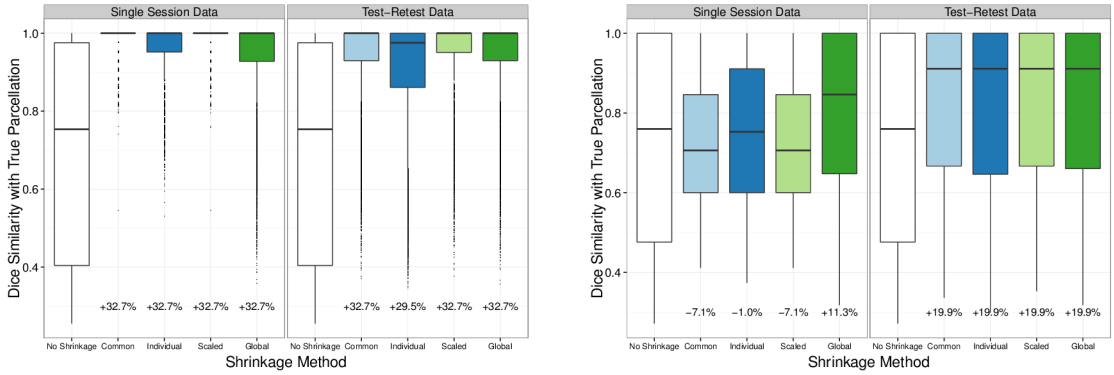
data to perform shrinkage, the median MSE of the shrinkage correlation estimates was 0.00119 (76.1% lower) with a common noise variance; 0.00134 (73.1% lower) with individual noise variance; 0.00118 (76.3% lower) with scaled noise variance; and 0.00121 (75.7% lower) with a global noise variance.

Figure 2.4b shows the Dice coefficient of similarity with the true parcellations of the parcellations generated from the raw and shrinkage correlation estimates. The improvement in Dice coefficient due to shrinkage was fairly uniform across shrinkage methods but was maximized when test-retest data was used to perform shrinkage and when the global noise variance estimator was used with single session data. The median Dice coefficient over all subjects and iterations of the raw parcellations was 0.750. Using a single session to perform shrinkage, the median Dice coefficient of the shrinkage-based parcellations was 0.924 (23.2% higher) with a common noise variance; 0.923 (23.1% higher) with individual noise variance; 0.924 (23.2% higher) with scaled noise variance; and 0.961 (28.1% higher) with a global noise variance. Using test-retest data to perform shrinkage, the median Dice coefficient of the shrinkage-based parcellations was 0.962 (28.3% higher) with a common noise variance; 0.961 (28.1% higher) with individual noise variance; 0.962 (28.3% higher) with scaled noise variance; and 0.962 (28.3% higher) with a global noise variance.

Figure 2.5 illustrates the performance of the shrinkage-based parcellations in two different regions of the images displayed in Figure 2.1. The first is the region in which all subjects share the same parcellation, namely the top four rows and bottom four

rows of the image. In this region, shrinkage towards the group mean will clearly be beneficial, since the group mean is representative of the truth for each subject. The second is the region in which subject-level differences in parcellations are allowed to occur, namely the middle two rows of the image. In this region, it is less clear whether shrinkage will result in parcellations that are closer to the true subject-level parcellations. In Figure 2.5a, we see that the improvement in the Dice coefficient within the first region was large for all shrinkage methods. We also see that the methods that use a single session resulted in the greatest improvement, since these methods tend to over-estimate the noise variance and thus over-shrink. In this region, since all subjects have the exact same parcellation, total shrinkage towards the group mean will be the most beneficial. In Figure 2.5b, as expected, we see that the improvement in the Dice coefficient within the second region was less dramatic. In fact, for the shrinkage methods that tend to over-shrink, there was a reduction in the median Dice coefficient compared with the raw parcellations. However, when test-retest data was used to perform shrinkage or the global noise variance estimator was used with single session data, there was an improvement in the median Dice coefficient. When test-retest data was used to perform shrinkage, the median Dice coefficient increased by 19.9% due to shrinkage for all noise variance estimators; when a single session was used to perform shrinkage and the global noise variance estimator was used, the median Dice coefficient increased by 11.3% due to shrinkage.

Analysis S2: Sensitivity to simulation parameters



(a) Region where subjects share parcellation

(b) Region where subjects differ

Figure 2.5: Dice similarity (with each subject's true parcellation) within two different regions: the region of the image where all subjects share the same parcellation (a), and the region of the image where subjects' parcellations differ (b). Results were computed under the default simulation parameters of $I = 20$, $T = 200$, $\rho = 0.05$, $\sigma_X^2 = 0.02$, using either single session or test-retest data.

Figure 2.6 shows how the degree of shrinkage (top row), MSE of the correlation estimates (middle row), and the Dice similarity of the parcellations (bottom row) change with varying simulation parameters. The first column shows the effect of varying the number of subjects; the second column shows the effect of varying the time series length; the third column shows the effect of varying the strength of inter-voxel correlations; and the fourth column shows the effect of varying the degree of similarity across subjects. Each color represents a different shrinkage method, and results are shown in the case of both single session data (top panel of each plot) and test-retest data (bottom panel of each plot). Each point represents the mean over all subjects and iterations, and 95% confidence intervals are shown as grey bands, which may not be visible due to their narrow width.

We see that the degree of shrinkage tends to decrease as the time series length

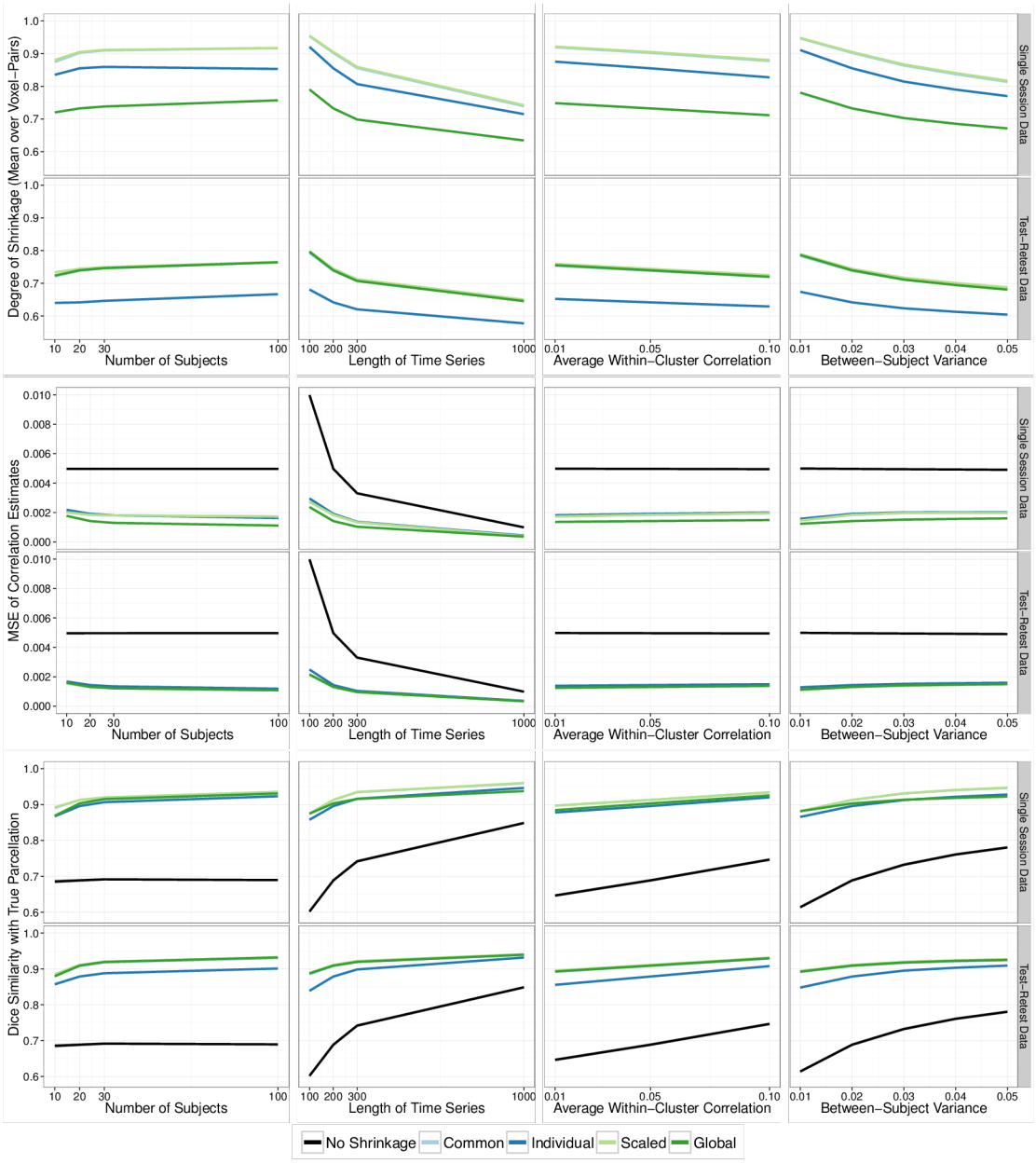


Figure 2.6: Effect of each simulation parameter on the degree of shrinkage, MSE of the correlation estimates, and Dice similarity of the subject-level parcellations. Each point shows the median value over all subjects and iterations. Error bands show ±2 standard errors around the median (may not be visible due to narrow width).

T increases, as the within-cluster correlation ρ increases, and as the between-subject variance σ_X^2 increases (Figure 2.6, top row). This is expected, since the shrinkage parameter is defined as the ratio of within-subject variance to total (within-subject plus between-subject) variance, and higher values of T and ρ reduce the within-subject variance. There is a weak increase in the degree of shrinkage as the number of subjects I increases. This reflects bias in the estimation of $\lambda_i(v, v')$, a non-linear function of variance components, a bias that diminishes as the sample size increases.

The MSE of the raw estimates is primarily related to the time series length T, since as T increases, sampling variability decreases (Figure 2.6, middle row). As T increases, the MSE of all estimators decreases, and the MSE of the raw estimator approaches, but does not achieve, the MSE of the shrinkage estimators. As the number of subjects I increases, there is also a reduction in the MSE of the shrinkage estimators, which is due to the increase in the degree of shrinkage associated with larger sample size.

The Dice coefficient of the raw parcellations increases as the time series length T increases, as the within-cluster correlation ρ increases, and as the between-subject variance σ_X^2 increases (Figure 2.6, bottom row). Similar to the results for MSE of the correlation estimates, as T increases, the Dice coefficient of the raw parcellations approaches that of the shrinkage-based parcellations; unlike the results for MSE, the shrinkage-based parcellations still dramatically outperform the raw parcellations even at $T = 1000$. The Dice coefficient of the shrinkage-based parcellations increases along

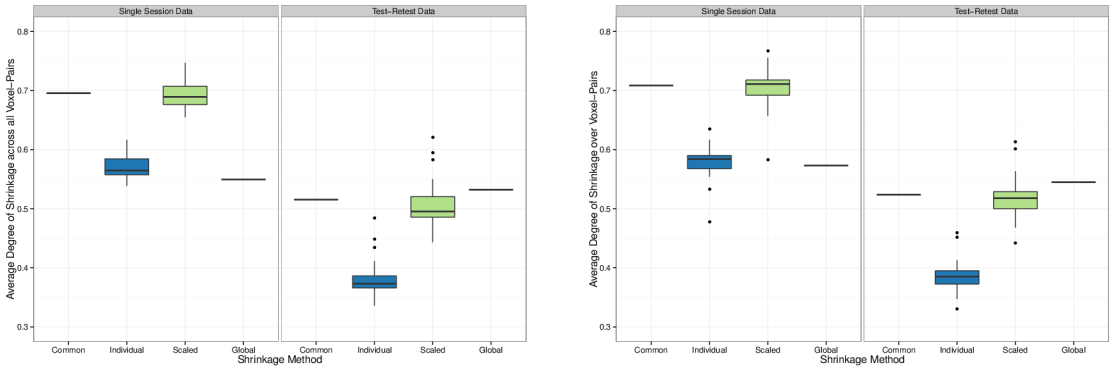
with the Dice coefficient of the raw parcellations as T, ρ, or σ_X^2 is increased. As the number of subjects I increases, the Dice coefficient of the shrinkage estimators increases, which is again due to the increase in the degree of shrinkage associated with larger sample size.

2.3.2 Real fMRI Dataset Results

Figure 2.7 shows the degree of shrinkage towards the group mean performed on the real rsfMRI dataset by noise variance estimation method and the type of dataset (single session or test-retest) used to perform shrinkage. As in the simulation results, the degree of shrinkage for subject i was computed as the mean value of the shrinkage parameter $\lambda_i(v, v')$ over all voxel-pairs (v, v'). Each boxplot shows the distribution of these values over subjects. For those methods that computed only a group-level shrinkage parameter (the common and global noise variance methods), the shrinkage parameter is the same over all subjects and the boxplot shows only a single value.

Below, results are reported for the case of shrinkage on Fisher-transformed correlations, followed in brackets by results for the case of shrinkage directly on correlations.

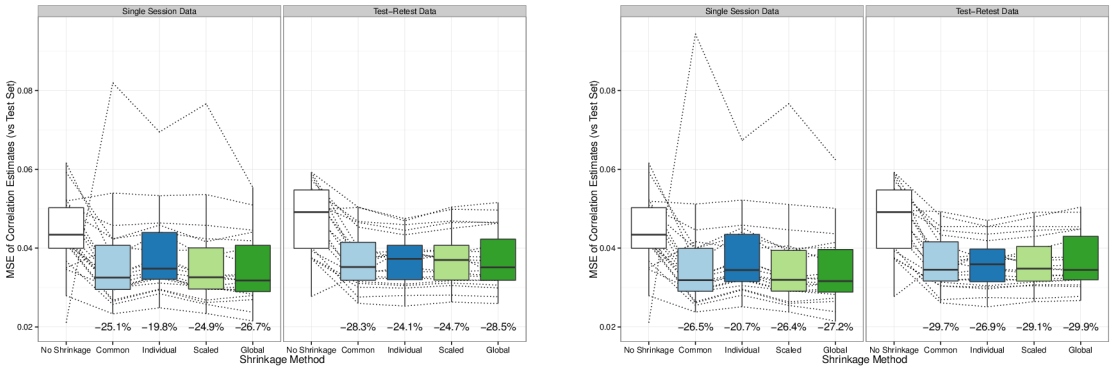
When a single session was used to compute the noise variance, the median degree of shrinkage over all subjects was 69.6% [70.8%] with a common noise variance; 56.5% [58.4%] with individual noise variance; 68.9% [71.1%] with scaled noise variance; and 55.0% [57.3%] with a global noise variance. When test-retest data was used to compute the noise variance, the median degree of shrinkage over all subjects was



(a) Shrinkage on Fisher-transformed correlations

(b) Shrinkage on correlations

Figure 2.7: Degree of shrinkage (percent weighting of the group mean, averaged over all voxel-pairs) by noise variance estimation method and type of dataset (single session or test-retest) used to perform shrinkage. For the "common" and "global" noise variance methods, all subjects have the same shrinkage parameter at each voxel-pair, so the boxplot shows a single value.



(a) Shrinkage on Fisher-transformed correlations

(b) Shrinkage on correlations

Figure 2.8: MSE of raw and shrinkage correlation estimates, by noise variance estimation method and the type of dataset (single session or test-retest) used to perform shrinkage. Each dotted line shows the MSE for a single subject's raw and shrinkage estimates, and the boxplots show the distributions over all subjects. The percent decrease in the median MSE of each shrinkage estimate (compared to the median MSE of the raw estimate) is reported below each boxplot.

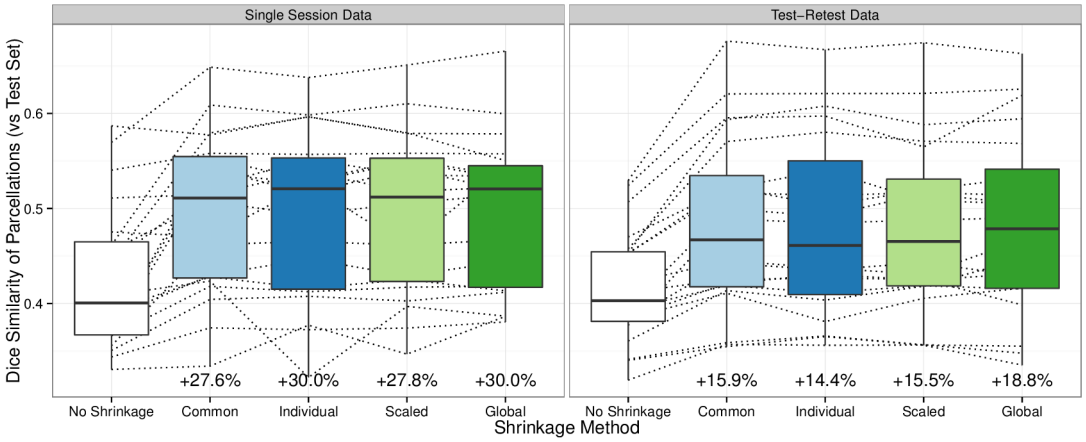


Figure 2.9: Dice coefficients of similarity (with test set parcellations) of parcellations produced using raw and shrinkage correlation estimates. Results are shown by noise variance estimation method and the type of dataset (single session or test-retest) used to perform shrinkage. Each dotted line shows the Dice coefficients for a single subject, and the boxplots show the distributions over subjects. The percent increase in the median Dice coefficient of each shrinkage parcellation (compared to the median Dice coefficient of the raw parcellation) is reported below each boxplot.

51.5% [52.4%] with a common noise variance; 37.3% [38.5%] with individual noise variance; 49.5% [51.8%] with scaled noise variance; and 53.2% [54.5%] with a global noise variance.

As expected, the degree of shrinkage was generally higher when a single session was used due to upward bias in the noise variance estimation. As the global noise variance estimator was designed to avoid this problem it does not suffer from inflated degree of shrinkage. By contrast, when the individual noise variance estimator was used, the degree of shrinkage was significantly lower compared with the common or scaled noise variance estimator. Since each individual noise variance estimate $\hat{\sigma}_i^2(v, v')$ is based on only two observations (rather than $S = 20$), the distribution around the truth $\sigma_i^2(v, v')$ is a highly skewed Chi-squared, which introduces bias into the shrinkage

parameter estimator, since it is a non-linear function of the variance estimators.

Analysis R1: Performance of shrinkage estimates

Figure 2.8 shows the MSE of each raw and shrinkage correlation estimate by the noise variance estimation method employed and the type of dataset (single session or test-retest) used to perform shrinkage. Each dotted line represents a single subject, and the boxplots show the distribution of values over all subjects. Below each boxplot we report the percent decrease in the median MSE of the shrinkage estimates compared to the median MSE of the raw estimates. Figure 2.8a shows the results from applying shrinkage on the Fisher-transformed correlations, and Figure 2.8b shows the results from applying shrinkage directly to the untransformed correlations.

Results are again reported for the case of shrinkage on Fisher-transformed correlations, followed in brackets by results for the case of shrinkage directly on correlations.

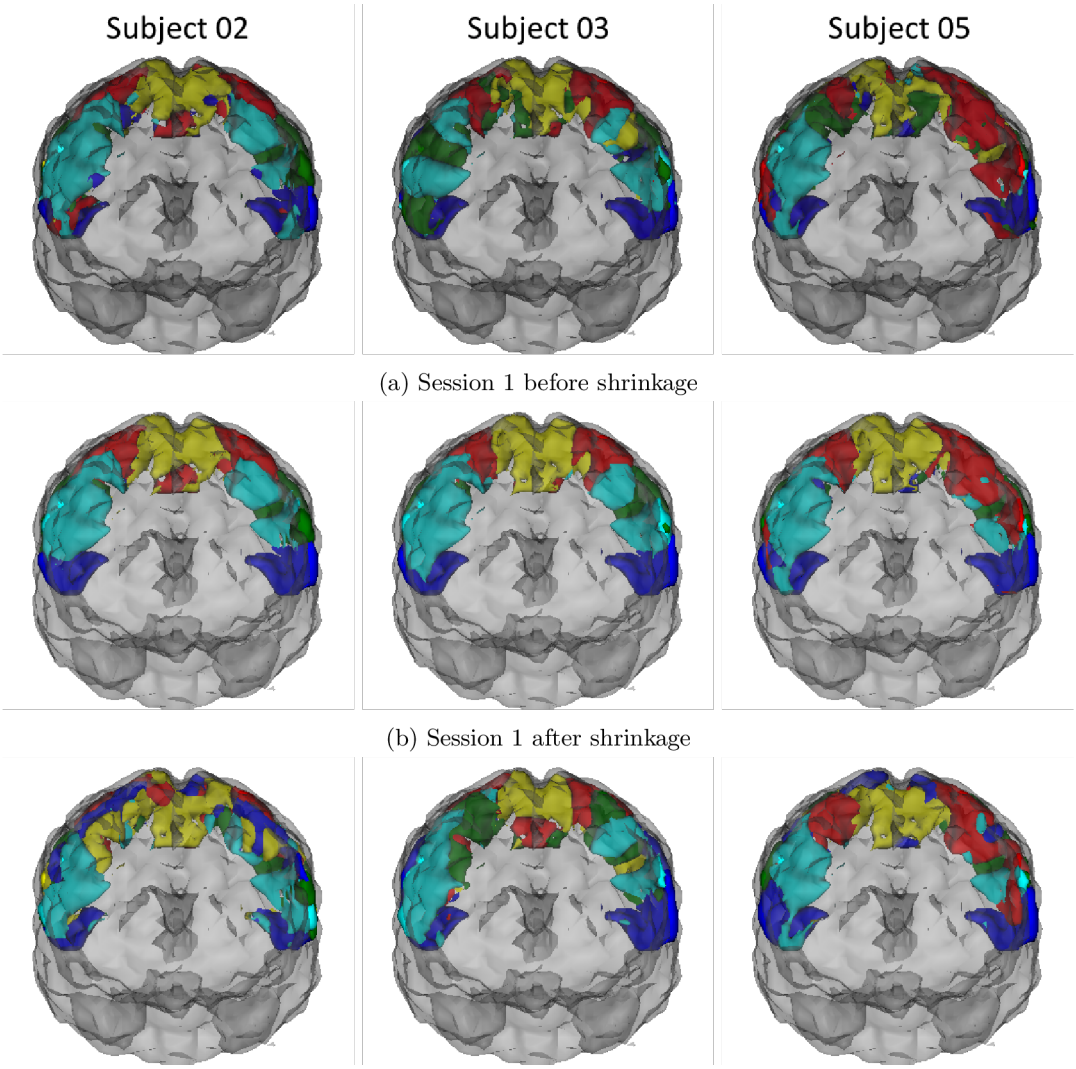
All shrinkage methods resulted in a decrease in the median MSE compared with the raw estimates. Recall that the data used to compute the raw estimates and parcellations was different in the single session case and the test-retest case (see Figure 2.2). Specifically, the full 7-minute scan from the first session was used in the single session case, and only the first 4 minutes and 40 seconds of that scan was used in the test-retest case. Therefore, the raw coefficients and parcellations, and their respective reliability measures, differ across the two cases. When a single session was used to compute the noise variance, the raw correlation estimates had a median MSE of 0.0434. The shrinkage estimates had a median MSE of 0.0325 (25.1% lower)

[0.0319 (26.5% lower)] with a common noise variance; 0.0348 (19.8% lower) [0.0344 (20.7% lower)] with individual noise variance; 0.0326 (24.9% lower) [0.0319 (26.4% lower)] with scaled noise variance; and 0.0318 (26.7% lower) [0.0316 (27.2% lower)] with a global noise variance. When test-retest data was used to compute the noise variance, the raw correlation estimates had a median MSE of 0.0491. The shrinkage estimates had a median MSE of 0.0352 (28.3% lower) [0.0345 (29.7% lower)] with a common noise variance; 0.0373 (24.1% lower) [0.0359 (26.9% lower)] with individual noise variance; 0.0370 (24.7% lower) [0.0348 (29.1% lower)] with scaled noise variance; and 0.0351 (28.5% lower) [0.0344 (29.9% lower)] with a global noise variance.

For all methods, applying shrinkage directly to the correlations resulted in greater reduction in MSE (compared with applying shrinkage to the Fisher-transformed correlations). Whether single session data or test-retest data was used, shrinkage using the global noise variance estimator resulted in the greatest reduction in median MSE. At the subject level, when test-retest data was used to perform shrinkage, shrinkage on untransformed correlations resulted in reduced MSE for 19 out of 20 subjects across all noise variance estimation methods; when single session data was used, shrinkage on untransformed correlations resulted in reduced MSE for 17 out of 20 subjects for the individual noise variance estimation method and 18 out of 20 subjects for all other noise variance estimation methods.

Analysis R2: Performance of parcellations

Figure 2.9 shows the Dice coefficients of similarity (with test set parcellations) of the

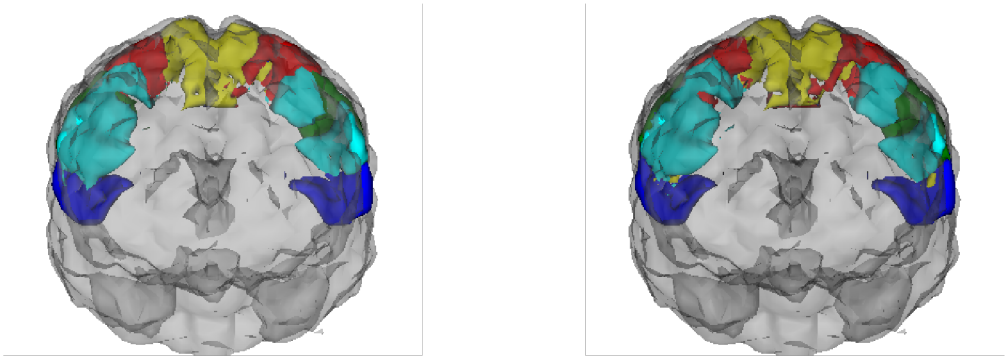


(a) Session 1 before shrinkage

(b) Session 1 after shrinkage

(c) Session 2 no shrinkage

Figure 2.10: Subject-level parcellations of the motor cortex from 3 example subjects.



(a) Session 1 (b) Session 2

Figure 2.11: Group-level parcellations from sessions 1 and 2.

parcellations generated from the raw and shrinkage correlation estimates, by noise variance estimation method and type of dataset (single session or test-retest) used to perform shrinkage. As in Figure 2.8, each line shows the results for a single subject, and the boxplots show the distributions of Dice coefficient values over all subjects. For each noise variance estimation method, the percent increase in the median Dice coefficient of the shrinkage parcellations, compared with the raw parcellations, is reported below each boxplot.

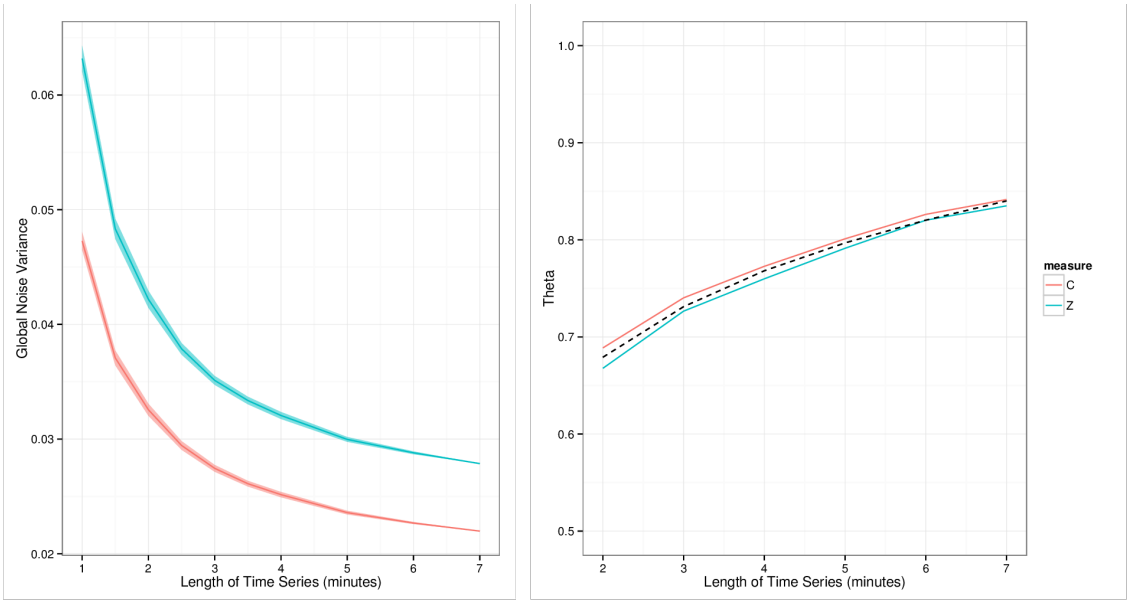
Since shrinkage on the correlation estimates without Fisher-transforming resulted in better performance than shrinkage on Fisher-transformed correlations, parcellations were only generated from the shrinkage estimates obtained by shrinking the correlations directly. The results below therefore reflect the performance of parcellations obtained using this method.

All shrinkage methods resulted in an increase in the median Dice coefficient of parcellations compared with raw parcellations. When a single session was used to

compute the noise variance, the raw parcellations had a median Dice coefficient of 0.401. The shrinkage-based parcellations had a median Dice coefficient of 0.511 (27.6% higher) with a common noise variance; 0.521 (30.0% higher) with individual noise variance; 0.512 (27.8% higher) with scaled noise variance; and 0.521 (30.0% higher) with a global noise variance. When test-retest data was used to compute the noise variance, the raw parcellations had a median Dice coefficient of 0.403. The shrinkage-based parcellations had a median Dice coefficient of 0.467 (15.9% higher) with a common noise variance; 0.461 (14.4% higher) with individual noise variance; 0.465 (15.5% higher) with scaled noise variance; and 0.479 (18.8% higher) with a global noise variance.

The improvement in test-retest reliability of parcellations due to shrinkage was remarkably similar across shrinkage methods. Whether single session data or test-retest data was used to compute the noise variance, the global noise variance estimator again showed the best performance, with an increase in the Dice coefficient of 30.0% using single session data or 18.8% using test-retest data.

Figure 2.10 shows the parcellations of the motor cortex of three subjects (from left to right: subjects 2, 3 and 5) resulting from the first scanning session before shrinkage (Figure 2.10a), the first scanning session after shrinkage (Figure 2.10b), and the second scanning session with no shrinkage (Figure 2.10c). The parcellations in Figure 2.10b were based on shrinkage using a single scanning session and the global noise variance estimator. These parcellations illustrate that subject-level differences



(a) Time series length t vs. noise variance $\sigma^2_U(t)$ (b) Time series length t vs. $\theta(t) = \sigma^2_U(t)/\sigma^2_U(t/2)$

Figure 2.12: The relationship between scan length and noise variance. Results are shown using untranformed correlations (C, shown in red) and Fisher-transformed correlations (Z, shown in teal) to compute the noise variance. Panel (a) shows diminishing session-to-session variance as scan length increases. Panel (b) shows how the adjustment factor $\theta(t) = \sigma^2_U(t)/\sigma^2_U(t/2)$ changes as t increases. The fitted line relating $\theta(t)$ to $\log(t)$ is shown in black.

in parcellations generated from raw correlation estimates are not always seen in subsequent scanning sessions. They also illustrate that while shrinkage-based parcellations are, by nature, more similar to the group-level parcellation, shown in Figure 2.11, subject-level differences can still be seen.

Time series length and noise variance

Figure 2.12 shows the estimated noise variance $\sigma^2_U(t)$ with 95% confidence interval (a) and estimated adjustment factor $\theta(t) = \sigma^2_U(t)/\sigma^2_U(t/2)$ (b) for scan lengths ranging from $t = 1$ to 7 minutes. Results are shown using untranformed correlations

and Fisher-transformed correlations to compute the noise variance. Figure 2.12b also shows the fitted line from the regression relating $\log(t)$ to $\theta(t)$. The coefficient estimates and standard errors from the regression model

$$\theta(t) = \beta_0 + \beta_1 \times \log(t) + \epsilon,$$

were $\hat{\beta}_0 = 0.590$ (s.e. 0.00732) and $\hat{\beta}_1 = 0.129$ (s.e. 0.00493). The adjusted R-squared of the model was 0.986.

2.4 Discussion

In this work we propose a new approach for improving the test-retest reliability of subject-level resting state parcellations based upon the use of shrinkage-based measures of similarity, or distance, between voxels as input to clustering. On 7-minute resting-state scans from 20 healthy adults, parcellations obtained using shrinkage correlation estimates were shown to have up to 30% improved test-retest reliability compared to those obtained using the raw correlation estimates. Through simulations, similar improvement in reliability were observed for a wide range of sample sizes, time series lengths, signal strengths, and degrees of similarity among subjects.

Shrinkage methods have found wide usage in the statistics literature, allowing noisy subject-level estimators to "borrow strength" from a larger population of subjects. The approach is implicit in penalized likelihood inference, multi-level models,[39]

empirical Bayes estimation,[40–42] and Bayesian inference. Indeed, most shrinkage estimators will correspond to the mode of some Bayesian posterior. In recent work, Shou et al. (2014) applied shrinkage to rsfMRI seed-based connectivity analysis and showed a nearly 30% average improvement in intra-subject reliability of correlation estimates. Our work extends these results by developing shrinkage estimators for the full voxel-by-voxel distance matrix required for clustering; proposing methods for constructing shrinkage estimators when only a single scan is available; exploring the utility of shrinkage estimators when the amount of shrinkage is subject-dependent to account for differences in intra-subject variability; and demonstrating improved test-retest reliability of subject-level parcellations based on shrinkage estimates.

Typically, subject-level parcellations derived from short rsfMRI scans (e.g., 5-10 minutes) tend to be highly unreliable due to their low SNR. Longer scans (e.g., 30-60 minutes) lead to more reliable results, and many subject-level parcellation methods are taking advantage of the increased availability of such data. However, there are still a number of reasons why it may be useful or necessary to use shorter scans to produce subject-level parcellations. First, many such rsfMRI scans have already been collected, from which a wealth of information is potentially available. Second, it may be infeasible to collect longer scans for certain populations, including children, the elderly, or diseased populations. While healthy adult controls are ideal candidates to undergo long resting-state scanning sessions, they are not always of primary interest to researchers. Third, the price of obtaining long scans may be prohibitive. By

borrowing strength from the group mean to enhance the quality of noisy subject-level functional connectivity estimates, our shrinkage methods are able to minimize the limitations of short rsfMRI scans and take advantage of the widespread availability of such scans for subject-level inference.

Furthermore, our simulations suggest that shrinkage-based parcellations derived from short scans (200 time points) are equivalent, in terms of reliability, to raw parcellations derived from much longer scans (over 1000 time points). This suggests that performing shrinkage may be comparable to collecting scans of considerably longer length, in terms of reliability. This is an important finding, due to the high costs involved with performing longer scans. Nonetheless, it may still be beneficial to collect longer rsfMRI scans when feasible. More data is almost always better, and subject-level differences will likely be more accurately expressed as the quantity of subject-level data increases. On the other hand, longer scans should not be viewed as mutually exclusive with shrinkage methods. Our simulation results suggest that shrinkage on longer scans can still lead to substantially more reliable subject-level parcellations. Further research on the benefits of shrinkage on longer scans using real rsfMRI data will be important to understand this interplay.

Finally, according to our simulations, the shrinkage methods we have proposed not only improve the reliability of the overall parcellations, but also the reliability within regions where subject-level differences exist. The ability to accurately parcellate these regions is vital to quantifying and studying subject-level differences in functional

organization.

Although we chose to demonstrate the applicability of our shrinkage methods using correlations as our similarity metric, it is important to note that they are applicable to almost any similarity or distance metric. As long as the assumptions of Normality and signal-noise independence are roughly satisfied, our methods can be applied to a wide variety of metrics. Interestingly, we observed more improvement in reliability by applying shrinkage directly to correlation estimates, rather than to Fisher-transformed correlations, even though the model assumptions are not strictly satisfied for correlation estimates. This is no doubt due to the fact that we are evaluating the error on the correlation scale. If we were instead to evaluate the error on the Fisher-transformed scale, it would be better to perform shrinkage directly on the Fisher-transformed data. We have also applied shrinkage to the inter-voxel similarity metric described in [34], which incorporates long-range correlations between voxels within the motor cortex and the rest of the brain, and found that shrinkage reduced the MSE of the estimates by approximately 40% (results not shown). Furthermore, though we used spectral clustering for parcellation, there is nothing that prevents other clustering methods from being used instead. For example, we have observed similar improvement in parcellation using K-means clustering (results not shown). The objective of the methods we have described is to maximize the reliability of the similarity or distance metric utilized in clustering. Therefore, any clustering method that depends on such a metric will benefit from using shrinkage estimates of that

metric in place of raw observed values.

Throughout the paper we have described a number of potential shrinkage methods. Based on our empirical findings we recommend utilizing the global noise variance estimator, since it is practical for settings where test-retest data is not available (or is only available for a subset of subjects), and it demonstrated the best performance in terms of improved reliability of correlation estimates and parcellations. This method has been implemented in both MATLAB and R and is available for download through Mathworks File Exchange[1] and GitHub[2], respectively. The strong performance of the global noise variance estimator does not necessarily imply that there is no spatial variability in noise levels, but rather that there may be too many parameters to estimate in a meaningful way. For example, in the motor cortex alone there are more than 27 million unique voxel-pairs, and estimating a shrinkage parameter for each can be prohibitive. This may also explain why the scaled noise variance method tends to out-perform the individual noise variance method. Though both noise variance estimators seek to quantify each subject's personal noise level, the scaled noise variance estimator requires the estimation of dramatically fewer parameters. For example, with 7396 voxels and 20 subjects, rather than estimating over $20\times27{,}000{,}000$ parameters, the scaled noise variance approach requires approximately 27,000,000+20 parameters. As more data is collected, more parameters can be reliably estimated. Therefore, with longer scans, the common or scaled noise variance estimators may

[1] http://www.mathworks.com/matlabcentral/fileexchange/48453-mandymejia-shrinkit
[2] https://github.com/mandymejia/shrinkR

ultimately begin to out-perform the global noise variance estimator. However, there is another issue to keep in mind when performing subject-specific shrinkage using the scaled or individual noise variance estimator. When all subjects share the same shrinkage parameter, the rank of the subjects' values relative to one other will not change as each subject's value changes. By contrast, if shrinkage is subject-specific, the rank of subjects may not be preserved. Therefore, even if subject-level parameters can be reliably estimated, care should be taken when performing subject-specific shrinkage.

Although we have strong evidence of the benefits of the proposed shrinkage methods, our analysis was limited to the motor cortex in a population of healthy adults. The benefits of shrinkage on reproducibility are likely to vary, depending on the inter-voxel similarity metric being estimated, the ROI being parcellated, and the population being studied, among other factors. Future research should focus on quantifying the benefits of shrinkage within other regions of interest, for whole-brain parcellation, and for other, potentially more diverse, populations. Finally, while the methods we have proposed can be easily applied to any distance or similarity metric, some parcellation methods employ other versions of subject-level data, such as the entire time series or principal components. Future research should focus on adapting the ideas presented in this paper to such settings.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1232825. This research was supported in part by NIH grants R01 EB016061 and P41 EB015909 from the National Institute of Biomedical Imaging and Bioengineering.

Appendix

Claim. The expected value of the common noise variance estimator and the expected value of the mean individual noise variance estimator are the same at each voxel-pair (v, v').

$$E\left[\hat{\sigma}_U^{2(C)}(v, v')\right] = E\left[\frac{1}{I}\sum_{i=1}^{I}\hat{\sigma}_{U,i}^{2(I)}(v, v')\right]$$

Starting from the LHS and dropping the (v, v') notation for conciseness, we see that

$$\begin{aligned}
E\left[\hat{\sigma}_U^{2(C)}\right] &= E\left[\frac{1}{2(I-1)}\sum_{i=1}^{I}\left(D_i - \bar{D}\right)^2\right] \\
&= \frac{1}{2(I-1)}E\left[\sum_{i=1}^{I}\left(D_i^2 - 2D_i\bar{D} + \bar{D}^2\right)\right] \\
&= \frac{1}{2(I-1)}E\left[\sum_{i=1}^{I}D_i^2 - 2I\bar{D}^2 + I\bar{D}^2\right] \\
&= \frac{1}{2(I-1)}E\left[\sum_{i=1}^{I}D_i^2 - I\bar{D}^2\right] \\
&= \frac{1}{2(I-1)}\sum_{i=1}^{I}E\left[D_i^2\right] - \frac{I}{2(I-1)}E\left[\bar{D}^2\right] \\
&= \frac{1}{2(I-1)}\sum_{i=1}^{I}E\left[2\hat{\sigma}_{U,i}^{2(I)}\right] - \frac{I}{2(I-1)}Var(\bar{D}) \\
&= \frac{1}{I-1}\sum_{i=1}^{I}\sigma_{U,i}^2 - \frac{I}{2(I-1)}\frac{1}{I^2}\sum_{i=1}^{I}Var\left(D_i\right) \\
&= \frac{1}{I-1}\sum_{i=1}^{I}\sigma_{U,i}^2 - \frac{1}{2I(I-1)}\sum_{i=1}^{I}Var\left(U_{i2} - U_{i1}\right) \\
&= \frac{1}{I-1}\sum_{i=1}^{I}\sigma_{U,i}^2 - \frac{1}{2I(I-1)}\sum_{i=1}^{I}2Var\left(U_{ij}\right) \\
&= \frac{1}{I-1}\sum_{i=1}^{I}\sigma_{U,i}^2 - \frac{1}{I(I-1)}\sum_{i=1}^{I}\sigma_{U,i}^2 \\
&= \frac{1}{I}\sum_{i=1}^{I}\sigma_{U,i}^2
\end{aligned}$$

Equality with the RHS clearly follows, since

$$E\left[\frac{1}{I}\sum_{i=1}^{I}\hat{\sigma}_{U,i}^{2(I)}\right] = \frac{1}{I}\sum_{i=1}^{I}E\left[\hat{\sigma}_{U,i}^{2(I)}\right] = \frac{1}{I}\sum_{i=1}^{I}\sigma_{U,i}^2.$$

Chapter 3

PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data

Outlier detection for high-dimensional data is a popular topic in modern statistical research. However, one source of high-dimensional data that has received relatively little attention is functional magnetic resonance images (fMRI), which consists of hundreds of thousands of measurements sampled at hundreds of time points. At a time when the availability of fMRI data is rapidly growing—primarily through large, publicly available grassroots datasets consisting of resting-state fMRI data—automated quality control and outlier detection methods are greatly needed. We propose PCA leverage and demonstrate how it can be used to identify outlying time

points in an fMRI scan. Furthermore, PCA leverage is a measure of the influence of each observation on the estimation of principal components, which forms the basis of independent component analysis (ICA) and seed connectivity, two of the most widely used methods for analyzing resting-state fMRI data. We also propose an alternative measure, PCA robust distance, which is less sensitive to outliers and has controllable statistical properties. The proposed methods are validated through simulation studies and are shown to be highly accurate. We also conduct a reliability study using resting-state fMRI data from the Autism Brain Imaging Data Exchange (ABIDE) and find that removal of outliers using the proposed methods results in more reliable estimation of subject-level resting-state networks using ICA.

3.1 Introduction

The presence of outliers in high-dimensional settings, such as genetics, medical imaging and chemometrics, is a common problem in modern statistics and has been the focus of much recent research.[51–56] One source of especially high-dimensional data that could benefit from improved outlier detection methods is functional magnetic resonance imaging (fMRI). A single fMRI scan usually contains 100,000-200,000 volumetric elements or voxels within the brain, which are sampled at hundreds of time points. Here, we consider voxels to be variables and time points to be observations, in which case the outlier problem is to identify time points that contain high levels of

systematic non-neuronal variation or artifacts. A second type of outlier problem—but one that is not the focus of this paper—is to identify outlying fMRI scans in a group analysis.

The acquired fMRI data contains multiple sources of noise related to both the hardware and the participants themselves.[3] These include gradient and magnetic field instability, head movement, and physiological effects, such as heartbeat and respiration. These sources of noise appear in the data as high-frequency spikes, image artifacts and distortions, as well as low-frequency (slow) drift. All fMRI data contain some artifacts, and it is critical to perform adequate quality control on the data prior to statistical analysis.

In recent years, a number of grassroots initiatives have resulted in large, publicly available fMRI databases, and this increased accessibility has emphasized the need for automated methods to perform accurate and reliable quality control on these data. Most of this data consists of resting-state fMRI, collected while the subject is not performing an explicit task. The promise of these datasets is great; many focus on a particular disease, including the Alzheimer's Disease Neuroimaging Initiative (ADNI),[4] the Autism Brain Imaging Data Exchange (ABIDE),[57] and the Attention Deficit Hyperactivity Disorder (ADHD) 200,[6] providing an enormous opportunity for researchers to better understand the neurological reasons behind such diseases. However, these datasets are often a collection of scans from multiple sites with varying acquisition protocols, preprocessing pipelines and quality control mechanisms, result-

ing in widely varying types and rates of outliers. Quality inspection often takes place in a manual or semi-automated manner by the individual research groups that use these datasets, but this is typically only a cursory process due to the large volume of data. This presents a timely opportunity for statisticians to adapt established methods of outlier detection to the unique and very high-dimensional setting of fMRI data.

Here we propose a fully automated method to detect outliers in fMRI data, which is based on dimension reduction through principal components analysis (PCA) and established measures of outlyingness, namely leverage and robust distances. While leverage has not typically been employed for outlier identification outside of the regression framework, here we argue for leverage as a meaningful measure when the principal components are themselves of interest, which is the case when PCA is used for dimension reduction prior to independent component analysis (ICA). We note that leverage has been previously combined with PCA in the context of principal components regression.[58]

A number of outlier detection methods for standard and high-dimensional data employ PCA. For example, PCA influence functions and other measures of the sensitivity of principal components to each observation[59,60] have been proposed as a way to identify outliers. However, such methods often rely on re-estimating the principal components with each observation left out, and can therefore be computationally demanding. Other outlier detection methods depend on robust covariance estimation

(see[53] for a review), but these methods are usually not well-suited for high-dimensional settings. One such method that has received significant attention is minimum covariance determinant (MCD) estimators, which identify a subset of observations that result in an estimated covariance matrix with the minimum determinant.[61]

Several outlier detection methods employing PCA have been proposed specifically for high-dimensional data.[51] proposed ROBPCA, a robust method of PCA for high-dimensional data that can also be used to identify outliers, which lie far from the robust principal components space.[52] proposed PCOut and Sign, two computationally efficient methods that perform standard PCA after robustly scaling the data and looking for outliers within the principal directions that explain 99% of the variance in the data.[56] proposed the minimum diagonal product estimator, which is related to the MCD but ignores off-diagonal elements and, unlike the MCD, is identifiable when there are more variables than observations.[55] proposed an adaptation of the MCD to high-dimensional data through regularization and applied the methods to summary statistics from a neuroimaging dataset.

However, many of the proposed methods for outlier detection in high-dimensional data have only been validated using moderately sized data, which often contain more observations than variables. One exception comes from the field of genetics, where[54] proposed a method for identifying outlying genes in microarray data by performing dimension reduction through PCA followed by computation of robust distances on the reduced data. The proposed method was validated using microarray datasets

with approximately 100 observations and 2,000 variables, and was shown to be highly accurate and improve class prediction accuracy. ROBPCA was also performed for comparison and resulted in more false positives and false negatives than the proposed method.

To identify outlying time points in fMRI data, several data-driven approaches have been proposed, and the removal of these time points (scrubbing or spike regression) has been shown to be beneficial.[62–64] However, the methods proposed thus far tend to focus on subject head motion and variability of intensities within and across time points. The focus on motion is well-founded, as it has been thoroughly demonstrated that motion results in spurious changes to fMRI measures of brain connectivity.[62,64–66] However, a more unified framework for outlier identification is needed, as motion is only one potential source of artifacts in fMRI data. Furthermore, many of the approaches proposed thus far result in a collection of related quality control measures that must be combined in some way to identify and remove outliers. Here, we propose using a single, universal measure of outlyingness that is directly related to the influence of each time point on the estimation of principal components, which are the foundation of several common measures of brain connectivity (as described in more detail in Section 3.2.2).

The remainder of this paper is organized as follows. We begin with a description of our statistical methodology. We then present a simulation study, which is used to assess the sensitivity and specificity of the proposed methods. Next, we present a

reliability analysis employing the ABIDE dataset. We use independent components analysis (ICA) to identify subject-level brain networks and evaluate the reliability of these networks with and without outlier removal using the proposed methods. We conclude with a brief discussion.

3.2 Methods

We propose two PCA-based measures of outlyingness: *PCA leverage* and *PCA robust distance*. We develop thresholding rules to label outliers using either measure. For both measures, we begin by reducing dimensionality through PCA. These measures and their corresponding thresholding rules are described in detail below. All computations are performed in the R statistical environment version 3.1.1.[67]

3.2.1 Dimension Reduction

We use the following notation: for an fMRI dataset, let T be the number of 3-dimensional volumes in a 4-dimensional scan and let V be the number of voxels in the brain, where $T \ll V$.

Let $\mathbf{Y}_{T\times V}$ represent the data from a single fMRI scan, where each row of $\mathbf{Y}$ is a vectorized volume after removing voxels outside of the brain. We begin by centering and scaling each column relative to its median and median absolute deviation,[68] respectively, to avoid the influence of outliers. The singular value decomposition

(SVD)[69] of $\mathbf{Y}$ is given by $\mathbf{Y} = \mathbf{UDV}^t$, where $\mathbf{U}$ is a $T \times T$ matrix, $\mathbf{D}$ is a diagonal $T \times T$ matrix, $\mathbf{V}^t$ is a $T \times V$ matrix, and $\mathbf{UU}^t = \mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}_T$. Here $\mathbf{A}^t$ denotes the transpose of matrix $\mathbf{A}$. The diagonal elements of $\mathbf{D}$ are $d_1 \geq d_2 \geq \cdots \geq d_T \geq 0$. The rows of $\mathbf{V}^t$ contain the principal components or *eigenimages* of $\mathbf{Y}$, and the columns of $\tilde{\mathbf{U}} = \mathbf{UD}$ contain the corresponding principal component scores. Note that to avoid memory limitations, rather than compute the SVD of $\mathbf{Y}$ directly, one generally computes the singular value decomposition (SVD) of $\mathbf{YY}^t$ to obtain $\mathbf{YY}^t = \mathbf{UD}^2\mathbf{U}^t$ and then solves for $\mathbf{V}^t$.

We retain $Q < T$ principal components, so that the "reduced data" are given by the submatrices of $\mathbf{U}$ and $\mathbf{D}$ corresponding to the first Q principal components. For ease of notation we redefine $\mathbf{U}_{T\times Q}$ and $\mathbf{D}_{Q\times Q}$ to represent these submatrices and $\tilde{\mathbf{U}}_{T\times Q} = \mathbf{UD}$. To choose the model order Q, we retain only those components with a greater-than-average eigenvalue, which explain a greater-than-average percent of the total variation. This is one of the simplest cutoff rules for principal components;[70–72] while more sophisticated cutoff methods exist we find that this works well in practice for outlier detection. To avoid extreme solutions, we limit the range of possible model orders to between 15 and 50.

3.2.2 Principal components leverage

3.2.2.1 Motivation and definition

In regression, leverage is defined as the diagonals of the "hat matrix" $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, where $\mathbf{X}$ is a matrix of explanatory variables.[73] The hat matrix projects the outcome variable or variables $\mathbf{Y}$ onto the column space of $\mathbf{X}$, yielding the projected data $\hat{\mathbf{Y}} = \mathbf{HY}$. Leverage, which is bounded between 0 and 1, is often used to assess the potential of an observation to influence the regression fit, since it is the proportion of a change in y_i that will be reflected in the fitted value $\hat{y}_i$. Similarly, it is proportional to the uncertainty in the estimate of $\hat{y}_i$, since $\text{Var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$. Of particular relevance for our context, leverage is also a measure of outlyingness among the explanatory variables.

Extending the idea of leverage to the PCA context, we treat $\tilde{\mathbf{U}} = \mathbf{UD}$ as a design matrix in the estimation of $\mathbf{V}^t$. Recall that for computational convenience we often first obtain $\mathbf{U}$ and $\mathbf{D}$ through the SVD of $\mathbf{YY}^t$. Using the reduced versions of $\mathbf{U}$ and $\mathbf{D}$, we then compute $\mathbf{V}^t = \mathbf{D}^{-1}\mathbf{U}^t\mathbf{Y}$, which is equivalent to the ordinary least squares (OLS) estimate $\hat{\mathbf{V}}^t$ in the multivariate regression model $\mathbf{Y} = \tilde{\mathbf{U}}\mathbf{V}^t + \mathbf{E}$, since $\hat{\mathbf{V}}^t = \mathbf{D}^{-1}\mathbf{U}^t\mathbf{Y}$. Note that this equality holds whether the dimension is reduced before or after $\mathbf{V}^t$ is estimated.

We therefore define PCA leverage as $\mathbf{h} = \{h_1, \ldots, h_T\} = \text{diag}\{\mathbf{H}\}$, where $\mathbf{H} = \tilde{\mathbf{U}}(\tilde{\mathbf{U}}^t\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^t = \mathbf{UU}^t$. Note that $\mathbf{D}$ is simply a scaling factor applied to each variable

and therefore has no effect on leverage, which is a standardized measure. Continuing the regression analogy, in PCA the matrix $\mathbf{H}$ projects $\mathbf{Y}$ onto the column space of $\tilde{\mathbf{U}}$ — the principal directions or reduced space of the data — as $\hat{\mathbf{Y}} = \mathbf{UDV}^t = \mathbf{UDD}^{-1}\mathbf{U}^t\mathbf{Y} = \mathbf{UU}^t\mathbf{Y} = \mathbf{HY}$. Conceptually, PCA leverage represents the potential of an observation to influence the estimation of the principal components in $\mathbf{V}^t$. Furthermore, PCA leverage is a measure of outlyingness not only among the PCA scores but also within the reduced data $\hat{\mathbf{Y}} = \mathbf{UDV}^t$, since $\hat{\mathbf{Y}}(\hat{\mathbf{Y}}^t\hat{\mathbf{Y}})^{-1}\hat{\mathbf{Y}}^t = \mathbf{UU}^t = \mathbf{H}$. Note that dimension reduction is essential for PCA leverage to be informative, since $\mathbf{UU}^t = \mathbf{I}$ when all T components are retained.

Before specifying thresholding rules for PCA leverage to identify outliers, we return briefly to the interpretation of PCA leverage and describe why it is a meaningful measure of outlyingness for fMRI data and especially for resting-state fMRI data. First, while in the regression context leverage only represents the *potential* influence of an observation on estimation of regression coefficients, PCA leverage is a more direct measure of influence: in regression, an influence point is defined as a point that is an outlier in the explanatory variables (a "leverage point") and in the response variables; by contrast, PCA leverage points are outliers in both $\tilde{\mathbf{U}}$ and the original data $\mathbf{Y}$, implying that we can consider leverage points to be influence points in the PCA context. Furthermore, while in regression we distinguish between "good" and "bad" leverage points, in fMRI data observations with high PCA leverage points are unlikely to represent true signal, since the signal change associated with neuronal

sources is very small compared with noise and artifacts. Therefore, we assume that all observations with high PCA leverage are "bad" influence points in the fMRI context.

Additionally, the interpretation of PCA leverage as the influence of each observation on the estimation of principal components is particularly meaningful for resting-state fMRI data, as the matrix $\mathbf{V}^t$ is the precursor of two of the most common types of analysis applied to such data, namely the computation of pairwise correlation between regions of interest and temporally coherent brain networks derived from ICA: the $V \times V$ pairwise correlation matrix is equal to $\mathbf{V}\mathbf{D}^2\mathbf{V}^t$; subject- and group-level ICA both begin with a PCA reduction of the subject-level data along the temporal dimension, giving $\mathbf{V}^t$. Therefore, the estimation of $\mathbf{V}^t$ is often of principal interest in the analysis of resting-state fMRI data, in which case PCA leverage is a meaningful measure of influence as well as a measure of outlyingness.

3.2.2.2 Thresholding rule

In setting a thresholding rule to identify outliers, it is important to recognize that leverage of an observation only has relative meaning, since the leverage of a set of observations sums to the number of variables. With only so much leverage to "go around", the leverage of normal observations will necessarily decrease in the presence of one or more high-leverage observations. This implies that the mean, which is fixed, may not be representative of normal observations. We therefore use the median leverage across all observations, $m_h = \text{med}(h_1, \ldots, h_T)$, as a reference for normal

values. Let $Z_t = I(h_t > \alpha \cdot m_h)$ indicate whether observation t exceeds α-times m_h, in which case it is labeled a "leverage outlier". We find that $\alpha = 3$ works well, but this may be varied to establish a more or less conservative threshold.

Leverage does not display convenient statistical properties, so rules of thumb such as the one we propose, rather than formal statistical tests, are usually employed to detect leverage points. While this may work reasonably well in practice, a formal statistical test for outliers with known and controllable properties is desirable. Furthermore, leverage is known to be subject to masking of outliers,[74,75] a phenomenon in which truly outlying observations have leverage values within the normal range due to the presence and influence of more extreme outliers. To address both of these limitations, in the following section we propose an alternative robust distance measure based on minimum covariance determinant (MCD) estimators.[61]

3.2.3 Principal components robust distance

For a design matrix with an intercept or centered variables, leverage is related to the squared empirical Mahalanobis distance,[76] which is defined for an $n \times p$ matrix $\mathbf{X}$ and observation i as $d_i^2 = (X_i - \bar{X})^t S^{-1} (X_i - \bar{X})\}$, where $\bar{X}$ and S are the empirical mean and covariance matrix of $\mathbf{X}$, respectively. The Mahalanobis distance is known to be sensitive to outliers, as they may have a strong influence on the empirical mean and covariance estimates; hence, leverage is often believed to be sensitive to outliers.[74] However, at first glance PCA leverage may appear to be less sensitive to outliers than

Mahalanobis distance, since the columns of $\mathbf{U}$ are not centered and the middle term cancels, so that the PCA leverage of each observation does not depend directly on any other observation.

However, outliers may in fact have a more subtle and indirect effect on the leverage of a set of observations, due its properties of being bounded between 0 and 1 and summing to a fixed value. For example, as described above, the presence of a single observation with high leverage will reduce the leverage of all other points. Furthermore, due to the fact that all leverage values must be less than 1, it may be difficult to distinguish amongst the relative importance of multiple outliers that each have leverage close to 1. Clearly, leverage may be subject to the influence of outliers.

As an alternate measure, we adopt the minimum covariance determinant (MCD) distance proposed by.[61] For a general dataset, let n be the number of observations and p be the number of variables. The MCD estimators of location, $\bar{X}^*$, and scale, S^*, are obtained by selecting a subset of the data of size $h < n$ for which the confidence ellipsoid determined by S^* and centered at $\bar{X}^*$ has minimal volume. Within this subset of the data, the traditional location and scale estimators are computed. The maximum breakdown point of MCD estimators is obtained by setting $h = \lfloor (n + p + 1)/2 \rfloor$ and approaches 50% as $n \to \infty$. MCD distance $d^2_{S^*}(X_i, \bar{X}^*)$ is computed as a Mahalanobis distance using the MCD estimates of location and shape in place of traditional estimates. For ease of notation, let d_i^2 denote the MCD distance of observation i.

Let $\mathcal{N} = 1, \ldots, n$, and let $\mathcal{N}^*$, $|\mathcal{N}^*| = h$, be the indices of the observations selected to compute the MCD estimates of location and scale. Let $\mathcal{N}^- = \mathcal{N} \backslash \mathcal{N}^*$ be the indices of the remaining observations. We look for outliers within these observations, as they represent nearly half of the data and are further from the center of the distribution than those in $\mathcal{N}^*$. For Gaussian data, the MCD distances of the observations in $\mathcal{N}^*$ approximately follow a Chi-squared distribution with p degrees of freedom,[51,54] while the MCD distances of the remaining observations follow an F distribution.[77] While previous work has simply assumed a Chi-squared distribution for all observations,[54] for fMRI data we found that this resulted in a large number of false positives. For $i \in \mathcal{N}^-$,

$$\tilde{d}_i^2 := \frac{c(m-p+1)}{pm} d_i^2 \sim F_{p,m-p+1}, \tag{3.1}$$

where c and m can be estimated asymptotically or through simulation. To estimate c we use the asymptotic form, $\hat{c} = Pr\left\{\chi^2_{p+2} < \chi^2_{p,h/n}\right\} / (h/n)$, which is known to perform well in small samples. To estimate m we use the small sample-corrected asymptotic form given in.[77] To improve the F-distribution fit, like[78] and[52] we find it helpful to transform the distances to match the median of the theoretical distribution:

$$\tilde{\tilde{d}}_i^2 := \tilde{d}_i^2 \frac{F_{p,m-p+1,0.5}}{\text{median}\{\tilde{d}_i^2 : i \in \mathcal{N}^-\}}. \tag{3.2}$$

We label a "distance outlier" any observation in $\mathcal{N}^-$ with $\tilde{\tilde{d}}_i^2$ greater than the $(1-\gamma)$th quantile of the theoretical distribution. We set $1 - \gamma = 0.999$ and find that this

performs well in both simulated and experimental data.

3.3 Simulation Study

3.3.1 Construction of baseline scans

Our simulated dataset is based on three different fMRI scans from the ABIDE dataset (described in Section 3.4). Each scan was acquired from a different subject at a different site in order to maximize sample heterogeneity. For each scan, we use a contiguous subset of volumes containing no visible artifacts or detected outliers, resulting in 141, 171 and 89 volumes in each scan, respectively. We reduce each 3D volume to 2 dimensions by using only the 45th axial (horizontal) slice, which corresponds roughly to the center of the brain.

To construct the simulated baseline (noise- and artifact-free) scans, we take the following steps. We first center each scan "across time" by subtracting the mean value of each voxel, which separates primarily anatomical information from functional information. For scan i, let T_i be the length of the scan and V_i be the number of voxels in the brain mask, so that scan i is represented by the $T_i \times V_i$ matrix $\mathbf{Y}_i$. After centering across time, we have $\mathbf{Y}_i = \mathbf{Z}_i + \mathbf{B}_i$, where $\mathbf{Z}_i$ represents primarily functional information and $\mathbf{B}_i$ represents primarily anatomical information (each row of $\mathbf{B}_i$ is identical).

As described in detail in Section 3.4, for each site we perform group ICA (GICA)

and identify the independent components corresponding to known resting-state brain networks. Let Q_i be the number of brain networks identified for the site of scan i. We then perform the first step of dual regression[79] to obtain scan-level temporal components for each brain network. This results in the decomposition $\mathbf{Z}_i = \mathbf{A}_i\mathbf{S}_i + \mathbf{E}_i =: \mathbf{X}_i + \mathbf{E}_i$, where $\mathbf{S}_i$ is $Q_i \times V_i$ and contains the vectorized resting-state networks identified through GICA for the site of scan i, and $\mathbf{A}_i$ is $T_i \times Q_i$ and contains the corresponding temporal components for scan i. $\mathbf{E}_i$ is the residual, which consists of structured (spatially and temporally correlated) noise.

3.3.2 Artifact-free images

For each scan i, we construct three simulation setups: baseline image ($\mathbf{B}_i$) plus white noise (setup 1); basline image plus functional signal ($\mathbf{B}_i + \mathbf{X}_i$) plus white noise (setup 2); and baseline image plus functional signal plus structured noise (setup 3).

To test the specificity of each outlier detection method in the artifact-free setting, we generate images with varying signal-to-noise ratio (SNR) in the following way. In general, we estimate the variance across time and take the average across voxels. For scan i, let the signal variance be $\hat{\sigma}^2_{i,X} = \frac{1}{V_i}\sum_{v=1}^{V_i} \hat{Var}\{X_i(v)\}$ and the noise variance be $\hat{\sigma}^2_{i,E} = \frac{1}{V_i}\sum_{v=1}^{V_i} \hat{Var}\{E_i(v)\}$. Define SNR as the ratio of signal variance to noise variance, and let λ be the desired SNR of the simulated scans. We simulate scans with $\lambda \in \{0.025, 0.050, 0.075, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$. For setups 1 and 2, we generate the white noise matrix $\mathbf{W}_i(\lambda)$ for scan i as independent, mean-zero Gaussian noise with

variance $\sigma^2_{i,E}(\lambda) = \hat{\sigma}^2_{i,X}/\lambda$. For setup 3, we generate the structured noise matrix $\mathbf{E}_i(\lambda) = \sqrt{SNR_i/\lambda} \times \mathbf{E}_i$, where $SNR_i = \hat{\sigma}^2_{i,X}/\hat{\sigma}^2_{i,E}$ is the baseline SNR of scan i.

Therefore, the simulated artifact-free data for each setup at SNR λ is $\mathbf{B}_i + \mathbf{W}_i(\lambda)$ for setup 1; $\mathbf{B}_i + \mathbf{X}_i + \mathbf{W}_i(\lambda)$ for setup 2; and $\mathbf{B}_i + \mathbf{X}_i + \mathbf{E}_i(\lambda)$ for setup 3. For setups 1 and 2, we randomly generate $\mathbf{W}_i(\lambda)$ 1000 times; for setup 3 the noise is fixed.

We are interested in quantifying the specificity, or the percentage of observations *not* labeled as leverage or distance outliers that are truly non-outliers. Since there are no artifacts in the simulated data, specificity in this case is simply the percentage of volumes in each scan not labeled as outliers. For each simulation setup, specificity is nearly 100% on average using either method, with very narrow confidence bands.

3.3.3 Images with artifacts

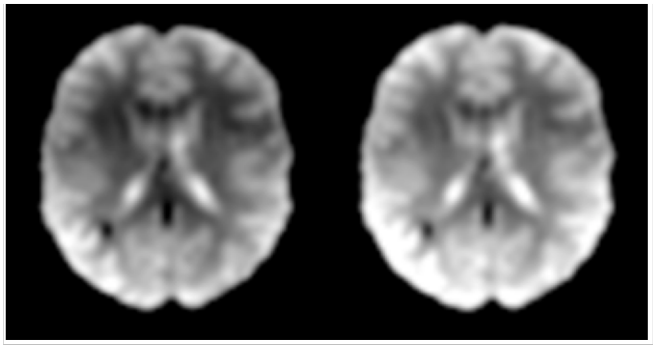
For each scan i, we again construct the three simulation setups described above, but with SNR fixed at SNR_i. We generate four comon types of artifacts found in fMRI data: spike artifacts, motion artifacts, banding artifacts, and ghosting artifacts. Spike artifacts are created by increasing the intensity of an entire volume by a given percentage. Motion artifacts are created by rotating a volume by a given angle. Banding artifacts are generated by changing the intensity in a particular location in "k-space", the Fourier transform of the image, resulting in a striped appearance in the image. Ghosting artifacts are created by superimposing an image of a "ghost" of a given intensity moving through space over time.

Artifact Type	Parameter	Range
Spike	percent intensity increase	$1\% - 10\%$
Motion	rotation angle	$1^\circ - 5^\circ$
Banding	change in k-space value at location (15,15)	$50 - 200$ times
Ghosting	mean ghost intensity/mean scan intensity	$0.06 - 0.32$

Table 3.1: Varying parameter and range of values for each artifact. At each iteration, the artifact intensity is generated from a uniform distribution with the specified range.

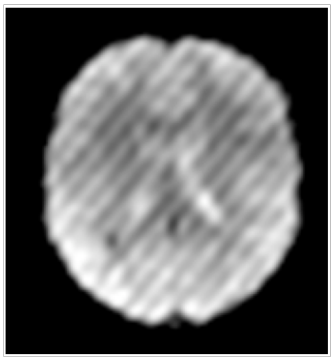
At each of 1000 iterations, one simulated fMRI scan is generated for each subject, artifact type and simulation setup. For spike, motion and banding artifacts, 10 volumes are randomly selected, and the artifact intensity for each volume is randomly generated from a uniform distribution. For ghosting artifacts, 9 sequential volumes are randomly selected, and the ghost intensity is randomly generated from a uniform distribution. The parameters and intensity range for each artifact type is given in Table 3.1. An example of each artifact type is displayed in Figure 3.1.

We are interested in both the specificity (defined in the previous section) and the sensitivity, or the percentage of true outliers that are identified as outliers. Figure 3.2 shows the mean and 95% bootstrap confidence intervals of sensitivity and specificity by outlier detection method, simulation setup, and artifact type. We see that as the simulation setup becomes more realistic, the sensitivity to outliers tends to decrease, while the specificity is relatively stable. The distance method has higher sensitivity and specificity than the leverage method across simulation setups and artifact types. While differences across artifact types are apparent, comparisons are not meaningful, as such differences may be driven by the range of intensities chosen (see Table 3.1).

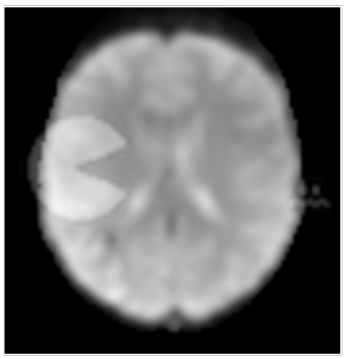


(a) Spike artifact

(b) Rotation artifact



(c) Banding artifact



(d) Ghosting artifact

Figure 3.1: Examples of each artifact type. Figure (a) shows a normal volume on the left and a volume with a spike artifact on the right. Figure (b) shows the image mask before and after rotation. The spike, rotation and ghosting artifacts are generated from the maximum artifact intensity as described in Table 3.1; the banding artifact is generated randomly as described in Table 3.1.

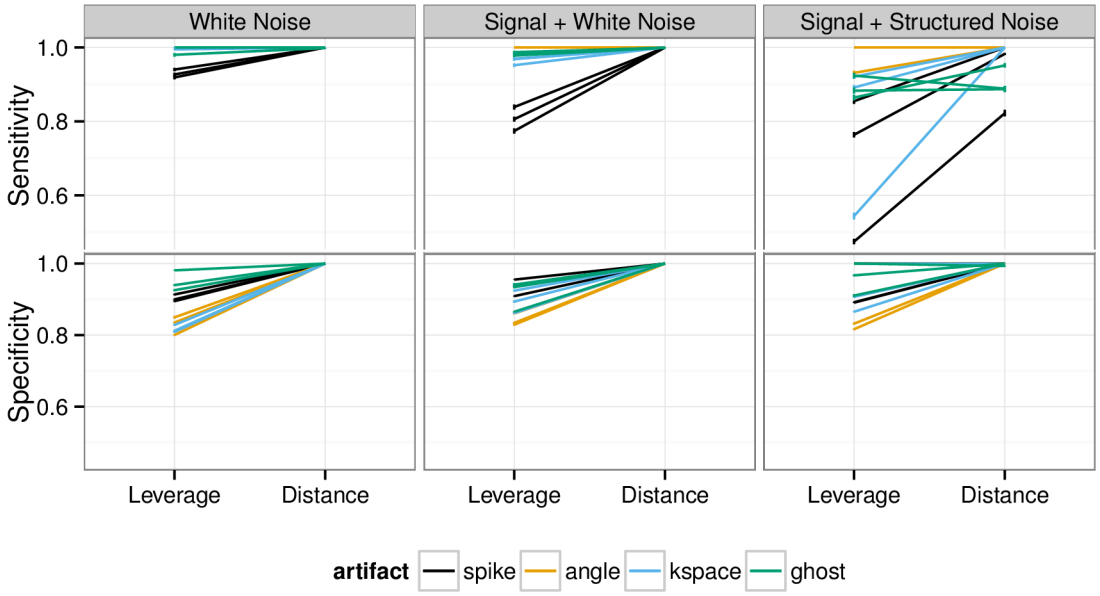


Figure 3.2: Sensitivity and specificity of each method in the presence of artifacts by simulation setup. Each line represents a different scan for each artifact type and shows the mean across 1000 iterations with a 95% bootstrap confidence interval.

3.4 Experimental Data Results

Using a large, multi-site fMRI dataset, we assess the result of outlier removal on the scan-rescan reproducibility of a common analysis, identification of subject-level brain networks through ICA. This section is organized as follows. We begin with a description of the dataset employed and show an example of the leverage and distance measures. We then describe how we identify subject-level brain networks through GICA and dual regression, and how we assess reproducibility of these networks. Finally, we quantify the improvement to reproducibility with the proposed outlier detection methods using a linear mixed model to account for subject and site effects.

3.4.1 fMRI Dataset

We use data from the publicly available Autism Brain Imaging Data Exchange (ABIDE), which contains neuroimaging and phenotypic information collected from 1203 subjects, including 583 subjects with autism spectrum disorder (ASD) and 620 typically developing (TD) subjects.[57] The ABIDE consists of 20 datasets collected at 16 international sites. Table 3.2 lists each dataset and the number of subjects in each. Image acquisition parameters and demographic information for each dataset are available at `http://fcon_1000.projects.nitrc.org/indi/abide/`. For each subject, a T_1-weighted MPRAGE volume and one or more resting-state fMRI (rs-fMRI) sessions were collected on the same day.

Image pre-processing consisted of the following steps. SPM12b's segmentation tool was first used to correct for broad intensity variations across the MPRAGE volume; the bias-corrected MPRAGE was then registered to the first (stabilized) functional volume and normalized to Montreal Neuological Institute (MNI) space. Volumes corresponding to the first 10 seconds of the rs-fMRI scan were dropped to allow for magnetization stabilization. The remaining volumes were slice-time adjusted using the slice acquired at the middle of the repetition time (which varied by site). Rigid body realignment parameters were estimated with respect to the first (stabilized) functional volume of the rs-fMRI scan and used to calculate mean framewise displacement (FD), a summary measure of between-volume participant motion.[62] The non-linear spatial transformation estimated from the co-registered MPRAGE was

Dataset	N	N^+	N_{ICA}	Q^+
California Institute of Technology (Caltech)	38	19	19	8
Carnegie Mellon University (CMU)	27	18	18	5
Kennedy Krieger Institute (KKI)	146	140	50	8
University of Leuven: Sample 1 (Leuven 1)	29	23	23	6
University of Leuven: Sample 2 (Leuven 2)	35	31	31	10
Ludwig Maximilians University Munich (LMU)	57	55	50	10
NYU Langone Medical Center (NYU)	184	108	50	13
Oregon Health and Science University (OHSU)	28	28	28	8
Olin Institute of Living at Hartford Hospital (Olin)	36	29	29	4
University of Pittsburgh School of Medicine (Pitt)	57	54	50	12
Social Brain Lab, the Netherlands (SBL)	30	30	30	9
San Diego State University (SDSU)	36	32	32	14
Stanford University (Stanford)	40	35	35	5
Trinity Centre for Health Sciences (Trinity)	49	47	47	11
University of California, Los Angeles: Sample 1 (UCLA 1)	82	44	44	11
University of California, Los Angeles: Sample 2 (UCLA 2)	27	18	18	6
University of Michigan: Sample 1 (UM 1)	110	89	50	6
University of Michigan: Sample 2 (UM 2)	35	34	34	9
University of Utah School of Medicine (USM)	101	94	50	10
Yale Child Study Center (Yale)	56	46	46	11

Table 3.2: For each dataset, the total number of subjects (N), number of subjects that passed quality inspection (N^+), number of subjects used to perform GICA (N_{ICA}), and the number of signal GICA networks identified (Q^+).

then applied to the functional data along with the estimated rigid body realignment parameters and resulted in 2-mm isotropic voxels in MNI space. Each resting state scan was temporally detrended on a voxelwise basis and spatially smoothed using a 5-mm full width at half maximum (FWHM) Gaussian kernel.[80]

After pre-processing, each rs-fMRI scan was quality inspected for motion and issues with registration and normalization using the following procedure. First, scans were flagged for quality if mean FD across the scan was greater than 2 standard deviations above the sample mean. We then calculated the Pearson spatial correlation between the first (stabilized) volume of each subject's MNI-registered data and SPM's EPI template.[81] In total, 229 subjects were found to have major quality problems. Table 3.2 displays the number of subjects from each data collecting site that passed quality inspection. All scans were included in our analysis to assess the effect of outlier removal; however, only those scans that passed quality inspection were used to create group-level ICA maps.

For a single example scan, Figure 3.3 shows the leverage and robust distance functions, along with 6 motion parameters (roll, pitch, yaw, and translation in each direction) and their derivatives. Volumes labeled as outliers using the leverage and distance functions are indicated in red. Leverage and distance outliers are indicated with a red dot. Below the plot, the volumes corresponding to the spikes at time points 60, 90, 135 and 150 (shaded on the plot) are shown. Three of the spikes are leverage and distance outliers, while the spike at time point 90 is only a leverage outlier.

Obvious banding artifacts are seen at time points 60 and 150, a moderate banding artifact is seen at time point 135, and no visible artifact is apparent at time point 90. While the artifact at time point 150 would be detected using motion measures, the other spikes would likely go undetected using only motion.

Figure 3.4 shows the distribution of the proportion of volumes in each scan identified as outliers using each method. For 17.6% of subjects, no leverage outliers are identified, while for 4.6% of subjects no distance outliers are identified. The percent of subjects in each dataset with no outliers identified using either method varies widely, reflecting the heterogeneity in data quality across ABIDE datasets.

3.4.2 Identification of subject-level brain networks through GICA and dual regression

We perform GICA separately for each of the 20 ABIDE datasets using the following procedure. For datasets containing more than 50 subjects that passed quality inspection, we use 50 randomly selected subjects from among these; otherwise we use all subjects that passed quality inspection. Let $\mathcal{N}_k$ be the resulting set of subjects for dataset k. For each subject $i \in \mathcal{N}_k$, let $\mathbf{Y}_i$ be the $T_i \times V_k$ data matrix after centering each voxel across time, where V_k is the number of voxels in the group-level brain mask for dataset k. For each subject $i \in \mathcal{N}_k$, we perform PCA and retain 50 PCs to obtain $\mathbf{Y}_i = \mathbf{U}_i\mathbf{D}_i\mathbf{V}_i^t + \mathbf{E}_i$, resulting in the reduced $50 \times V$ subject-level data $\tilde{\mathbf{Y}}_i = \mathbf{D}_i\mathbf{V}_i^t$.

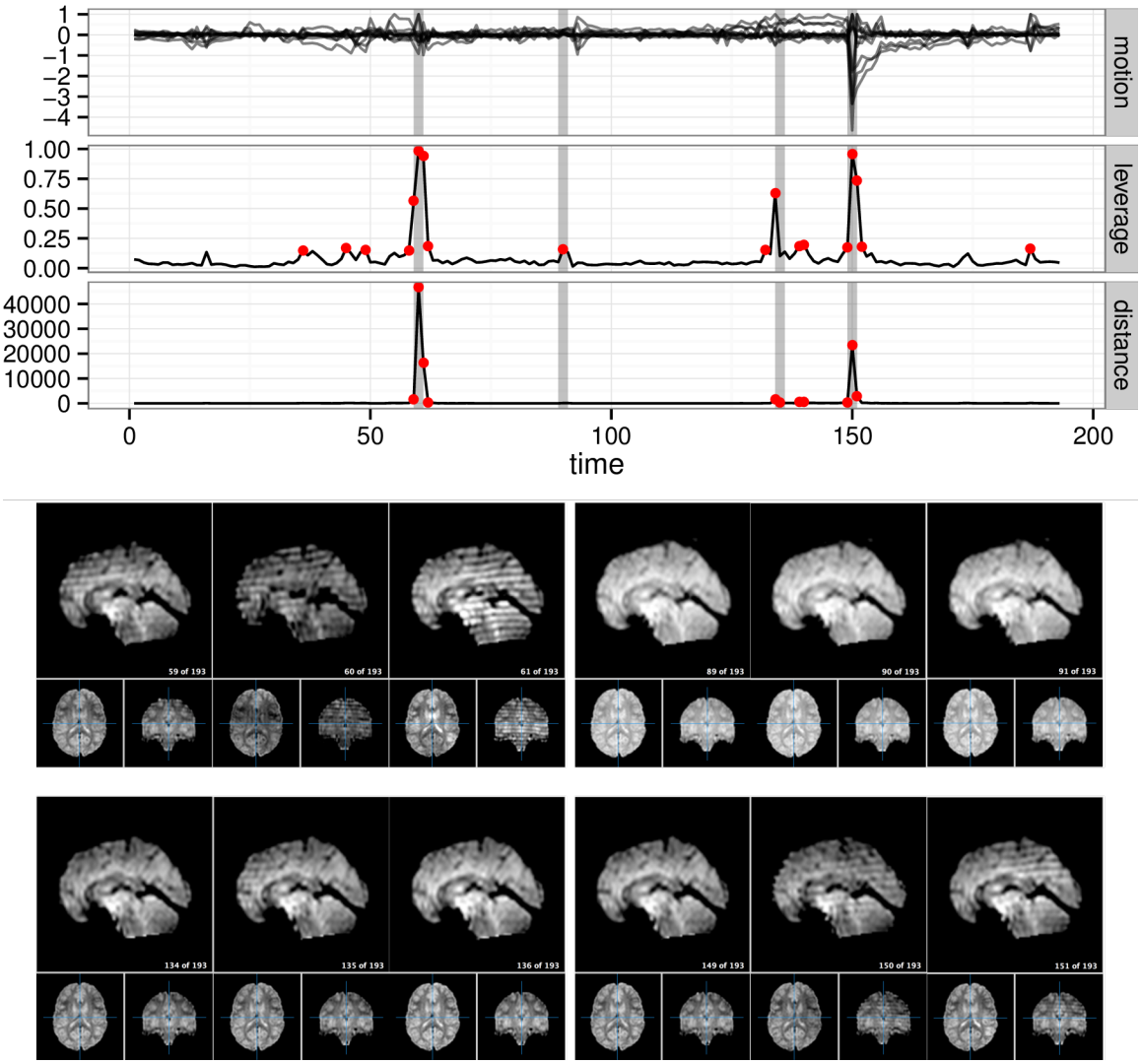


Figure 3.3: For a single subject, the motion parameters, leverage function, and robust distance function. Leverage and distance outliers are indicated in red. Below the plot, the volumes corresponding to the spikes at time points 60, 90, 135 and 150 (shaded on the plot) are shown. Three of the spikes are leverage and distance outliers, while the spike at time point 90 is only a leverage outlier. Obvious banding artifacts are seen at time points 60 and 150, a moderate banding artifact is seen at time point 135, and no visible artifact is apparent at time point 90. While the artifact at time point 150 would be detected using motion measures, the other spikes would likely go undetected using only motion.

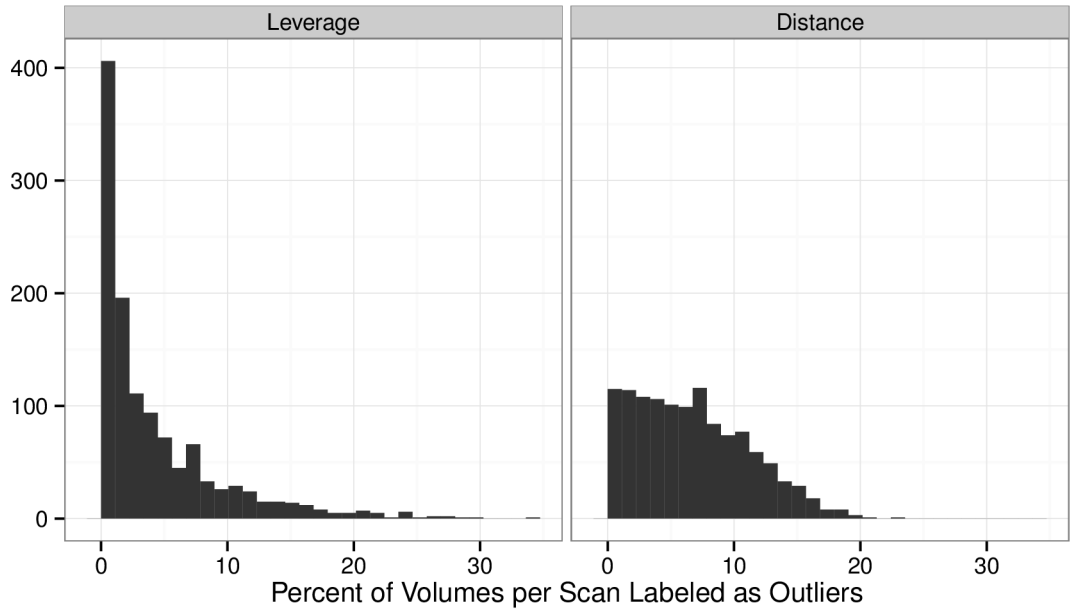


Figure 3.4: Distribution of the percent of volumes in each scan identified as outliers using each outlier removal method. The leverage method tends to identify no or very few outliers more often than the distance method.

(Note that for ICA we use PCA to reduce dimensionality along the temporal dimension, whereas for outlier detection we reduce along the spatial dimension.) Next, we temporally concatenate all subjects to form the $50\mathcal{N}_k \times V$ matrix $\mathbf{Y}_k$. We then perform PCA again with $Q = 30$ components to obtain $\mathbf{Y}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^t + \mathbf{E}_k$, resulting in the reduced $Q \times V$ group-level data $\tilde{\mathbf{Y}}_k = \mathbf{D}_k\mathbf{V}_k^t$. Finally, we perform the fastICA algorithm[82] to obtain $\tilde{\mathbf{Y}}_k = \mathbf{A}_k\mathbf{S}_k$, where $\mathbf{S}_k$ is a $Q \times V$ matrix whose rows contain the group-level spatial independent components and $\mathbf{A}_k$ is the $Q \times Q$ mixing matrix.

To identify "signal" independent components (ICs), we first standardize each spatial IC by subtracting its mean and dividing by its standard deviation and threshold the result at ± 2. We then visually inspect each spatial IC and label those corre-

sponding to known resting-state brain networks. This results in 4-14 signal ICs per dataset (Table 3.2). While the number of signal ICs identified for some datasets is quite small, this is not surprising given the widely varying quality and quantity of data in each dataset. We observe a positive association between sample size and number of signal ICs: based on a simple no-intercept linear model, we estimate that for every additional subject included in analysis, on average 0.23 (95% CI: [0.20, 0.26]) additional signal ICs are identified through GICA. As more subjects are included in GICA, more resting-state brain networks can be clearly identified. Let $\mathbf{S}_k^+$ denote the $Q_k^+ \times V$ matrix containing only the Q_k^+ signal components identified for dataset k.

To obtain subject-level ICs, we perform dual regression[79] as follows. Let $\mathbf{S}_k^+$ be centered and scaled across voxels and $\mathbf{Y}_i$ be centered and scaled across time as described above. Let $\dot{\mathbf{Y}}_i$ be $\mathbf{Y}_i$ after also centering each time point across voxels. In the first regression, temporal ICs for subject $i \in \mathcal{N}_k$ are obtained by regressing $\dot{\mathbf{Y}}_i^t$ against $\mathbf{S}_k^{+t}$ to obtain $\mathbf{A}_i^t = \left(\mathbf{S}_k^+ \mathbf{S}_k^{+t}\right)^{-1} \mathbf{S}_k^+ \dot{\mathbf{Y}}_i^t$. In the second regression, subject-level spatial ICs for subject $i \in \mathcal{N}_k$ are obtained by regressing $\dot{\mathbf{Y}}_i$ against $\mathbf{A}_i$ to obtain $\mathbf{S}_i = \left(\mathbf{A}_i^t \mathbf{A}_i\right)^{-1} \mathbf{A}_i^t \dot{\mathbf{Y}}_i$. This results in the ICA decomposition $\dot{\mathbf{Y}}_i \approx \mathbf{A}_i \mathbf{S}_i$, where $\mathbf{A}_i$ is $T_i \times Q_k^+$ and $\mathbf{S}_i$ is $Q_k^+ \times V_k$. We are interested in $\mathbf{S}_i$, whose rows contain the vectorized ICA-estimated brain networks for subject i.

3.4.3 Measuring reproducibility of subject-level brain networks

Let $\mathbf{S}_{i1}$ and $\mathbf{S}_{i2}$ be two sets of spatial ICs for subject i obtained by performing dual regression separately for two different scanning sessions of subject $i \in \mathcal{N}_k$ (but using the same group ICs $\mathbf{S}_k^+$). Note that using the same group ICs avoids the "ICA matching problem", since the ICs in $\mathbf{S}_{i1}$ and $\mathbf{S}_{i2}$ correspond to the same group ICs in $\mathbf{S}_k^+$. To assess reliability of subject-level spatial ICs, for each subject i and component q we compute the number of overlapping voxels between $\mathbf{S}_{i1}(q)$ and $\mathbf{S}_{i2}(q)$ after both have been thresholded at ± 2 standard deviations. We then average over all signal components to obtain the average scan-rescan overlap per component for each subject, denoted Z_{ikm} for subject i in dataset k using outlier removal method $m \in \{$none, leverage, robust distance$\}$. This is our measure of reproducibility of subject-level brain networks identified through ICA.

Note that in the ABIDE dataset most subjects have only a single scanning session. To simulate scan-rescan data we simply split each subject's data into two "sessions", consisting of the first $\lfloor T_i/2 \rfloor$ volumes and the remaining volumes. While this may over-estimate the true scan-rescan overlap, we are primarily interested in the change in overlap due to outlier removal, rather than the absolute amount of overlap.

To test for changes in scan-rescan overlap due to each outlier removal method, we fit a linear mixed effects model with a fixed effect for each method, a fixed effect for

each dataset, and a random intercept for each subject. We employ this model for its ability to test several groups and methods simultaneously and to account for within-subject correlation across methods. We fit the model on all subjects for whom at least one leverage or distance outlier was identified (96.3% of subjects). We therefore estimate the following model:

$$Z_{ikm} = b_{i0} + \gamma_k + \alpha_m I_{m>0} + \epsilon_{ikm},\ \epsilon_{ikm} \sim N(0, \sigma_e^2),\ b_{i0} \sim N(0, \sigma_u^2) \qquad (3.3)$$

where $m = 0$ indicates no outlier removal. Here, γ_k represents the average scan-rescan overlap per component for subjects in dataset k when no outlier removal is performed, and α_m represents the change in overlap per component when outlier removal method m is used. To obtain coefficient estimates, we fit this model using the `lme` function from the `nmle` package.[83] Since we have a large sample size, we compute Normal 95% confidence intervals.

3.4.4 Effect of outlier removal on reproducibility

Figure 3.5 shows estimates and 95% confidence intervals for the model coefficients. The left-hand plot displays the coefficients for the outlier removal methods (α_m). Both methods significantly improve reproducibility of subject-level brain networks identified through ICA. However, removal of leverage outliers results in greater improvement than removal of distance outliers. Specifically, removal of leverage outliers

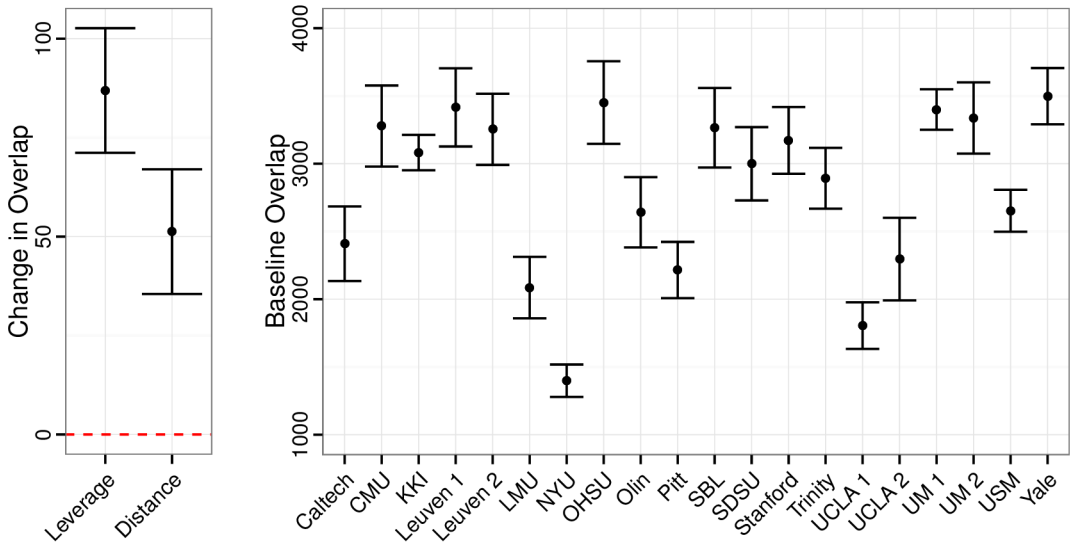


Figure 3.5: Estimates and 95% confidence intervals for the model coefficients. The left-hand plot displays the coefficients for each outlier removal method (α_m); both methods result in statistically significant improvement to reproducibility of subject-level brain networks identified through ICA, and removal of leverage outliers results in greater improvement than removal of distance outliers. The right-hand plot displays the fixed effects for each dataset (γ_k) and illustrates the heterogeneity in reproducibility of ICA results across ABIDE datasets.

results in an increase in scan-rescan overlap of 87 [95% CI: 71-103] voxels per component, while removal of distance outliers results in an increase of 51 [95% CI: 36-67] voxels per component. The right-hand plot of Figure 3.5 displays the fixed effects for each dataset (γ_k) and illustrates the heterogeneity in reproducibility of ICA results across ABIDE datasets before outlier removal.

We also stratify the model by those subjects who passed and did not pass quality inspection. Figure 3.6 shows estimates and 95% confidence intervals for the model coefficients after stratification. The left-hand plot shows that subjects who failed

quality inspection (QC = 0) tend to improve more than those who passed quality inspection (QC = 1). However, data from both groups of subjects benefit from outlier removal. Specifically, removal of leverage outliers results in an increase in scan-rescan overlap of 103 [95% CI: 67-140] voxels per component for subjects who failed quality inspection, and of 83 [95% CI: 66-101] voxels per component for subjects who passed quality inspection. Removal of distance outliers increases scan-rescan overlap by 70 [95% CI: 33-106] voxels per component for subjects who failed quality inspection, and by 47 [95% CI: 30-64] voxels per component for subjects who passed quality inspection. The right-hand plot of Figure 3.6 shows that, in general, subjects who failed quality inspection tend to have lower baseline reproducibility than those who passed quality inspection. It also again illustrates the heterogeneity in data quality across ABIDE datasets, even within subjects who passed quality inspection.

3.5 Discussion

We have proposed a method to detect outlying time points in an fMRI scan by drawing on the traditional statistical ideas of PCA, leverage, and outlier detection. The proposed methods have been validated through simulated data and a large, diverse fMRI dataset. We have demonstrated that the proposed methods are accurate and result in improved reproducibility of a common type of analysis for resting-state fMRI data, identification of brain networks through ICA. While other types of analysis

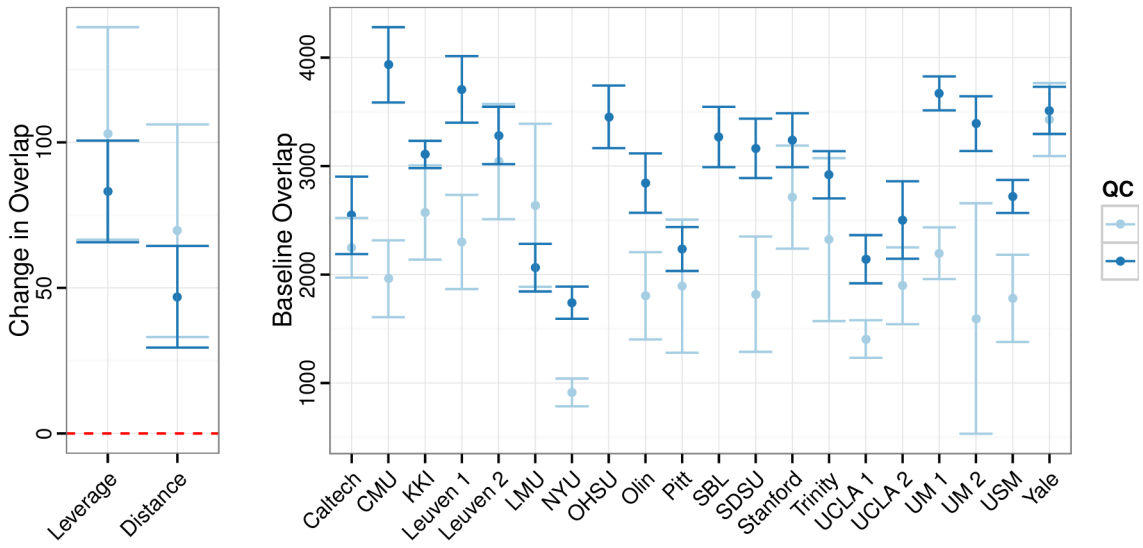


Figure 3.6: Estimates and 95% confidence intervals for the model coefficients after stratifying by quality inspection results. The left-hand plot shows that while both groups of subjects benefit from outlier removal, subjects who failed quality inspection (QC= 0) tend to benefit more than those who passed quality inspection (QC= 1). The right-hand plot shows that subjects who failed quality inspection tend to have lower baseline reproducibility than those who passed quality inspection.

could have also been used for this purpose, such as estimation of pairwise correlation between voxels or regions of interest, we chose to focus on a more difficult problem.

The proposed techniques are, to the best of our knowledge, the first to provide a single measure of outlyingness for time points in an fMRI scan, which can be easily thresholded to identify outliers. Unlike motion-based outlier detection methods for fMRI, they are agnostic to the source of artifact. Furthermore, PCA leverage is directly related to the estimation of principal components, which are used to compute independent components and are also a representation of the covariance matrix used in seed connectivity analysis.

One limitation of our approach is that we perform validation on a single dataset, the ABIDE. However, this dataset is in fact a diverse collection of 20 datasets from 16 international sites, which strengthens the generalizability of our results. Another limitation of the proposed methods is that they may be sensitive to the number of principal components retained. However, we have found that the method performs well with different model orders (e.g. 20 or 30), and we propose an automated method of selecting model order, in order to provide a fully automated approach. A limitation of the PCA leverage method is that the thresholding rule is, as in regression, somewhat arbitrary. However, use of the median leverage across observations as a benchmark is a sensible approach, and we tested a range of values before selecting the proposed cutoff of 3 times the median. We believe this value will work well in general for resting-state fMRI, but for different types of data the researcher may wish to re-evaluate this

choice.

While the proposed methods have been designed and validated for resting-state fMRI data, they may be easily extended to other types of medical imaging data, such as task fMRI and EEG data, as well as other types of high-dimensional data. Furthermore, they may also be extended to group analyses; future work should focus on exploring these directions.

As the availability of large fMRI datasets continues to grow, automated outlier detection methods are becoming essential for the effective use of such data. In particular, the reliability of analyses employing these diverse datasets may be negatively impacted by the presence of poor quality data. The outlier detection approach we propose stands to improve the quality of such datasets, thus enhancing the possibilities to use these data to understand neurological diseases and brain function in general.

Funding

This work was supported by the National Institutes of Health [R01 EB016061] and the National Institute of Biomedical Imaging and Bioengineering [P41 EB015909]. *Conflict of Interest*: None declared.

Chapter 4

Scan Length, Shrinkage and Reliability of Resting-State Functional Connectivity in the Human Connectome Project

4.1 Introduction

In recent years, there has been a proliferation of studies of brain connectivity utilizing resting-state functional magnetic resonance imaging (rsfMRI). Interest in estimating resting-state functional connectivity (rsFC) for individual subjects continues to grow, due in part to its potential to shed light on cognitive changes associated

with certain psychological disorders, typical neurological development, and aging-related cognitive decline. For example, group differences in rsFC have been observed in autism,[84,85] attention deficit hyperactivity disorder,[86,87] depression,[88] Alzheimer's disease,[89,90] mild cognitive impairment,[91] and schizophrenia,[92,93] while normal aging has also been shown to coincide with changes in rsFC.[94]

Meanwhile, there is a growing recognition of the importance of accuracy of results in psychological science[95–97] and specifically for the study of group and inter-individual differences in rsFC.[98] Accuracy of rsFC estimates is typically a greater concern for the study of inter-individual differences, as subject-level estimates are typically much noisier than group averages. While sampling variability does play a role in the accuracy of rsFC estimates, it is also widely recognized that rsFC varies both within and across scanning sessions due to changes in the cognitive and emotional state of the subject.[19,98] This introduces a challenge in the study of associations between subject-level rsFC and behavioral measures or traits, which are expected to be relatively stable over time. Another important use of subject-level rsFC is in identifying important functional areas for presurgical planning,[94] for which accuracy is clearly vital. There is also growing interest in achieving sufficient within-subject accuracy to be able to distinguish an individual from a group of subjects based on rsFC (e.g. *fingerprinting*) which may be useful in the clinical setting for diagnosis and treatment.[99–101]

Broadly speaking, the accuracy of an estimate can be assessed in terms of validity or reliability. Here validity refers to the degree an estimate measures what it purports

to measure, and reliability to the consistency of an estimate across different samples. In the context of measures derived from fMRI data, validity is often difficult to assess due to the absence of a ground truth measure. Still, there have been some efforts to assess validity of the existence of connections through simulations[102] or careful modeling of anatomical connectivity.[103] These efforts have provided insight into the accuracy of various methods of modeling connectivity, including full correlation, partial correlation, inverse covariance, lag-based models and Bayes? net. The accuracy of the estimated *strength* of connections, which is often of interest in studies of group and inter-individual differences, has been assessed primarily through reliability, which only requires multiple observations rather than knowledge of the true rsFC of each subject. Due to the growing availability of test-retest rsfMRI datasets and the continued interest in the study of brain connectivity, there has been a recent proliferation of studies focusing on reliability of rsFC. Here we provide a brief review of the methods and findings of this literature.

The manner in which reliability is defined across studies varies widely, but can grouped into three broad categories: intersession reliability, intrasession reliability, and end-point reliability. *Intersession reliability* asks, how similar is the rsFC of a subject across multiple scanning sessions occurring days or even months apart.[101] This is usually the true measure of interest, especially for the study of group differences or inter-individual differences in rsFC associated with traits or behavioral measures that display minimal variation over time. *Intrasession reliability*, on the other hand, asks

how similar is the rsFC of a subject across multiple runs within the same session. This tends to overestimate true intersession reliability,[19,98,99,104] but can serve as a reasonable alternative when multiple rsfMRI sessions are not available. In some cases, intrasession reliability be of interest, for example for the study of associations between rsFC and behavioral measures that vary across sessions. A related measure of reliability is based on alternating epochs within the same session. This tends to overestimate true intrasession reliability, since variability in true rsFC is likely to be greater across two contiguous epochs than across two sets of alternating epochs spanning the same time period.[19] Finally, *end-point reliability* asks, how similar is an estimate of rsFC produced from the first t minutes of a scanning session to the estimate produced using the full session of $T \geq t$ minutes. While this may provide insight into the rate of convergence of a series of estimates, it is not a true measure of reliability in the strictest sense, as the estimates being compared are not independent. Specifically, as t approaches the full scan duration T the two estimates become more dependent, and as a consequence the estimate produced from the first t minutes becomes more similar to the estimate produced using the full T minutes, eventually reaching equality at $t = T$. End-point reliability therefore may greatly overestimate the true intrasession and intersession reliability of rsFC estimates.

Studies also differ in terms of the metric used to quantify reliability. Several common metrics are intra-class correlation coefficient (ICC), correlation, and mean squared error (MSE).[105–108] Among these measures, ICC has the advantage of tak-

ing into account both within-subject variability and between-subject variability. It therefore can tell us something about how similar multiple observations from the same subject are to each other, relative to the differences between subjects. However, ICC itself can be defined in various ways, and may range from 0 to 1 or -1 to 1 depending on the definition employed One should therefore take care when interpreting and comparing estimated reliability levels across different studies.

Many studies have focused on determining the scanning duration needed to accurately estimate rsFC[19,94,99] and related measures, including spatial topography of resting-state networks,[94,109,110] regional homogeneity,[104,111] graph theoretic metrics,[112,113] and task activation.[114] While it has been clearly shown that increased scan duration leads to improved reliability of rsFC and related measures,[94,99,102–104] different studies have reached vastly different conclusions about what scan duration is sufficient, with recommendations ranging from 5 minutes,[112,113] to 90 minutes or more.[18] This discrepancy may be attributed to the fact that there are many other factors than scan length that have been shown to influence reliability , in addition to the differences in the definition of reliability described above. A number of studies have provided evidence for best practices to improve reliability through the methods employed for acquisition, preprocessing, and connectivity analysis.

In terms of acquisition, increasing temporal resolution[19,104,113] and having subjects lie with eyes open rather than closed[94] have been shown to result in improved reliability. The result of increasing spatial resolution is more controversial, with[104]

finding that it results in improved reliability but[114] finding that while it allows for better separation of physiological noise and spatial localization of activation, it also results in reduced temporal signal-to-noise ratio, thereby requiring a simultaneous increase in scan duration to maintain a given level of power. In terms of acquisition length and timing, Shehzad et al. (2009)[98] and Laumann et al. (2015)[18] find that combining estimates of rsFC from multiple sessions produces more reliable estimates of rsFC than using data from a single scan of the same (combined) duration.

In terms of preprocessing, several studies have found that global signal regression tends to worsen reliability,[104,113] while nuisance regression tends to improve reliability.[104] In addition, performing analysis in surface rather than volumetric space may result in improved reliability.[104]

In terms of the methods used to perform connectivity analyses, the use of functional versus anatomical regions of interest (ROIs)[99] and accurate identification of functional ROIs[102] have been found to improve reliability. Furthermore, different methods of estimating connectivity (e.g. full correlation, inverse covariance, partial correlation) vary in terms of their ability to identify true connections.[102,103] Reliability has also been shown to vary with the size of the regions considered[110] and the specific connections being studied.[18,94,98,99,115] For example, higher reliability has been observed for correlations that are statistically significant at the group level; for within-network versus between-network correlations; and for connections within the DMN network versus within the task positive, attention, motor and visual net-

works.[18,94,98]

As many of the factors described above may not always be subject to manipulation, in particular the specific connections of interest, there is unlikely to be a single answer to the question of how long to scan to produce sufficiently reliable estimates of rsFC. It therefore remains important to both increase the duration and/or number of scans whenever possible and to adopt best practices for improving reliability. One practice that has been shown to improve reliability of subject-level rsFC and parcellations is *shrinkage*, in which subject-level observations "borrow strength" from the group mean. In Mejia et al. (2015),[116] we considered voxel-level rsFC produced using relatively short (5-7 minute) scans and showed that performing shrinkage resulted in improvement in reliability of rsFC by 25-30%. While this provided clear evidence for the benefits of shrinkage for voxel-level rsFC in cases when longer scans are not feasible, it was not clear whether shrinkage would remain beneficial with increased scan duration and ROI size, or with the adoption of "best practices" such as increasing temporal resolution and transformation to surface space. It was also not clear whether shrinkage would be beneficial for the most reliable connections, such as connections between regions within the DMN. To address these questions, in this paper we use the Human Connectome Project (HCP) to explore the effect of shrinkage on reliability of whole-brain rsFC between regions identified through ICA, using much longer (up to 30 minutes) scans and an acquisition and preprocessing pipeline that has been designed to produce very high-quality rsfMRI data. Furthermore, the data has been

"denoised" to remove sources of non-neuronal variation and has been transformed to surface space. Therefore, the HCP is nearly ideal in terms of the best practices outlined above and longer scan duration, providing an opportunity to evaluate the benefits of shrinkage in a such a setting.

4.2 Methods

4.2.1 Data and connectivity estimation

4.2.1.1 Human Connectome Project Data

The Human Connectome Project (HCP) is a collection of neuroimaging and phenotypic information for over a thousand healthy adult subjects, 523 of which were publicly released to this date (`http://humanconnectome.org`).[8] All data were acquired on a customized 3T Siemens connectome-Skyra 3T scanner, designed to achieve 100 mT/m gradient strength. For the analyses described below, we use the following data provided as part of the HCP data release. For 461 of the 523 subjects, a multi-band / multi-slice pulse sequence with an acceleration factor of eight[117–121] was used to acquire four roughly 15-minute rsfMRI sessions, each consisting of 1200 volumes sampled every 0.72 seconds at 2 mm isotropic spatial resolution. The sessions were collected over two visits that occurred on separate days, with two sessions collected

at each visit. Across sessions at each visit, phase encoding directions were alternated between right-to-left (RL) and left-to-right (LR) directions. Before October 1, 2012, the first session of each visit was acquired with RL phase encoding, and the second session was acquired with LR phase encoding (RL/LR). After this date, the first visit continued to be acquired in the RL/LR order, but the second visit was acquired in the opposite order, with the LR acquisition followed by the RL acquisition (LR/RL).

Spatial preprocessing was performed using the minimal preprocessing pipeline as described by Glasser et al. (2013),[122] which includes correcting for spatial distortions and artifacts and projection of the data time series to the standard grayordinate space. Structured artifacts in the time series were removed using ICA + FIX (independent component analysis followed by FMRIB's ICA-based X-noiseifier),[123,124] and each data set was temporally demeaned with variance normalization according to Beckmann and Smith (2004).[125] Group independent component analysis (GICA) was performed on the full rsfMRI time series for all 461 subjects to estimate a set of spatial independent components (ICs) that represent population-average resting-state networks.[125] GICA was performed using model orders of 25, 50, 100, 200 and 300 independent components (ICs). After identification of spatial ICs at each model order, time courses were estimated for each subject and IC by performing the first stage of dual regression.[79] Specifically, the group IC spatial maps were used as predictors in a multivariate linear regression model against the full rsfMRI time series, which was created by concatenating the four sessions of each subject into a single time series in

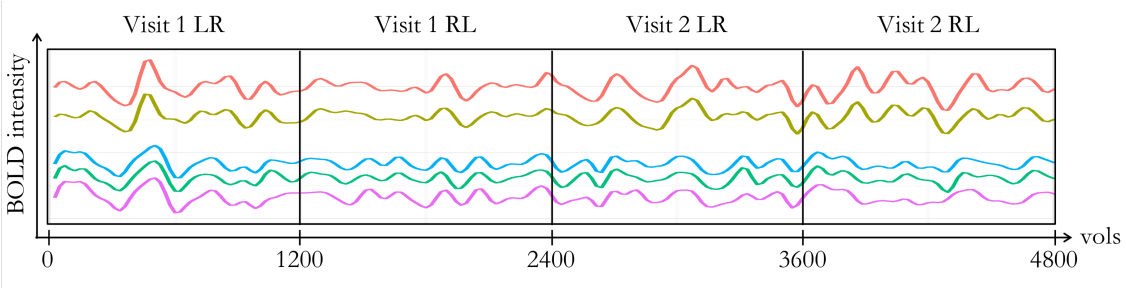


Figure 4.1: Illustration of the data for a single subject. For five regions, the full time series, consisting of 4800 volumes, is shown. Each time series consists of four sessions, occurring over two visits. In the GICA provided in the HCP data release, the sessions were reordered so that both visits are concatenated in the LR/RL order.

the following order: visit 1 LR, visit 1 RL, visit 2 LR, visit 2 RL (see Figure 4.1).

4.2.1.2 Connectivity matrix estimation

The quantity of interest for each subject is the true $Q \times Q$ connectivity matrix, representing the pairwise connectivity during rest between each of the $Q \in \{25, 50, 100, 200, 300\}$ regions identified through GICA. We are interested in how the reliability of estimates of this connectivity matrix changes with longer scan duration. To this end, we estimate the connectivity matrix using the first ℓ volumes of the time series for each subject at both visits, with $\ell \in \{300, 600, \ldots, 2400\}$. At TR of 0.72 seconds, the resulting time series range from 3.6 to 28.8 minutes in duration.

Shrinkage of connectivity estimates. Shrinkage estimators, which "borrow strength" from the population to improve subject-level estimates, have been shown to improve reliability of voxel-level connectivity estimates based on short rsfMRI scans.[43,116] Here, we assess the ability of shrinkage estimators to improve reliability of connectiv-

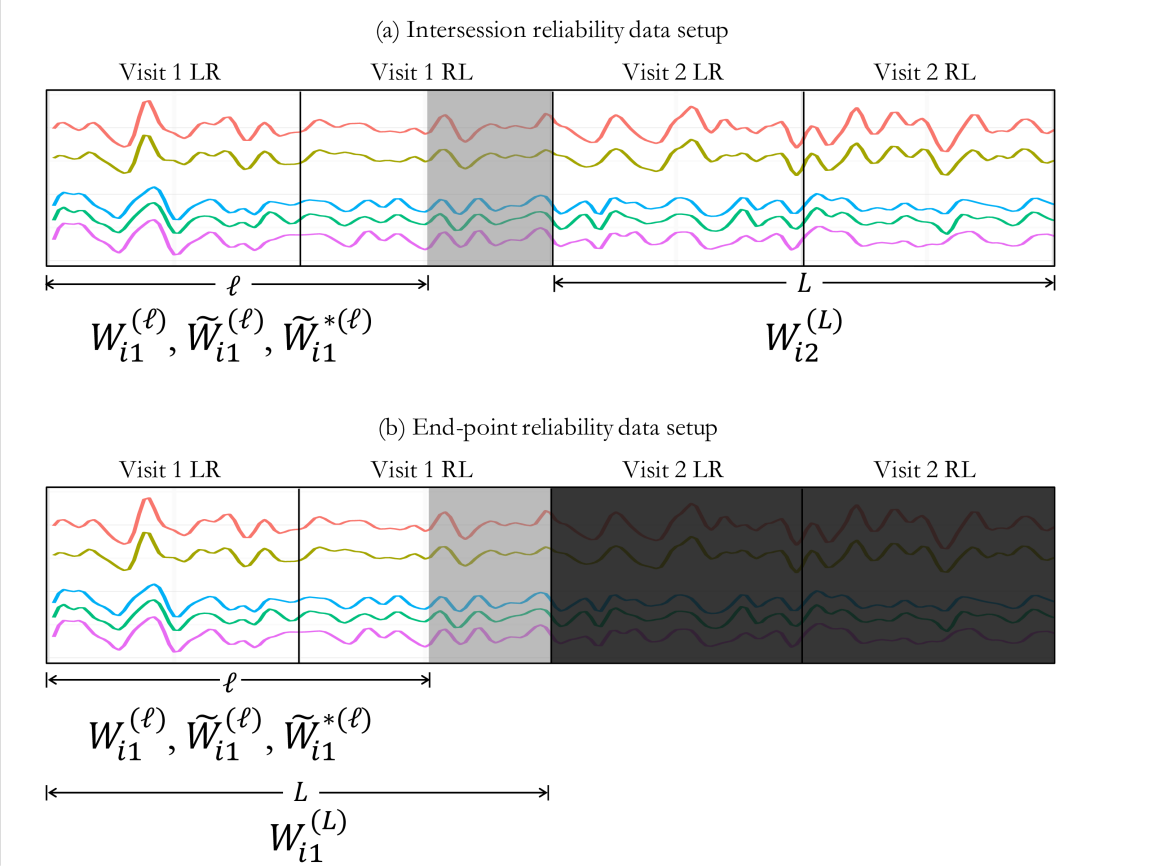


Figure 4.2: Illustration of data setup for intersession and end-point reliability analysis for a single subject i. For intersession reliability analysis, we are interested in how similar the connectivity estimates $W_{i1}^{(\ell)}$, $\tilde{W}_{i1}^{(\ell)}$ and $\tilde{W}_{i1}^{*(\ell)}$ are to the full visit 2 raw estimate $W_{i2}^{(L)}$, $L = 2400$, as ℓ varies from 300 to 2400. For end-point reliability analysis, we are interested in how close the connectivity estimates $W_{i1}^{(\ell)}$, $\tilde{W}_{i1}^{(\ell)}$ and $\tilde{W}_{i1}^{*(\ell)}$ are to the full visit 1 raw estimate $W_{i1}^{(L)}$, as ℓ varies from 300 to 2400.

ity estimates produced from longer scans.

We first provide a brief introduction to empirical Bayes shrinkage estimators. We start with a simple measurement error model.[44] For subjects $i = 1, \ldots, n$, let the true connectivity between two regions q and q' be denoted $X_i(q, q')$. For visit j and scan length ℓ, we have an estimate of $X_i(q, q')$, which we denote $W_{ij}^{(\ell)}(q, q')$. The measurement error model assumes that the estimate $W_{ij}^{(\ell)}(q, q')$ can be decomposed into a signal $X_i(q, q')$ and a noise term $U_{ij}^{(\ell)}(q, q')$:

$$W_{ij}^{(\ell)}(q, q') = X_i(q, q') + U_{ij}^{(\ell)}(q, q'), \tag{4.1}$$

where $X_i(q, q') \sim N\{\mu(q, q'), \sigma_x^2(q, q')\}$ and $U_{ij}^{(\ell)}(q, q') \sim N\{0, \sigma_u^{2(\ell)}(q, q')\}$. We assume that $X_i(q, q')$ and $U_{ij}^{(\ell)}(q, q')$ are independent, the $X_i(q, q')$ are independent across subjects, and the $U_{ij}^{(\ell)}(q, q')$ are independent across subjects and repetitions. Then the shrinkage estimator of $X_i(q, q')$ is equal to the empirical posterior mean,

$$\tilde{W}_{ij}^{(\ell)}(q, q') = \lambda_t(q, q')\bar{W}_{\cdot j}^{(\ell)}(q, q') + \{1 - \lambda_t(q, q')\} W_{ij}^{(\ell)}(q, q'),$$

where $\bar{W}_{\cdot j}^{(\ell)}(q, q') = \frac{1}{n} \sum i = 1^n W_{ij}^{(\ell)}(q, q')$. The shrinkage parameter $\lambda_t(q, q')$ is given by

$$\lambda_t(q, q') = \frac{\sigma_u^{2(\ell)}(q, q')}{\sigma_u^{2(\ell)}(q, q') + \sigma_x^2(q, q')}$$

and ranges from 0 (no shrinkage) to 1 (complete shrinkage to the group mean), de-

pending on the relative size of the within-subject and between-subject variance terms. Estimation of these variance components is, in theory, straightforward: defining $\sigma_w^{2(\ell)}(q, q')$ as the total variance, which is estimated as the variance of the $W_{ij}^{(\ell)}(q, q')$, averaged over visits, the noise variance can be estimated as

$$\hat{\sigma}_u^{2(\ell)}(q, q') = \frac{1}{2} Var_i\{W_{i2}^{(\ell)}(q, q') - W_{i1}^{(\ell)}(q, q')\},$$

and the signal variance can be estimated as

$$\hat{\sigma}_x^2(q, q') = \hat{\sigma}_w^{2(\ell)}(q, q') - \hat{\sigma}_u^{2(\ell)}(q, q').$$

However, this presupposes the availability of multiple observations or visits for each subject. While this is true in our case, it is not a reasonable assumption in general, as many studies only collect a single resting-state fMRI scan for each subject. Furthermore, when multiple visits or sessions are available, they can be combined to create a single estimate of connectivity with improved accuracy.[18] Therefore, the salient problem is to estimate the within-subject variance of a *single* connectivity estimate produced using *all* of the rsfMRI data available for each subject. This clearly precludes the availability of multiple observations of the quantity of interest. In Mejia et al. (2015),[116] we proposed a solution based on the idea of "pseudo scan-rescan" data, in which a single scanning session is treated as two sessions, composed respectively of the first and second halves of the time series. This approach was also

recently applied in the context of reliability correction by Mueller et al. (2015).[115] In Mejia et al. (2015),[116] we found the within-subject variance estimate produced from this approach to be upwardly biased, and proposed using an empirical adjustment factor to correct for this. However, this adjustment method assumes that the majority of within-subject variance can be attributed to sampling variance, and it is best suited to connectivity estimates produced from noisier time series as in the case of Mejia et al. (2015),[116] where voxel-level connectivity based on short (7.5 minute) scans were estimated. We now present a more general method of estimating within-subject variance from a single session, which seeks to separate sampling variance from other sources of within-subject variance.

Consider a set of independent measurements $\{X_{it}\}$ from subjects $i = 1, \ldots, n$ at time points $t = 1, \ldots, T$. Suppose that the quantity of interest for each subject is some summary statistic across time points, such as the sample mean or variance. Let $Y_{i,\Omega}$ represent the true value of this quantity during the continuous time period $\Omega = [1, T]$, and let $\widehat{Y}_{i,S}$ represent the estimate produced using a discrete set of observations $S \subset \Omega$. For example, let $\widehat{Y}_{i,\mathcal{T}}$ be the estimate of $Y_{i,\Omega}$ based on the full set of measurements $\mathcal{T} = \{1, \ldots, T\}$.

We consider that the true signal can be written as $Y_{i,\Omega} = Z_i + W_{i,\Omega}$, where $Z_i \sim N(\mu, \sigma_z^2)$ is the long-term average of the subject i and $W_{i,\Omega} \sim N(0, \sigma_w^2)$ is the true deviation from that value during time period Ω. We assume that Z_i are independent across subjects and $W_{i,\Omega}$ are independent across non-overlapping time periods Ω. For

any evenly spaced sampling S of Ω, we consider that the estimate $\widehat{Y}_{i,S}$ can be written

$$\widehat{Y}_{i,S} = Y_{i,\Omega} + U_{i,S} = Z_i + W_{i,\Omega} + U_{i,S},$$

where $U_{i,S} \overset{\text{ind}}{\sim} N\left(0, \sigma^2_{u,S}\right)$ and $\sigma^2_{u,S}$ depends upon the sampling S. We further assume that Z_i, $W_{i,\Omega}$ and $U_{i,S}$ are mutually independent.

We are interested in performing empirical Bayes shrinkage on the estimate $\widehat{Y}_{i,\mathcal{T}}$, where the quantity of interest is the true long-term average Z_i. There are two sources of variance around Z_i associated with the estimate $\widehat{Y}_{i,\mathcal{T}}$, the *signal variance* σ^2_w and the *sampling variance* $\sigma^2_{u,\mathcal{T}}$. We must therefore estimate both within-subject variance terms as well as the *population variance* σ^2_z in order to produce the shrinkage parameter,

$$\lambda = \frac{\sigma^2_w + \sigma^2_{u,\mathcal{T}}}{\sigma^2_w + \sigma^2_{u,\mathcal{T}} + \sigma^2_z}.$$

The denominator can simply be estimated as $\widehat{Var}_i\{\widehat{Y}_{i,\mathcal{T}}\}$.

Sampling Variance Estimation. Without loss of generality, assume that T is even and let $S_o = \{1, 3, \ldots, T-1\}$ and $S_e = \{2, 4, \ldots, T\}$. Consider $\widehat{Y}_{i,S_o}$ and $\widehat{Y}_{i,S_e}$, which can be written as

$$\begin{cases} \widehat{Y}_{i,S_o} &= Z_i + W_{i,\Omega} + U_{i,S_o}; \\ \widehat{Y}_{i,S_e} &= Z_i + W_{i,\Omega} + U_{i,S_e} \end{cases}$$

If sampling variance is inversely proportional to the number of observations in the sample (which follows by the central limit theorem for any summary statistic that

can be written as a mean), then U_{i,S_o} and U_{i,S_e} each have variance $2\sigma^2_{u,S}$, since S_o and S_e each contain half the number of observations as S. Observe that

$$\begin{aligned} Var_i\{\widehat{Y}_{i,S_o} - \widehat{Y}_{i,S_e}\} &= Var_i\{U_{i,S_o} - U_{i,S_e}\} \\ &= Var_i\{U_{i,S_o}\} + Var_i\{U_{i,S_e}\} \\ &= 4\sigma^2_{u,S}. \end{aligned}$$

Therefore, the sampling variance can be estimated as $\hat{\sigma}^2_{u,S} = \frac{1}{4}\widehat{Var}_i\{\widehat{Y}_{i,S_o} - \widehat{Y}_{i,S_e}\}$. We note that a bootstrap approach could also be used to estimate the sampling variance with greater efficiency. However, when enough subjects are available, the proposed subsampling approach can also result in efficient estimation. Furthermore, the proposed approach is less computationally demanding.

Signal Variance Estimation. Let $S_1 = \{1, \ldots, T/2\}$ and $S_2 = \{T/2+1, \ldots, T\}$, and consider $\widehat{Y}_{i,S_1}$ and $\widehat{Y}_{i,S_2}$, which can be written as

$$\begin{cases} \widehat{Y}_{i,S_1} &= Z_i + W_{i,\Omega_1} + U_{i,S_1}; \\ \widehat{Y}_{i,S_2} &= Z_i + W_{i,\Omega_2} + U_{i,S_2} \end{cases},$$

where $\Omega_1 = [1, T/2]$ and $\Omega_2 = (T/2, T]$. Since Ω_1 and Ω_2 are non-overlapping, by assumption W_{i,Ω_1} and W_{i,Ω_2} are independent. Observe that

$$
\begin{aligned}
Var_i\{\widehat{Y}_{i,S_1} - \widehat{Y}_{i,S_2}\} &= Var_i\{(W_{i,\Omega_1} + U_{i,S_1}) - (W_{i,\Omega_2} + U_{i,S_2})\} \\
&= Var_i\{W_{i,\Omega_1}\} + Var_i\{W_{i,\Omega_2}\} + Var_i\{U_{i,S_1}\} + Var_i\{U_{i,S_2}\} \\
&= 2\sigma_w^2 + 4\sigma_{u,S}^2.
\end{aligned}
$$

$$
Var_i\{\widehat{Y}_{i,S_1} - \widehat{Y}_{i,S_2}\} \quad = Var_i\{(W_{i,\Omega_1} + U_{i,S_1}) - (W_{i,\Omega_2} + U_{i,S_2})\} \quad = Var_i\{W_{i,\Omega_1}\} + Var_i\{W_{i,\Omega_2}\} \quad + V
$$

Therefore, the signal variance can be estimated as $\hat{\sigma}_w^2 = \frac{1}{2}\widehat{Var}_i\{\widehat{Y}_{i,S_1} - \widehat{Y}_{i,S_2}\} - 2\hat{\sigma}_{u,S}^2$.

We also assess the performance of an "oracle" shrinkage estimator, which uses both visits from each subject to estimate the variance components. While this is not realistic (since, again, if multiple visits are available they would be combined into a single, improved estimator), it provides an upper bound on the performance of shrinkage estimators, since it is based on the best—if realistically unattainable—estimate of within-subject variance. In the continuation we denote this estimate $\tilde{W}_{ij}^{*(\ell)}(q, q')$. We refer to the proposed shrinkage methods for single-session fMRI data as "single-session shrinkage" or simply "shrinkage".

4.2.2 Reliability of connectivity estimates

We now describe the methods used to assess reliability of the raw and shrinkage estimates of connectivity for each subject. We are primarily interested in *intersession reliability* but we also assess *end-point reliability* to illustrate the bias inherent in this approach.

As illustrated in Figure 4.2a, in order to assess intersession reliability of the raw and shrinkage connectivity estimates for each subject, we compare the estimates produced using the first ℓ volumes of visit 1 to the raw connectivity estimate produced using all $L = 2400$ volumes of visit 2. As shown in Figure 4.2b, to assess end-point reliability for each subject, we compare the raw and shrinkage connectivity estimates produced using the first ℓ volumes of the first visit to the raw estimate produced using all L volumes of the same visit.

The metric we use to quantify reliability is absolute percent error. Specifically, the intersession reliability of the raw estimate of connectivity between regions q and q' for subject i and scan length ℓ is the absolute value of $\{W_{i1}^{(\ell)}(q,q') - W_{i2}^{(L)}(q,q')\} \Big/ W_{i2}^{(L)}(q,q')$, while the intersession reliability of the corresponding shrinkage estimate is the absolute value of $\{\tilde{W}_{i1}^{(\ell)}(q,q') - W_{i2}^{(L)}(q,q')\} \Big/ W_{i2}^{(L)}(q,q')$. End-point reliability of raw and shrinkage estimates is computed in a similar way by plugging in $W_{i1}^{(L)}(q,q')$ for $W_{i2}^{(L)}(q,q')$.

As illustrated in Figure 4.3, we summarize reliability over subjects at three different resolutions: omnibus, seed-level and edge-level. This organization provides both

a high-level view of how reliability changes with additional scan duration and the use of shrinkage estimates, and a detailed view of how reliability varies across different pairs of regions and how scan duration and shrinkage affect reliability of specific connections. We first compute the median reliability across all subjects for each edge (or pair of regions). The result is *edge-level reliability*, illustrated in the bottom panel of Figure 4.3. Edge-level reliability can be visualized as a set of images, each showing the reliability of connectivity between a single seed and all other regions in the brain. As visualization of all regions and model orders is impractical, we select for visualization four seed regions lying within well-known resting state networks, including the visual cortex, the somatomotor cortex, the default mode network (DMN), and the basal ganglia (Figure 4.4). Next, we compute *seed-level* reliability by treating each region as a seed and computing the median edge-level reliability of connectivity with all other regions in the brain. Illustrated in the middle panel of Figure 4.3, seed-level reliability can be visualized as a single image, illustrating for each region the overall reliability of connectivity with every other region in the brain. Finally, we compute *omnibus* reliability as the median edge-level reliability across all unique pairs of regions, resulting in a single scalar summary of reliability as illustrated in the top panel of Figure 4.3.

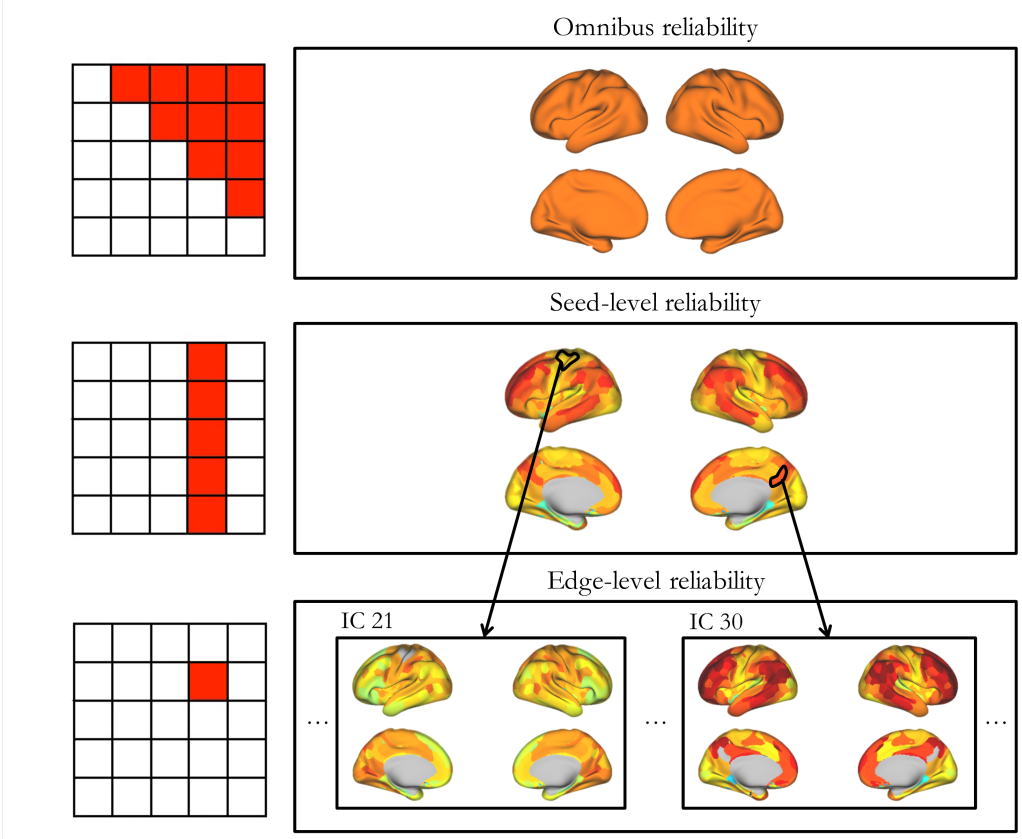


Figure 4.3: We summarize reliability at three different resolutions: omnibus, seed-level and edge-level. *Omnibus reliability* is computed as the median edge-level reliability across all unique pairs of regions, resulting in a single scalar summary measure of reliability. *Seed-level reliability* is computed as the median edge-level reliability within each seed, resulting in a single map of reliability. *Edge-level reliability* is computed as the median reliability across subjects at each edge, resulting in a map of reliability for each seed region.

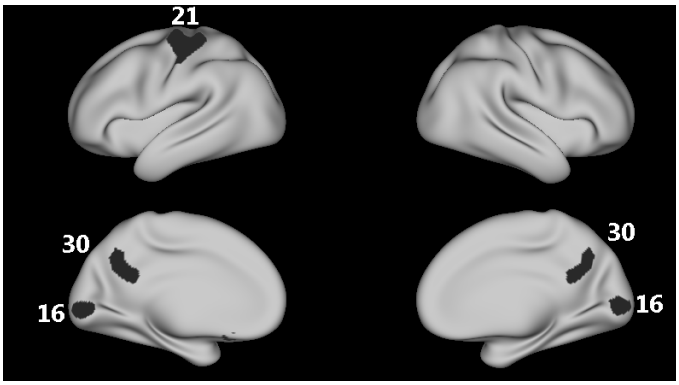


Figure 4.4: Three selected seed regions selected from model order 300, lying respectively in the visual cortex (IC 16), the somatomotor cortex (IC 21), and the DMN (IC 30). The visual seed is located in the bilateral lingual gyrus; the somatomotor seed is located in the left dorsolateral pre- and post-central gyri; the DMN seed is located in the posterior cingulate cortex (PCC).

4.3 Results

Figure 4.5 compares the omnibus intersession and end-point reliability measures of the raw connectivity matrix estimates as a function of scan length and model order. Here smaller values of percent error signify greater reliability. The two measures of reliability paint very different pictures of the accuracy of connectivity estimates at each scan length, with end-point reliability underestimating the true intersession error of connectivity estimates by 50 percent or more. This bias increases as scan length ℓ approaches $L = 2400$, since a subset of the data used to compute the reference $W_{i1}^{(L)}(q, q')$ is used to compute the estimate $W_{i1}^{(\ell)}(q, q')$, with the two quantities reaching equality at $\ell = L$. For this reason, end-point reliability is guaranteed to converge artificially to "perfect" reliability. Furthermore, Figure 4.5 suggests that the end-point approach tends to significantly overestimate reliability even for short

relative scan length, where ℓ is much less than L.

Figure 4.5 also illustrates the effect of scan length and ICA model order on reliability. Unsurprisingly, greater scan length results in increased intersession reliability. Specifically, intersession error decreases by approximately 30% as scan length increases from 300 (3.6 minutes) to 2400 (28.8 minutes). This suggests that while observing the functional connectivity of a subject over a longer period of time in a given session results in a connectivity estimate that is closer to the long-term average for that subject, session-to-session differences in rsFC limit the reproducibility of rsFC estimates produced from a single scanning session.

Finally, Figure 4.5 also shows that reliability of rsFC between a few larger regions (e.g. model order 25) tends to be greater than reliability of rsFC between many smaller regions (e.g. model order 300). This may be counter-intuitive, as smaller regions might be expected to have more coherent signals than larger regions and hence result in better estimates of rsFC. There are a number of possible drivers of the observed, opposite effect. First, the regions were defined using group ICA, and smaller group-level regions may be less spatially accurate than larger regions for individual subjects. For example, the entire somatomotor region may be spatially similar across subjects, while its subregions may exhibit greater subject-to-subject differences. There may also be visit-to-visit deviations in the true spatial location of small functional ROIs due to errors in registration or normalization. Second, the difficulty of model identification in ICA increases with the model order, and therefore

there may be more error associated with estimation of a greater number of ICs. Finally, the ICA time courses are essentially a weighted average across voxels, and averaging a greater number of voxels will tend to in reduced noise levels, and hence less noisy estimates of rsFC.

Figure 4.6 compares the omnibus intersession reliability of raw and shrinkage estimates of rsFC (bottom panel), as well as the degree of shrinkage (top panel), as a function of scan length and model order. Using oracle shrinkage, the degree of shrinkage decreases with increasing scan length, indicating greater weighting towards subject-level estimates as those estimates become more reliable. However, using single-session shrinkage, the degree of shrinkage exhibits an initial decrease followed by an increase after $\ell = 1200$. This may be due to the fact that the direction of phase encoding switches midway through acquisition of each session. Since differences in the phase encoding direction introduce an artificial source of variability, this results in overestimation of the true within-subject (intersession) variance based on single-session data, thus inflating the degree of shrinkage. However, this is likely an artifact of the unique HCP acquisition protocol, rather than a failing of the proposed single-session shrinkage methods.

Figure 4.6 clearly illustrates that both single-session and oracle shrinkage estimates exhibit greater intersession reliability than raw estimates across all model orders and scan lengths. Notably, shrinkage estimates produced using short scans ($\ell = 300$, 3.6 minutes) display similar reliability to raw estimates produced using much longer

scans ($\ell = 2400$, 28.8 minutes). Somewhat surprisingly, oracle shrinkage estimators only marginally outperform single-session shrinkage estimators, which suggests that the benefits of shrinkage are somewhat robust to differences in the degree of shrinkage. The degree of shrinkage towards the group mean using oracle shrinkage is fairly high (approximately 0.5 or greater), and undershrinkage appears to be worse than overshrinkage in terms of improving reliability. For example, at model order 300 and scan length 2400, single-session shrinkage results in nearly complete shrinkage to the group mean, while oracle shrinkage results in approximately equal weighting of the subject-level estimate and group mean. However, the reliability of the two resulting shrinkage estimates is nearly identical.

Figures 4.7 to 4.10 display intersession reliability maps of raw and shrinkage estimates at model order 300 as a function of scan length.[1] Reliability is computed at the subject level and is summarized at the edge and seed levels as illustrated in Figure 4.3. Figure 4.7a shows seed-level reliability; Figures 4.8a to 4.10a show edge-level reliability for the three seed regions shown in Figure 4.4. For each seed, the quantity displayed is the absolute percent error. Part (b) of each figure also shows the percent change in intersession error after shrinkage, relative to the raw estimates; negative values indicate improved reliability.

Figure 4.7 illustrates that raw and shrinkage estimates exhibit improved reliability with increased scan length. Both single-session shrinkage and oracle shrinkage esti-

[1]For Figures 4.7 to 4.14, the subcortical and left-hemispheric surface grayordinates are not displayed but show similar trends.

mates exhibit greater reliability than raw maps for all scan lengths and seed regions. Oracle shrinkage estimators only marginally outperform single-session shrinkage estimators; therefore, only the results of oracle shrinkage are displayed in the remainder of this document. Improvement due to shrinkage is greatest for shorter scans, with approximately 30-40% decrease in error in most regions at $\ell = 300$. However, improvement is still significant for longer scans, with approximately 10-20% decrease in error in most regions at $\ell = 2400$.

Figures 4.8 to 4.10 illustrate that, for a given seed region, edge-level reliability varies dramatically across connections. In general, within-network connections exhibit greater reliability than across-network connections, with proximal or contralateral connections within the motor network (see Figure 4.9) and the DMN (see Figure 4.10) exhibiting the greatest reliability. For all three regions, shrinkage results in improved reliability for the vast majority of connections.

Figures 4.11 to 4.14 display maps of within-subject variance, between-subject variance, and degree of shrinkage of connectivity estimates at model order 300 as a function of scan length. Both variance components and the degree of shrinkage are population-level parameters, so for each quantity there is one value for each edge or connection. For each seed, Figure 4.11 summarizes the variance components and degree of shrinkage as the median across all connections; Figures 4.12 to 4.14 show the variance components and degree of shrinkage for each connection with the three seed regions shown in Figure 4.4.

Figures 4.11 to 4.14 illustrate that as scan length increases, within-subject variance decreases as subject-level estimates of rsFC become more accurate, while between-subject variance remains similar across different scan lengths. Furthermore, there are clear spatial patterns of within-subject and between-subject variance, which are most apparent at the edge level. For example, the highest between-subject variance is exhibited by connections within the DMN and visual networks, while moderately high between-subject variance is exhibited by connections within the motor network and between the motor and visual networks. Similar spatial patterns are also observed for within-subject variance.

As detailed in Section 4.2.1.2, the degree of shrinkage is determined by the ratio of within-subject variance to total (within-subject plus between-subject) variance. The degree of shrinkage can be seen as a measure of reliability of raw estimates of connectivity, with lower values signifying greater reliability of subject-level estimates and therefore less need for shrinkage towards the group mean. Figures 4.12 to 4.14 show that the degree of shrinkage is lowest for connections within the DMN, motor network and visual network, and for connections between the motor and visual networks. Other between-network connections tend to have a higher degree of shrinkage. Figure 4.11 shows that overall, the degree of shrinkage is lower for connections with frontal and temporal/occipital networks and higher for connections with the visual network, motor network, medial temporal lobe, and the temporal pole.

4.4 Discussion

In this paper we investigate the effect of scan length on reliability of estimates of rsFC. We further study the benefits of shrinkage to improve reliability of these estimates. We find that shrinkage improves reliability of rsFC by approximately 30-40% for short scans and 10-20% for longer scans (up to 30 minutes). Our results illustrate that while longer scans can improve reliability of rsFC, shrinkage of subject-level estimates towards the group mean also results in a significant improvement in reliability, and this benefit is not unique to shorter scans. This study also shows that the benefits of shrinkage are not limited to noisy, voxel-level fMRI data but are also observed for regional time series that have been denoised using FIX and are based on high-quality and high-resolution fMRI data collected through the Human Connectome Project.

Our results suggest that increasing scan length beyond 20 or 30 minutes is likely to yield diminishing returns in terms of improved reliability of rsFC. As seen in Figure 4.6, intersession reliability appears to improve slowly as scan duration approaches 30 minutes, and at $\ell = 2400$ (28.8 minutes), the rsFC estimates still exhibit 50-100% error, depending on the model order. This suggests that session-to-session variations in rsFC are far from negligible, which is consistent with previous findings that intrasession reliability tends to be greater than intersession reliability.[19,98,99,104] If the goal is to produce the best estimate of the long-term average rsFC of a given subject, combining data from multiple sessions, ideally occurring on different days, may result in more accurate estimates of rsFC as previously observed by Shehzad et

al. (2009)[98] and Laumann et al. (2015).[18]

Furthermore, as illustrated in Figure 4.6, shrinkage estimators of rsFC appear to be quite robust to overshrinkage (e.g. single-session shrinkage estimator at $\ell \geq 2100$). This indicates that repeated measures of rsFC for a given subject tend to be closer to the group average than to each other, and true subject-level differences in rsFC may therefore be quite subtle and difficult to estimate accurately. This again suggests that combining multiple rsfMRI scanning sessions may improve estimation of subject-level rsFC.

Our results also show that end-point reliability is a poor proxy for true intersession reliability (see Figure 4.5. To realistically assess reliability in contexts where only a single session of data is available for each subject, an alternative approach is to use split-half data. For example, for a single fMRI session of length L, reliability can be assessed by estimating rsFC using the first $L/2$ volumes of the session and the last $L/2$ volumes, and comparing the two estimates.

Acknowledgements

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at

Washington University.

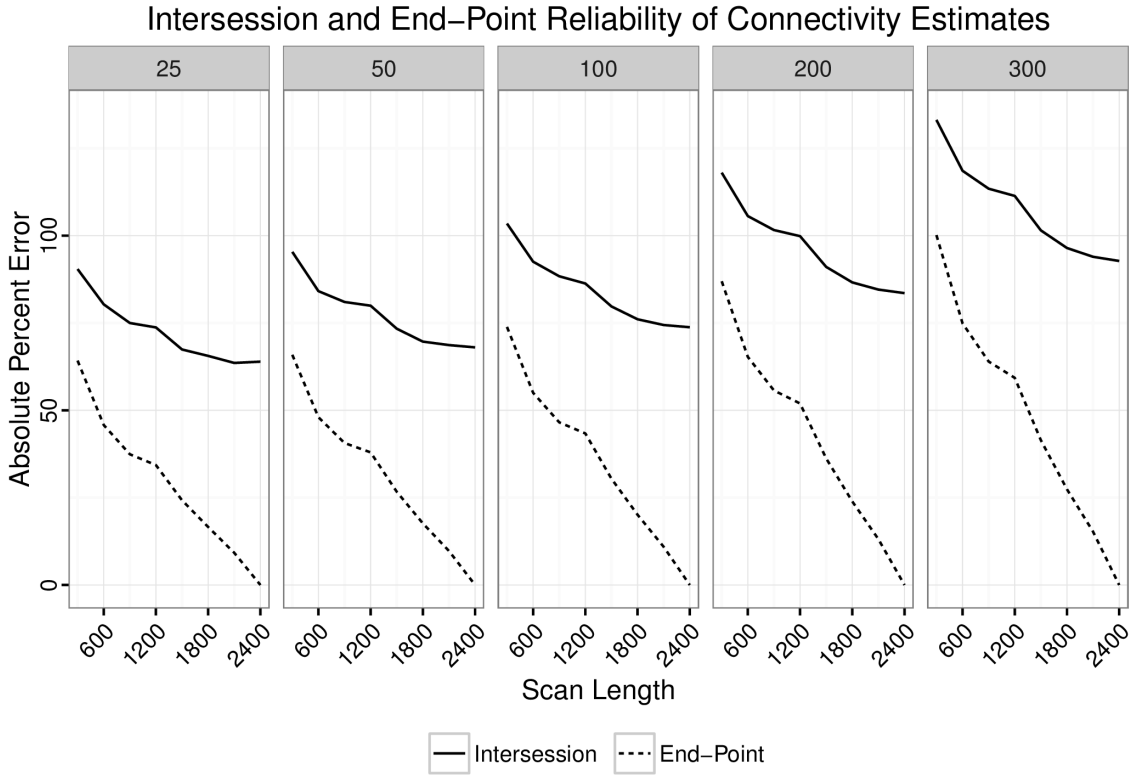


Figure 4.5: Comparison of omnibus intersession and end-point reliability of raw connectivity estimates by scan length at each model order (25, 50, 100, 200, 300). Smaller values of absolute percent error signify greater reliability. For each model order and at every scan length, end-point reliability is a poor proxy for intersession reliability, as end-point error greatly underestimates the true intersession error. Furthermore, this bias increases sharply as the scan length increases to $L = 2400$. This is because the estimate produced from $\ell < L$ volumes becomes more similar to the reference produced from L volumes, eventually reaching equality at $\ell = L$.

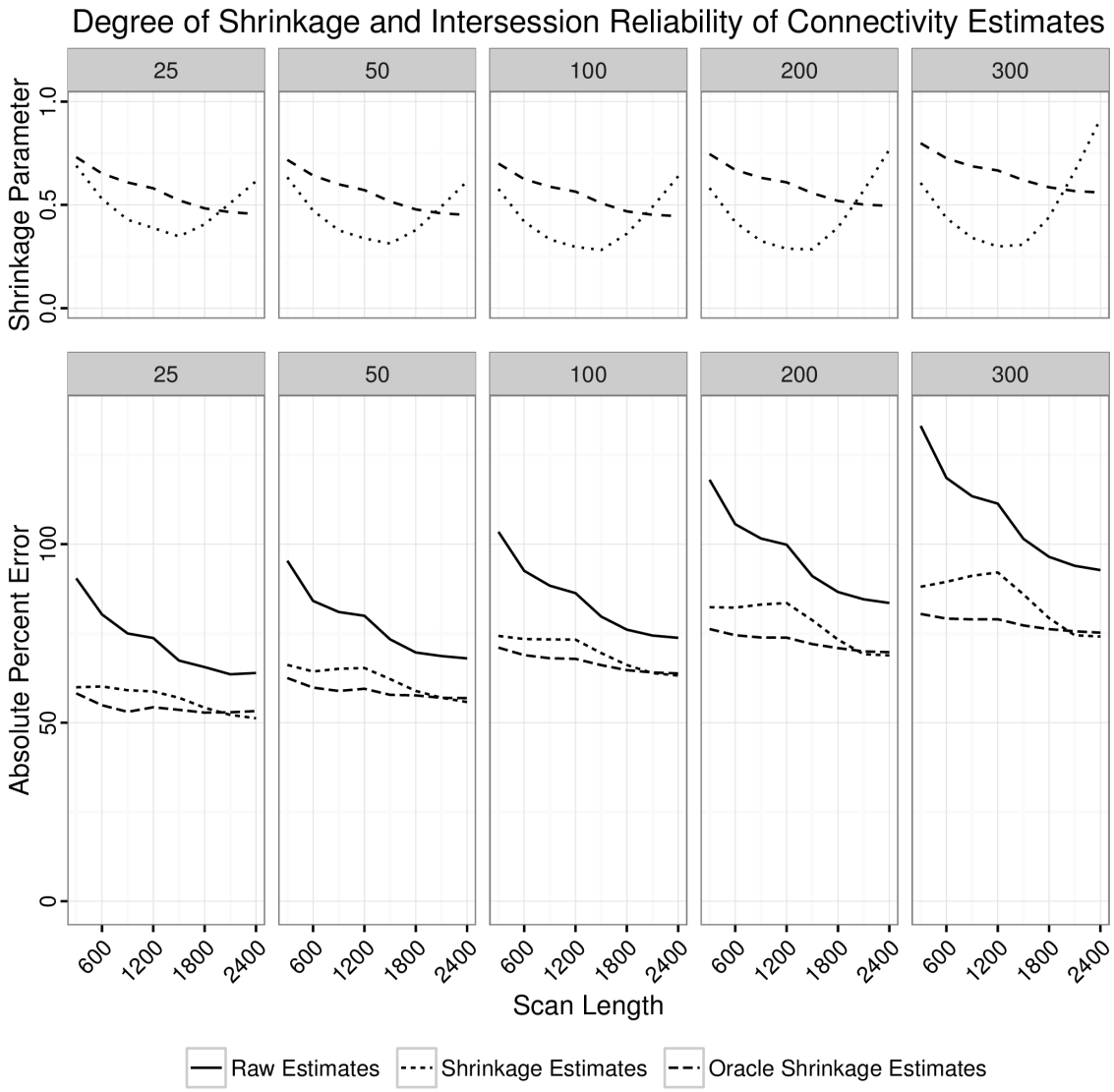
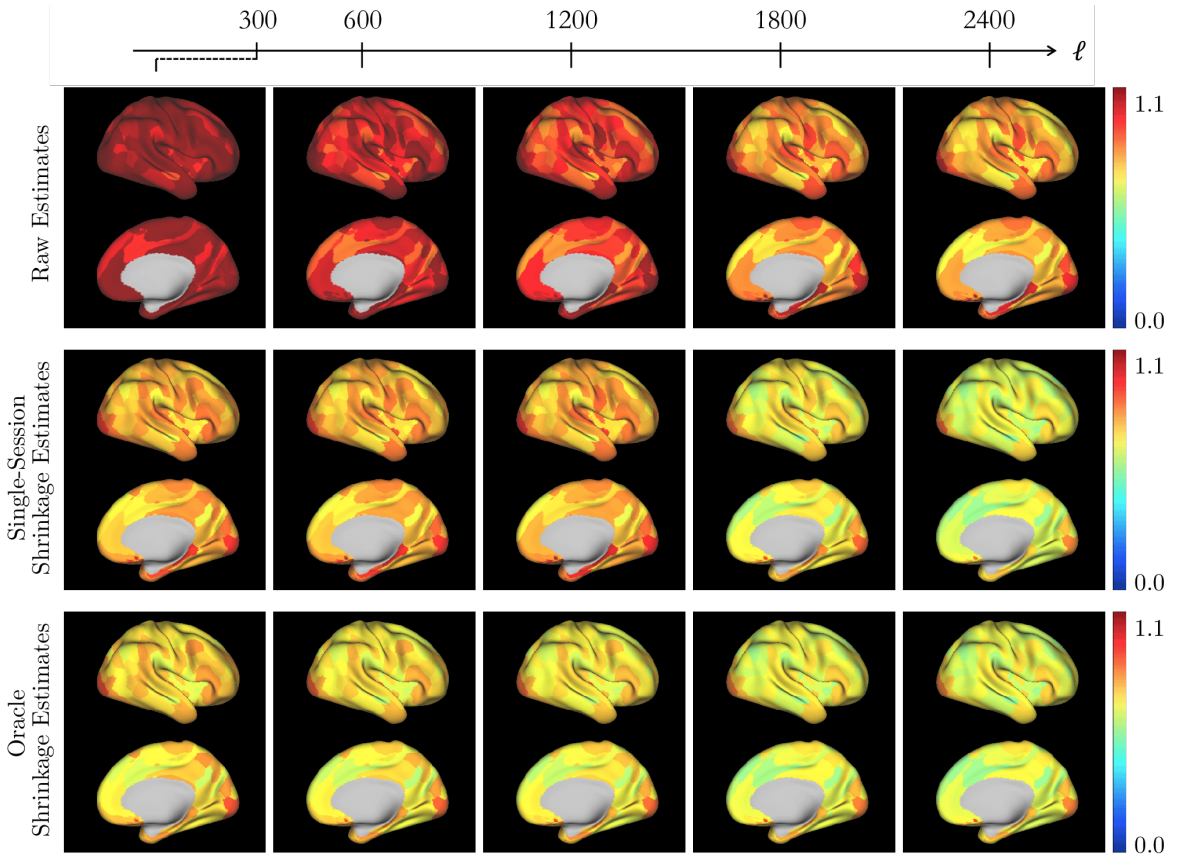
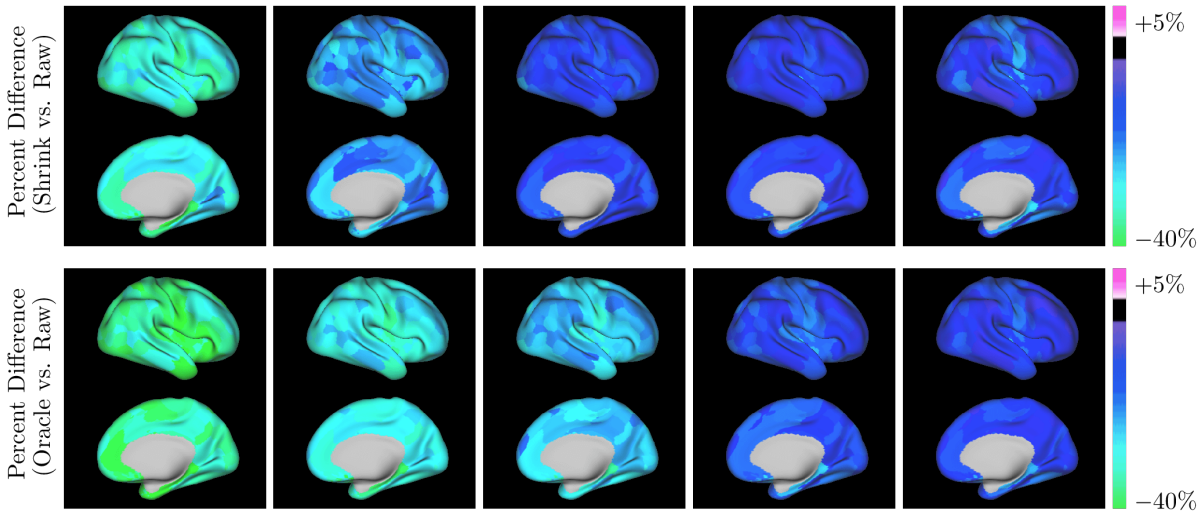


Figure 4.6: Top panel: Degree of shrinkage versus scan length for each model order (25, 50, 100, 200, 300) (median over all connections). For oracle shrinkage, the degree of shrinkage tends to decrease as scan length increases. However, for single-session shrinkage, the degree of shrinkage exhibits an initial decrease followed by an increase. This is likely artifact of the unique HCP acquisition protocol, in which the phase encoding method changes midway through each session, leading to an inflation of the within-subject variance estimated from a single session. **Bottom panel:** Comparison of omnibus intersession reliability of raw and shrinkage connectivity estimates by scan length for each model order. Smaller values of absolute percent error signify greater reliability. Both single-session and oracle shrinkage estimates exhibit greater intersession reliabilty than raw estimates across all model orders and scan lengths.

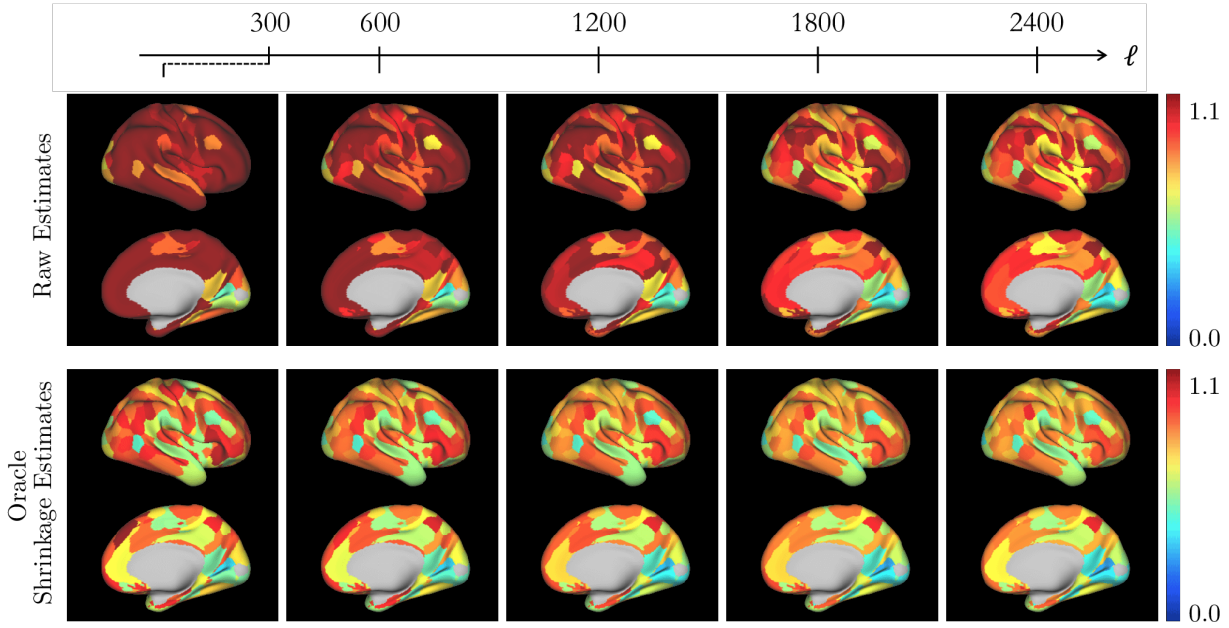


(a) Seed-level reliability. Absolute percent error (APE) of raw, shrinkage and oracle shrinkage connectivity estimates at model order 300. For each seed, the value displayed is the median APE over subjects and connections. Smaller values indicate greater reliability.

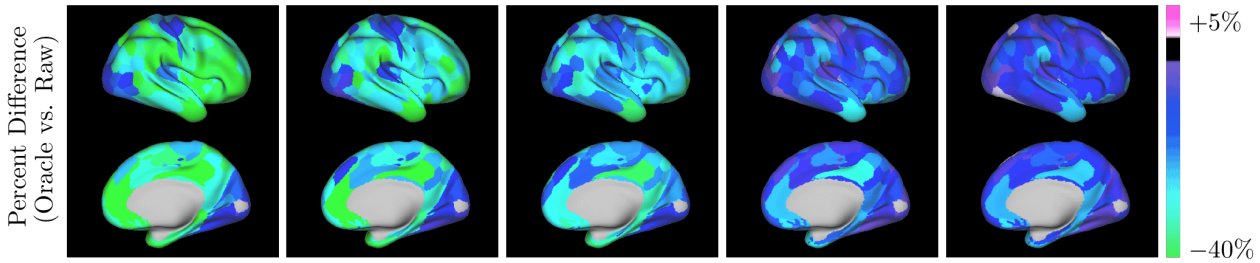


(b) Change in reliability due to shrinkage. The value shown is the percent change in APE after shrinkage, relative to that of the raw estimates. Negative values indicate improved reliability. The top row shows the effect of single-session shrinkage; the bottom row shows the effect of oracle shrinkage.

Figure 4.7: Seed-level reliability.

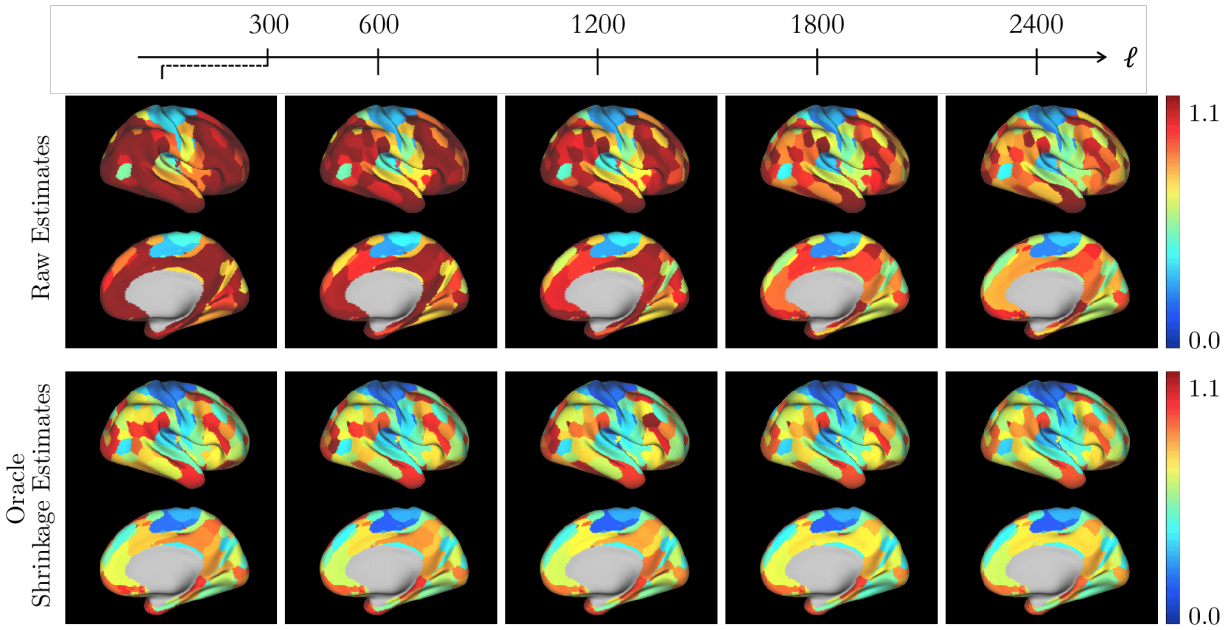


(a) Edge-level reliability, visual seed. Absolute percent error (APE) of raw and oracle shrinkage connectivity estimates between the visual seed shown in Figure 4 and each of the 299 other ICA regions. For each region, the value displayed is the median APE over subjects. Smaller values indicate greater reliability.

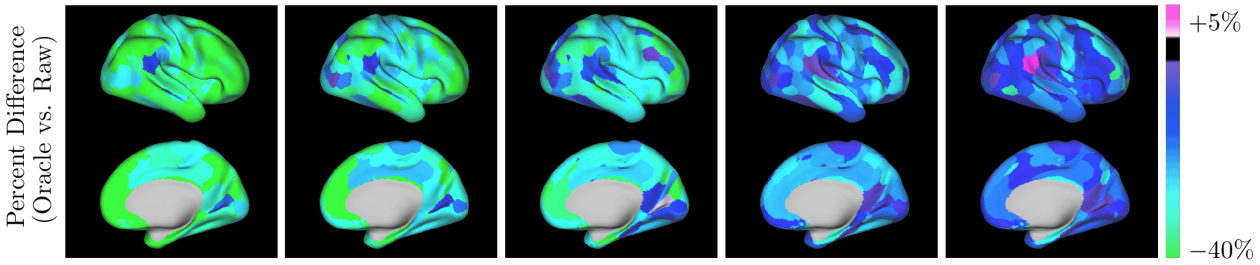


(b) Change in reliability due to shrinkage. The value shown is the percent change in APE after shrinkage, relative to that of the raw estimates. Negative values indicate improved reliability.

Figure 4.8: Edge-level reliability, visual seed.

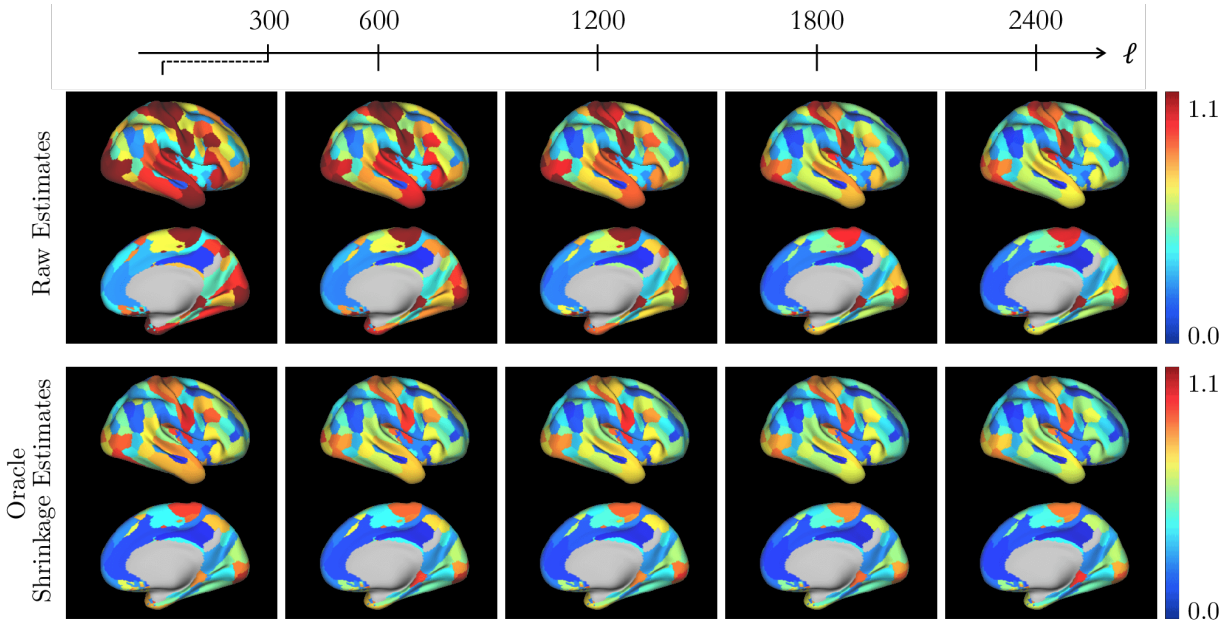


(a) Edge-level reliability, somatomotor seed. Absolute percent error (APE) of raw, shrinkage and oracle shrinkage connectivity estimates between the somatomotor seed shown in Figure 4 and each of the 299 other ICA regions. For each region, the value displayed is the median APE over subjects. Smaller values indicate greater reliability.

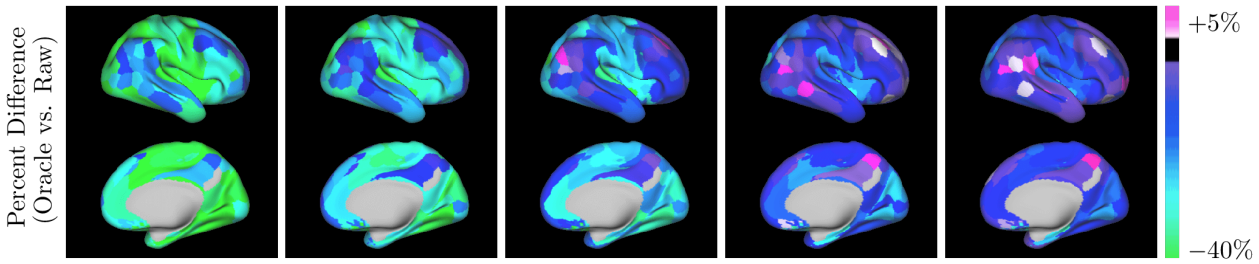


(b) Change in reliability due to shrinkage. The value shown is the percent change in APE after shrinkage, relative to that of the raw estimates. Negative values indicate improved reliability.

Figure 4.9: Edge-level reliability, somatomotor seed.



(a) Edge-level reliability, DMN seed. Absolute percent error (APE) of raw, shrinkage and oracle shrinkage connectivity estimates between the DMN seed shown in Figure 4 and each of the 299 other ICA regions. For each region, the value displayed is the median APE over subjects. Smaller values indicate greater reliability.



(b) Change in reliability due to shrinkage. The value shown is the percent change in APE after shrinkage, relative to that of the raw estimates. Negative values indicate improved reliability.

Figure 4.10: Edge-level reliability, DMN seed.

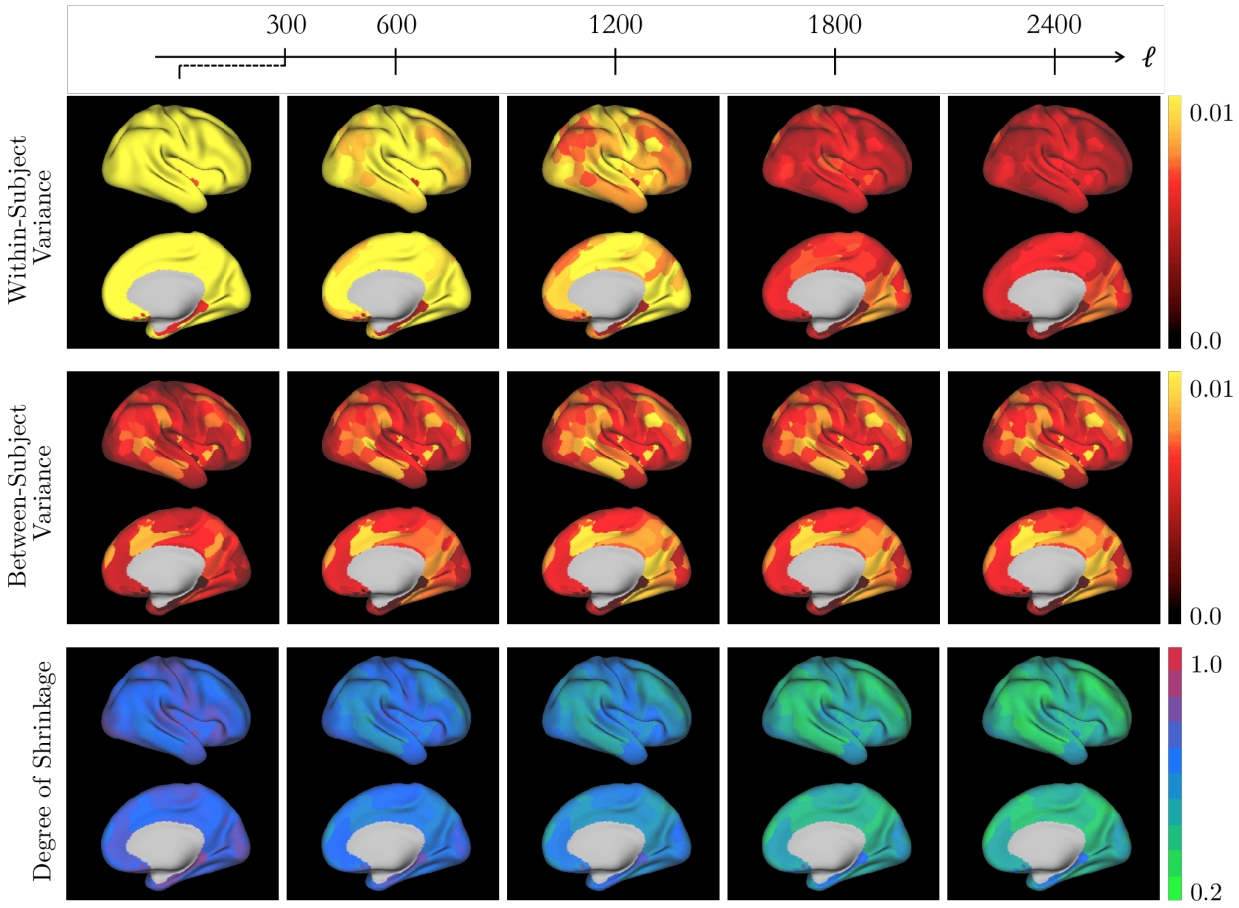


Figure 4.11: Seed-level variance components and degree of shrinkage. For each quantity, the median value over all connections with a given seed is displayed. As scan length increases, between-subject variance stays relatively constant, while within-subject variance and hence the degree of shrinkage decreases.

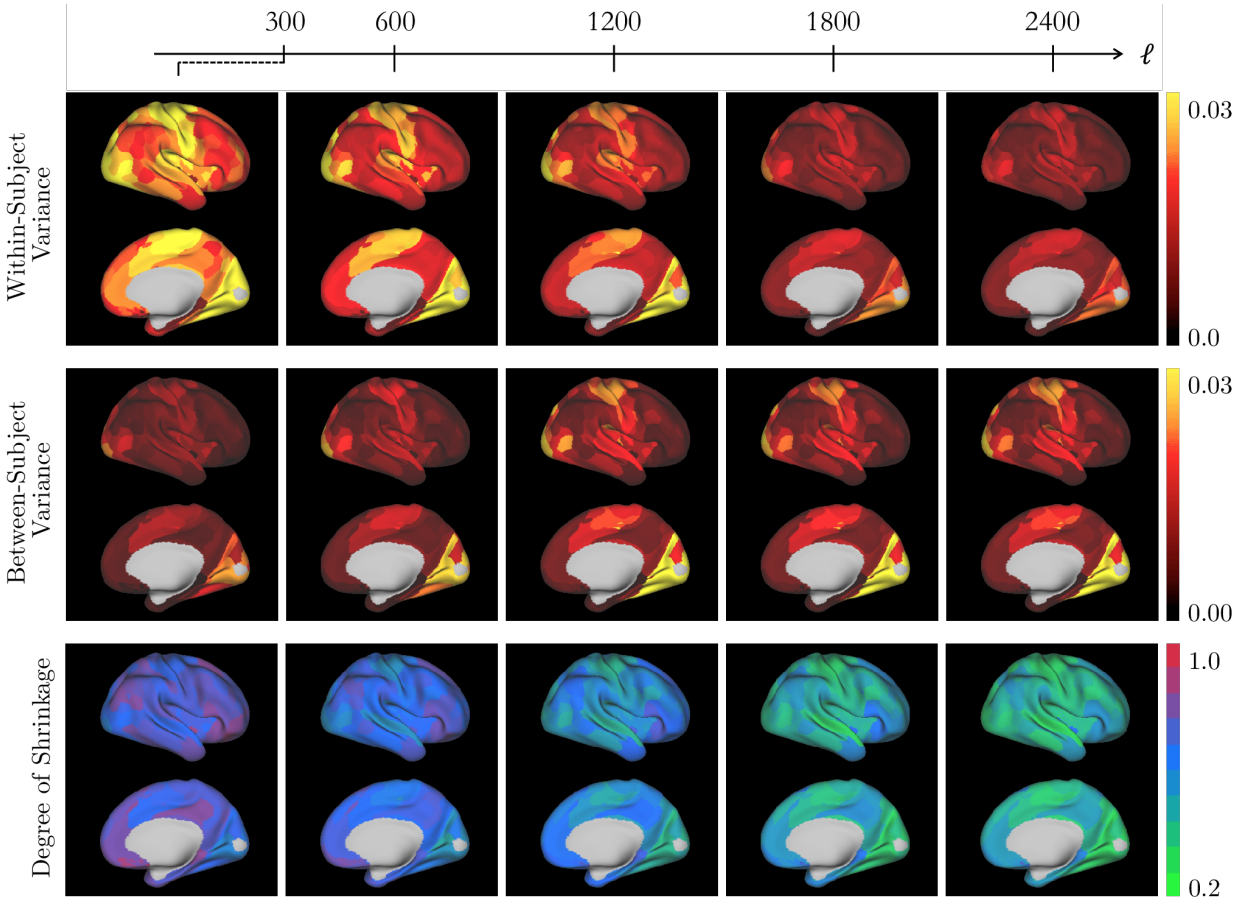


Figure 4.12: Edge-level variance components and degree of shrinkage, visual seed.

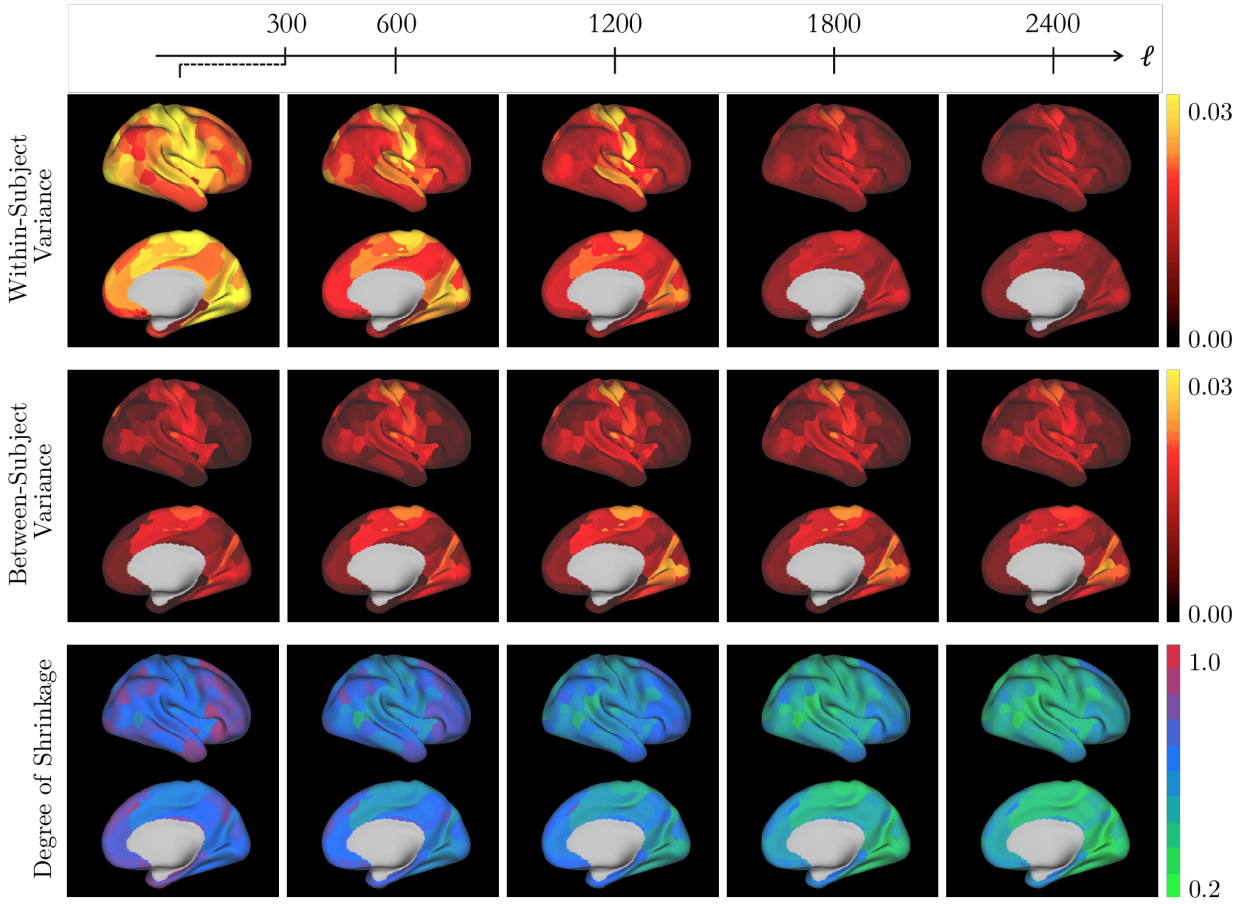


Figure 4.13: Edge-level variance components and degree of shrinkage, somatomotor seed.

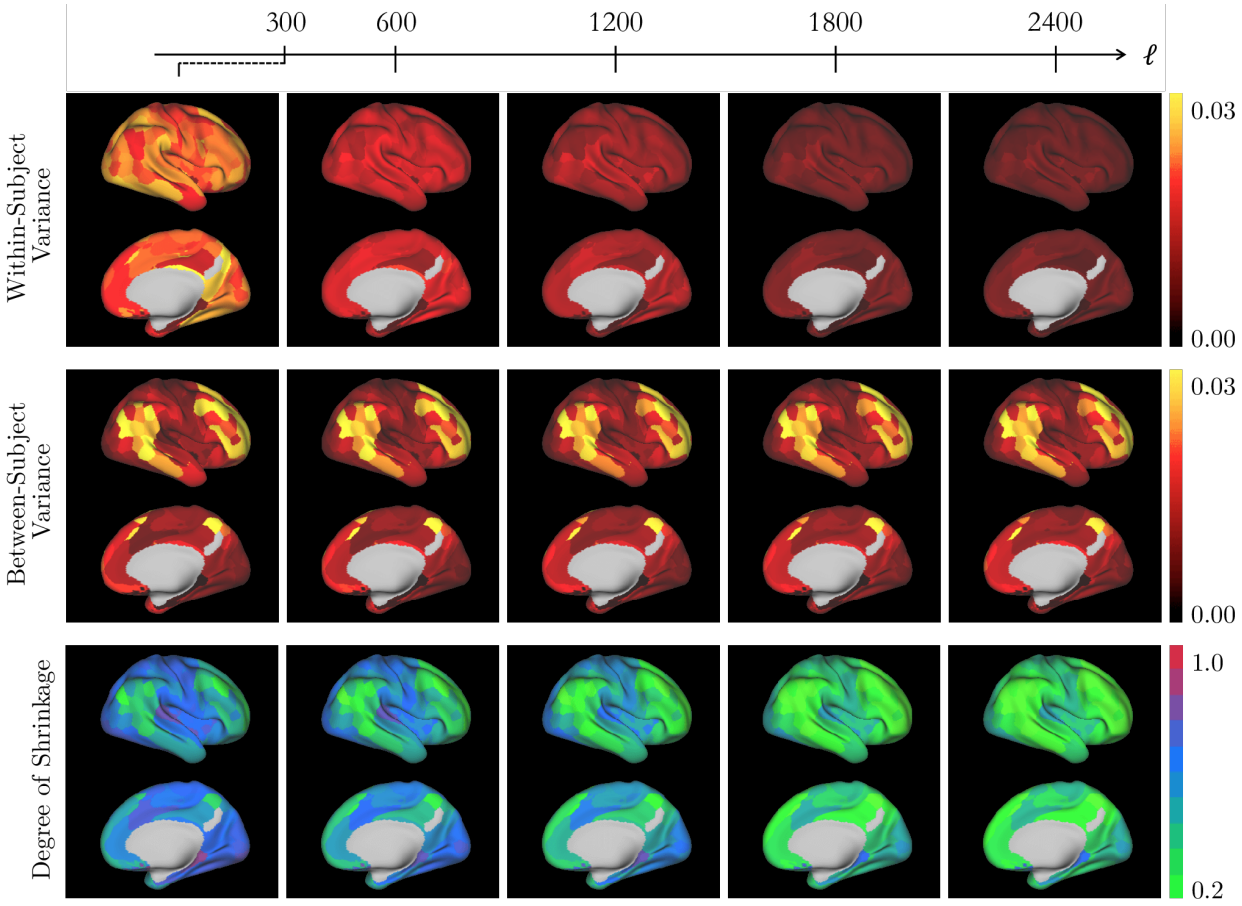


Figure 4.14: Edge-level variance components and degree of shrinkage, DMN seed.

Chapter 5

Discussion

In Chapters 2 and 4, I proposed shrinkage methods for estimation of subject-level functional connectivity based on resting-state fMRI data. The empirical Bayes methods employed rely on estimation of both between-subject and within-subject variance of the estimators on which shrinkage is to be performed. In Chapter 2, I introduced the idea of estimating within-subject (inter-session) variance from single-session data. Specifically, I proposed a psuedo-scan-rescan technique, combined with an adjustment factor estimated from the data for varying scan lengths. However, several limitations of the method and analysis led to the extension presented in Chapter 4. Specifically, the empirically-estimated adjustment factor may become less appropriate as conditions vary, such as the specific connections being considered and the quality of the fMRI data. In Chapter 4, I therefore propose a new technique to estimate within-subject variance from single-session data, in which two different sources of

within-subject variation are identified: sampling variance and variance of the signal over time. This new technique is more general and is designed to work in a variety of cases, including settings where sampling variance dominates (e.g. voxel-level time courses produced from short, noisy fMRI sessions) and settings where signal variance dominates (e.g. region-level time courses produced from long, denoised fMRI sessions).

Furthermore, in Chapter 2 reliability of shrinkage estimates was assessed using a single fMRI study containing relatively short scans from a small population of subjects, and only connections within the motor cortex were considered. The benefits of shrinkage for different regions, datasets, and scan lengths remained to be determined. To this end, in Chapter 4 I employed data from the Human Connectome Project, where for each of 461 subjects nearly 60 minutes of resting-state fMRI data was acquired at a TR of 0.72 seconds, resulting in an unusually high number of observations for each subject. Furthermore, the acquisition and processing methods employed in the HCP were designed to optimize data quality and reduce sources of noise, and region-level connectivity rather than voxel-level connectivity was considered. Therefore, the raw connectivity estimates considered in Chapter 4 would be expected to exhibit much greater reliability than those considered in Chapter 2. However, the results indicate that shrinkage remains strongly beneficial to the reliability of connectivity estimates across different scan lengths, even when using the neary optimal fMRI data of the HCP.

In Chapter 3, I considered another source of error in analyses of fMRI data, namely artifacts, which are common and may arise from a number of sources, including subject motion, scanner instabilities, and processing errors. In order to identify artifacts, I proposed a novel high-dimensional outlier detection method. Artifact detection and removal may be performed as a pre-processing step, and can therefore be combined with shrinkage methods to maximize reliability of connectivity estimates. However, artifact detection can also be used on fMRI data prior to other types of analysis, including ICA or task activation studies. The proposed methods can also be tailored to many different contexts, including other types of functional neuroimaging, group analyses of structural MR images, and other types of high-dimensional data not related to imaging. As the availability of "big data" continues to grow, quality control will become increasingly important, and outlier detection methods designed for high-dimensional data such as those proposed in Chapter 3 will continue to be an important topic of research. This is especially true in the context of fMRI data, as publicly available datasets such as the ABIDE, ADHD-200, 1000 FCP, ADNI and HCP become increasingly large, more widely available and more highly utilized. As the BRAIN initiative exemplifies, understanding of brain function and organization is one of the most important topics of scientific research today, and these datasets, combined with appropriate quality-control methods, are a major component of that effort.

Bibliography

[1] T. R. Insel, S. C. Landis, F. S. Collins *et al.*, "The NIH brain initiative," *Science*, vol. 340, no. 6133, pp. 687–688, 2013.

[2] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional magnetic resonance imaging*. Sinauer Associates Sunderland, 2004, vol. 1.

[3] M. A. Lindquist *et al.*, "The statistical analysis of fMRI data," *Statistical Science*, vol. 23, no. 4, pp. 439–464, 2008.

[4] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[5] M. N. Hallquist, K. Hwang, and B. Luna, "The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity," *NeuroImage*, vol. 82, pp. 208–225, 2013.

BIBLIOGRAPHY

[6] A.-. Consortium *et al.*, "The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience," *Frontiers in systems neuroscience*, vol. 6, 2012.

[7] A. Di Martino, F. Castellanos, J. Anderson, K. Alaerts, M. Assaf, D. Deen, D. Fair, L. Gallagher, J. Lainhart, C. Lord *et al.*, "The autism brain imaging data exchange (abide) consortium: open sharing of autism resting state fMRI data," 2012.

[8] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The WU-Minn human connectome project: an overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[9] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe *et al.*, "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4734–4739, 2010.

[10] E. Sharer, D. Crocetti, J. Muschelli, A. D. Barber, M. B. Nebel, B. S. Caffo, J. J. Pekar, and S. H. Mostofsky, "Neural correlates of visuomotor learning in autism," *Journal of child neurology*, p. 0883073815600869, 2015.

[11] M. P. Van Den Heuvel and H. E. H. Pol, "Exploring the brain network: a review on resting-state fMRI functional connectivity," *European Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

BIBLIOGRAPHY

[12] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magnetic resonance in medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[13] C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith, "Investigations into resting-state connectivity using independent component analysis," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 360, no. 1457, pp. 1001–1013, 2005.

[14] K. J. Friston, "Functional and effective connectivity in neuroimaging: a synthesis," *Human brain mapping*, vol. 2, no. 1-2, pp. 56–78, 1994.

[15] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J.-B. Poline, "Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets," *Human brain mapping*, vol. 27, no. 8, pp. 678–693, 2006.

[16] A. L. Cohen, D. A. Fair, N. U. Dosenbach, F. M. Miezin, D. Dierker, D. C. Van Essen, B. L. Schlaggar, and S. E. Petersen, "Defining functional areas in individual human brains using resting functional connectivity MRI," *NeuroImage*, vol. 41, no. 1, p. 45, 2008.

[17] G. S. Wig, T. O. Laumann, A. L. Cohen, J. D. Power, S. M. Nelson, M. F. Glasser, F. M. Miezin, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen, "Parcellating an individual subject's cortical and subcortical brain structures using

snowball sampling of resting-state correlations," *Cerebral Cortex*, p. bht056, 2013.

[18] T. O. Laumann, E. M. Gordon, B. Adeyemo, A. Z. Snyder, S. J. Joo, M.-Y. Chen, A. W. Gilmore, K. B. McDermott, S. M. Nelson, N. U. Dosenbach *et al.*, "Functional system and areal organization of a highly sampled individual human brain," *Neuron*, vol. 87, no. 3, pp. 657–670, 2015.

[19] R. M. Birn, E. K. Molloy, R. Patriat, T. Parker, T. B. Meier, G. R. Kirk, V. A. Nair, M. E. Meyerand, and V. Prabhakaran, "The effect of scan length on the reliability of resting-state fMRI connectivity estimates," *NeuroImage*, vol. 83, pp. 550–558, 2013.

[20] K. Zilles and K. Amunts, "Centenary of brodmann's map —- conception and fate," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 139–145, 2010.

[21] P. Bellec, P. Rosa-Neto, O. C. Lyttelton, H. Benali, and A. C. Evans, "Multi-level bootstrap analysis of stable clusters in resting-state fMRI," *NeuroImage*, vol. 51, no. 3, pp. 1126–1139, 2010.

[22] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Human brain mapping*, vol. 33, no. 8, pp. 1914–1928, 2011.

[23] B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari,

M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni *et al.*, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of neurophysiology*, vol. 106, no. 3, pp. 1125–1165, 2011.

[24] M. A. Lindquist and A. Mejia, "Zen and the art of multiple comparisons," *Psychosomatic Medicine*, 2014, in press.

[25] O. Sporns, G. Tononi, and R. Kötter, "The human connectome: a structural description of the human brain," *PLoS computational biology*, vol. 1, no. 4, p. e42, 2005.

[26] R. Baumgartner, G. Scarth, C. Teichtmeister, R. Somorjai, and E. Moser, "Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. part i: Reproducibility," *Journal of Magnetic Resonance Imaging*, vol. 7, no. 6, pp. 1094–1101, 1997.

[27] M. J. McKeown, T. J. Sejnowski *et al.*, "Independent component analysis of fMRI data: examining the assumptions," *Human brain mapping*, vol. 6, no. 5-6, pp. 368–372, 1998.

[28] J. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. Stam, S. M. Smith, and C. Beckmann, "Consistent resting-state networks across healthy subjects," *Proceedings of the national academy of sciences*, vol. 103, no. 37, pp. 13 848–13 853, 2006.

BIBLIOGRAPHY

[29] M. De Luca, C. Beckmann, N. De Stefano, P. Matthews, S. M. Smith *et al.*, "fMRI resting state networks define distinct modes of long-distance interactions in the human brain," *NeuroImage*, vol. 29, no. 4, pp. 1359–1367, 2006.

[30] S. Ryali, T. Chen, K. Supekar, and V. Menon, "A parcellation scheme based on von Mises-Fisher distributions and Markov random fields for segmenting brain regions using resting-state fMRI," *NeuroImage*, 2012.

[31] D. Cordes, V. Haughton, J. D. Carew, K. Arfanakis, and K. Maravilla, "Hierarchical clustering to measure connectivity in fMRI resting-state data," *Magnetic resonance imaging*, vol. 20, no. 4, pp. 305–317, 2002.

[32] S. Salvador, A. Brovelli, and R. Longo, "A simple and fast technique for on-line fMRI data analysis," *Magnetic resonance imaging*, vol. 20, no. 2, pp. 207–213, 2002.

[33] T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. Van Essen, K. Ugurbil, T. E. Behrens, and S. M. Smith, "Spatially constrained hierarchical parcellation of the brain with resting-state fMRI," *NeuroImage*, vol. 76, pp. 313–324, 2013.

[34] M. B. Nebel, S. E. Joel, J. Muschelli, A. D. Barber, B. S. Caffo, J. J. Pekar, and S. H. Mostofsky, "Disruption of functional organization within the primary motor cortex in children with autism," *Human Brain Mapping*, vol. 35, no. 2, 2014.

BIBLIOGRAPHY

[35] J.-H. Kim, J.-M. Lee, H. J. Jo, S. H. Kim, J. H. Lee, S. T. Kim, S. W. Seo, R. W. Cox, D. L. Na, S. I. Kim *et al.*, "Defining functional SMA and pre-SMA subregions in human MFC using resting state fMRI: functional connectivity-based parcellation method," *NeuroImage*, vol. 49, no. 3, p. 2375, 2010.

[36] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *ICML*, vol. 2, 2002, pp. 27–34.

[37] W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 1961, 1961, pp. 361–379.

[38] B. Efron and C. Morris, "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 311–319, 1975.

[39] M. A. Lindquist and A. Gelman, "Correlations and multiple comparisons in functional imaging: a statistical perspective (commentary on Vul et al., 2009)," *Perspectives on Psychological Science*, vol. 4, no. 3, pp. 310–313, 2009.

[40] K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and bayesian inference in neuroimaging: theory," *NeuroImage*, vol. 16, no. 2, pp. 465–483, 2002.

BIBLIOGRAPHY

[41] K. Friston and W. Penny, "Posterior probability maps and SPMs," *NeuroImage*, vol. 19, no. 3, pp. 1240–1249, 2003.

[42] S.-C. Su, B. Caffo, E. Garrett-Mayer, and S. S. Bassett, "Modified test statistics by inter-voxel variance shrinkage with an application to fMRI," *Biostatistics*, vol. 10, no. 2, pp. 219–227, 2009.

[43] H. Shou, A. Eloyan, M. B. Nebel, A. Mejia, J. J. Pekar, S. Mostofsky, B. Caffo, M. A. Lindquist, and C. M. Crainiceanu, "Shrinkage prediction of seed-voxel brain connectivity using resting state fMRI," *NeuroImage*, 2014.

[44] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[45] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering analysis and an algorithm," *Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press*, vol. 14, pp. 849–856, 2001.

[46] B. A. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. L. Lim, J. A. Farrell, J. A. Bogovic, J. Hua, M. Chen, S. Jarso *et al.*, "Multi-parametric neuroimaging reproducibility: A 3-T resource study," *NeuroImage*, vol. 54, no. 4, pp. 2854–2866, 2011.

[47] Y. Behzadi, K. Restom, J. Liau, and T. T. Liu, "A component based noise cor-

rection method (CompCor) for BOLD and perfusion based fMRI," *NeuroImage*, vol. 37, no. 1, pp. 90–101, 2007.

[48] J. Muschelli, M. B. Nebel, B. S. Caffo, A. D. Barber, J. J. Pekar, and S. H. Mostofsky, "Reduction of motion-related artifacts in resting state fMRI using aCompCor," *NeuroImage*, vol. 96, pp. 22–35, 2014.

[49] W. Penfield and E. Boldrey, "Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation." *Brain: A journal of neurology*, 1937.

[50] K. Oishi, A. Faria, H. Jiang, X. Li, K. Akhter, J. Zhang, J. T. Hsu, M. I. Miller, P. van Zijl, M. Albert *et al.*, "Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants," *NeuroImage*, vol. 46, no. 2, pp. 486–499, 2009.

[51] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.

[52] P. Filzmoser, R. Maronna, and M. Werner, "Outlier identification in high dimensions," *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1694–1711, 2008.

BIBLIOGRAPHY

[53] A. S. Hadi, A. Imon, and M. Werner, "Detection of outliers," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 57–70, 2009.

[54] A. D. Shieh and Y. S. Hung, "Detecting outlier samples in microarray data," *Statistical applications in genetics and molecular biology*, vol. 8, no. 1, pp. 1–24, 2009.

[55] V. Fritsch, G. Varoquaux, B. Thyreau, J.-B. Poline, and B. Thirion, "Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators," *Medical image analysis*, vol. 16, no. 7, pp. 1359–1370, 2012.

[56] K. Ro, C. Zou, Z. Wang, and G. Yin, "Outlier detection for high-dimensional data," *Biometrika*, 2015.

[57] A. Di Martino, C. Yan, Q. Li, E. Denio, F. Castellanos, K. Alaerts, J. Anderson, M. Assaf, S. Bookheimer, M. Dapretto *et al.*, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[58] T. Næs, "Leverage and influence measures for principal component regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 5, no. 2, pp. 155–168, 1989.

[59] S. P. Brooks, "Diagnostics for principal components: influence functions as diagnostic tools," *The Statistician*, pp. 483–494, 1994.

[60] S. Gao, G. Li, and D. Wang, "A new approach for detecting multivariate outliers," *Communications in Statistics: Theory and Methods*, vol. 34, no. 8, pp. 1857–1865, 2005.

[61] P. J. Rousseeuw, "Multivariate estimation with high breakdown point," *Mathematical statistics and applications*, vol. 8, pp. 283–297, 1985.

[62] J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen, "Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion," *NeuroImage*, vol. 59, no. 3, pp. 2142–2154, 2012.

[63] T. D. Satterthwaite, M. A. Elliott, R. T. Gerraty, K. Ruparel, J. Loughead, M. E. Calkins, S. B. Eickhoff, H. Hakonarson, R. C. Gur, R. E. Gur *et al.*, "An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data," *NeuroImage*, vol. 64, pp. 240–256, 2013.

[64] J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen, "Methods to detect, characterize, and remove motion artifact in resting state fMRI," *NeuroImage*, vol. 84, pp. 320–341, 2014.

[65] K. R. Van Dijk, M. R. Sabuncu, and R. L. Buckner, "The influence of head motion on intrinsic functional connectivity MRI," *NeuroImage*, vol. 59, no. 1, pp. 431–438, 2012.

BIBLIOGRAPHY

[66] T. D. Satterthwaite, D. H. Wolf, J. Loughead, K. Ruparel, M. A. Elliott, H. Hakonarson, R. C. Gur, and R. E. Gur, "Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth," *NeuroImage*, vol. 60, no. 1, pp. 623–632, 2012.

[67] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[68] F. R. Hampel, E. M. Ronchetti, P. Rousseeuw, and W. A. Stahel, *Robust Statistics: the Approach based on Influence Functions*. New York: John Wiley, 1986.

[69] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[70] L. Guttman, "Some necessary conditions for common-factor analysis," *Psychometrika*, vol. 19, no. 2, pp. 149–161, 1954.

[71] H. F. Kaiser, "The application of electronic computers to factor analysis." *Educational and psychological measurement*, 1960.

[72] D. A. Jackson, "Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches," *Ecology*, pp. 2204–2214, 1993.

[73] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied linear statistical models*. Irwin Chicago, 1996, vol. 4.

[74] P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.

[75] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, 2011.

[76] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[77] J. Hardin and D. M. Rocke, "The distribution of robust distances," *Journal of Computational and Graphical Statistics*, vol. 14, no. 4, 2005.

[78] R. A. Maronna and R. H. Zamar, "Robust estimates of location and dispersion for high-dimensional datasets," *Technometrics*, vol. 44, no. 4, 2002.

[79] C. F. Beckmann, C. E. Mackay, N. Filippini, and S. M. Smith, "Group comparison of resting-state fMRI data using multi-subject ICA and dual regression," *NeuroImage*, vol. 47, no. Suppl 1, p. S148, 2009.

[80] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. S208–S219, 2004.

BIBLIOGRAPHY

[81] E. A. Allen, E. B. Erhardt, E. Damaraju, W. Gruner, J. M. Segall, R. F. Silva, M. Havlicek, S. Rachakonda, J. Fries, R. Kalyanam *et al.*, "A baseline for the multivariate comparison of resting-state networks," *Frontiers in Systems Neuroscience*, vol. 5, 2011.

[82] J. L. Marchini, C. Heaton, and B. D. Ripley, *fastICA: FastICA algorithms to perform ICA and projection pursuit*, 2013, R package version 1.2-0. [Online]. Available: http://CRAN.R-project.org/package=fastICA

[83] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team, *nlme: Linear and nonlinear mixed effects models*, 2014, R package version 3.1-118. [Online]. Available: http://CRAN.R-project.org/package=nlme

[84] A. E. Abbott, A. Nair, C. L. Keown, M. Datko, A. Jahedi, I. Fishman, and R.-A. Müller, "Patterns of atypical functional connectivity and behavioral links in autism differ between default, salience, and executive networks," *Cerebral Cortex*, p. bhv191, 2015.

[85] M. B. Nebel, A. Eloyan, C. A. Nettles, K. L. Sweeney, K. Ament, R. E. Ward, A. S. Choe, A. D. Barber, J. J. Pekar, and S. H. Mostofsky, "Intrinsic visual-motor synchrony correlates with social deficits in autism," *Biological psychiatry*, 2015.

[86] A. D. Barber, L. A. Jacobson, J. L. Wexler, M. B. Nebel, B. S. Caffo, J. J. Pekar, and S. H. Mostofsky, "Connectivity supporting attention in children

with attention deficit hyperactivity disorder," *NeuroImage: Clinical*, vol. 7, pp. 68–81, 2015.

[87] F. X. Castellanos, D. S. Margulies, C. Kelly, L. Q. Uddin, M. Ghaffari, A. Kirsch, D. Shaw, Z. Shehzad, A. Di Martino, B. Biswal *et al.*, "Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder," *Biological psychiatry*, vol. 63, no. 3, pp. 332–337, 2008.

[88] M. D. Greicius, B. H. Flores, V. Menon, G. H. Glover, H. B. Solvason, H. Kenna, A. L. Reiss, and A. F. Schatzberg, "Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus," *Biological psychiatry*, vol. 62, no. 5, pp. 429–437, 2007.

[89] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon, "Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 13, pp. 4637–4642, 2004.

[90] W. W. Seeley, J. M. Allman, D. A. Carlin, R. K. Crawford, M. N. Macedo, M. D. Greicius, S. J. Dearmond, and B. L. Miller, "Divergent social functioning in behavioral variant frontotemporal dementia and Alzheimer disease: reciprocal networks and neuronal evolution," *Alzheimer Disease & Associated Disorders*, vol. 21, no. 4, pp. S50–S57, 2007.

BIBLIOGRAPHY

[91] F. Bai, Z. Zhang, D. R. Watson, H. Yu, Y. Shi, Y. Yuan, Y. Zang, C. Zhu, and Y. Qian, "Abnormal functional connectivity of hippocampus during episodic memory retrieval processing network in amnestic mild cognitive impairment," *Biological psychiatry*, vol. 65, no. 11, pp. 951–958, 2009.

[92] T. Satterthwaite, S. Vandekar, D. Wolf, D. Bassett, K. Ruparel, Z. Shehzad, R. Craddock, R. Shinohara, T. Moore, E. Gennatas *et al.*, "Connectome-wide network analysis of youth with psychosis-spectrum symptoms," *Molecular psychiatry*, 2015.

[93] S. Whitfield-Gabrieli, H. W. Thermenos, S. Milanovic, M. T. Tsuang, S. V. Faraone, R. W. McCarley, M. E. Shenton, A. I. Green, A. Nieto-Castanon, P. LaViolette *et al.*, "Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia," *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 1279–1284, 2009.

[94] K. R. Van Dijk, T. Hedden, A. Venkataraman, K. C. Evans, S. W. Lazar, and R. L. Buckner, "Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization," *Journal of neurophysiology*, vol. 103, no. 1, pp. 297–321, 2010.

[95] O. S. Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015.

BIBLIOGRAPHY

[96] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.

[97] M. Munafò, S. Noble, W. J. Browne, D. Brunner, K. Button, J. Ferreira, P. Holmans, D. Langbehn, G. Lewis, M. Lindquist *et al.*, "Scientific rigor and the art of motorcycle maintenance," *Nature biotechnology*, vol. 32, no. 9, pp. 871–873, 2014.

[98] Z. Shehzad, A. C. Kelly, P. T. Reiss, D. G. Gee, K. Gotimer, L. Q. Uddin, S. H. Lee, D. S. Margulies, A. K. Roy, B. B. Biswal *et al.*, "The resting brain: unconstrained yet reliable," *Cerebral cortex*, vol. 19, no. 10, pp. 2209–2229, 2009.

[99] J. S. Anderson, M. A. Ferguson, M. Lopez-Larson, and D. Yurgelun-Todd, "Reproducibility of single-subject functional connectivity measurements," *American journal of neuroradiology*, vol. 32, no. 3, pp. 548–555, 2011.

[100] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nature neuroscience*, 2015.

[101] A. S. Choe, C. K. Jones, S. E. Joel, J. Muschelli, V. Belegu, B. S. Caffo,

M. A. Lindquist, P. C. van Zijl, and J. J. Pekar, "Reproducibility and temporal structure in weekly resting-state fMRI over a period of 3.5 years," *PloS one*, vol. 10, no. 10, p. e0140134, 2015.

[102] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for fMRI," *NeuroImage*, vol. 54, no. 2, pp. 875–891, 2011.

[103] D. A. Dawson, K. Cha, L. B. Lewis, J. D. Mendola, and A. Shmuel, "Evaluation and calibration of functional network modeling methods based on known anatomical connections," *NeuroImage*, vol. 67, pp. 331–343, 2013.

[104] X.-N. Zuo, T. Xu, L. Jiang, Z. Yang, X.-Y. Cao, Y. He, Y.-F. Zang, F. X. Castellanos, and M. P. Milham, "Toward reliable characterization of functional homogeneity in the human brain: preprocessing, scan duration, imaging resolution and computational space," *NeuroImage*, vol. 65, pp. 374–386, 2013.

[105] C. C. Guo, F. Kurth, J. Zhou, E. A. Mayer, S. B. Eickhoff, J. H. Kramer, and W. W. Seeley, "One-year test–retest reliability of intrinsic connectivity network fMRI in older adults," *NeuroImage*, vol. 61, no. 4, pp. 1471–1483, 2012.

[106] J. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. Stam, S. M. Smith, and C. Beckmann, "Consistent resting-state networks across healthy subjects," *Proceedings of the national academy of sciences*, vol. 103, no. 37, pp. 13 848–13 853, 2006.

[107] S. Chen, T. J. Ross, W. Zhan, C. S. Myers, K.-S. Chuang, S. J. Heishman, E. A. Stein, and Y. Yang, "Group independent component analysis reveals consistent resting-state networks across multiple sessions," *Brain research*, vol. 1239, pp. 141–151, 2008.

[108] Y.-h. Chou, L. P. Panych, C. C. Dickey, J. R. Petrella, and N.-k. Chen, "Investigation of long-term reproducibility of intrinsic connectivity network mapping: a resting-state fMRI study," *American Journal of Neuroradiology*, vol. 33, no. 5, pp. 833–838, 2012.

[109] K. Kalcher, W. Huf, R. N. Boubela, P. Filzmoser, L. Pezawas, B. Biswal, S. Kasper, E. Moser, and C. Windischberger, "Fully exploratory network independent component analysis of the 1000 functional connectomes database," *Frontiers in human neuroscience*, vol. 6, 2012.

[110] C. D. Hacker, T. O. Laumann, N. P. Szrama, A. Baldassarre, A. Z. Snyder, E. C. Leuthardt, and M. Corbetta, "Resting state network estimation in individual subjects," *NeuroImage*, vol. 82, pp. 616–633, 2013.

[111] X. Li, J. Zhou, and X. Liu, "Influence of scan duration on the reliability of resting-state fMRI regional homogeneity," in *Pattern Recognition*. Springer, 2014, pp. 402–408.

[112] C. T. Whitlow, R. Casanova, and J. A. Maldjian, "Effect of resting-state func-

tional MR imaging duration on stability of graph theory metrics of brain network connectivity," *Radiology*, vol. 259, no. 2, pp. 516–524, 2011.

[113] X.-H. Liao, M.-R. Xia, T. Xu, Z.-J. Dai, X.-Y. Cao, H.-J. Niu, X.-N. Zuo, Y.-F. Zang, and Y. He, "Functional brain hubs and their test–retest reliability: a multiband resting-state functional MRI study," *NeuroImage*, vol. 83, pp. 969–982, 2013.

[114] K. Murphy, J. Bodurka, and P. A. Bandettini, "How long to scan? the relationship between fMRI temporal signal to noise ratio and necessary scan duration," *NeuroImage*, vol. 34, no. 2, pp. 565–574, 2007.

[115] S. Mueller, D. Wang, M. D. Fox, R. Pan, J. Lu, K. Li, W. Sun, R. L. Buckner, and H. Liu, "Reliability correction for functional connectivity: Theory and implementation," *Human Brain Mapping*, 2015.

[116] A. F. Mejia, M. B. Nebel, H. Shou, C. M. Crainiceanu, J. J. Pekar, S. Mostofsky, B. Caffo, and M. A. Lindquist, "Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators," *NeuroImage*, vol. 112, pp. 14–29, 2015.

[117] S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, and K. Uğurbil, "Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal

whole-brain fMRI," *Magnetic Resonance in Medicine*, vol. 63, no. 5, pp. 1144–1153, 2010.

[118] D. A. Feinberg, S. Moeller, S. M. Smith, E. Auerbach, S. Ramanna, M. Gunther, M. F. Glasser, K. L. Miller, K. Ugurbil, and E. Yacoub, "Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging," *PloS one*, vol. 5, no. 12, p. e15710, 2010.

[119] K. Setsompop, B. A. Gagoski, J. R. Polimeni, T. Witzel, V. J. Wedeen, and L. L. Wald, "Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty," *Magnetic Resonance in Medicine*, vol. 67, no. 5, pp. 1210–1224, 2012.

[120] J. Xu, S. Moeller, J. Strupp, E. Auerbach, L. Chen, D. Feinberg, K. Ugurbil, and E. Yacoub, "Highly accelerated whole brain imaging using aligned-blipped-controlled-aliasing multiband EPI," in *Proceedings of the 20th Annual Meeting of ISMRM*, vol. 2306, 2012.

[121] K. Uğurbil, J. Xu, E. J. Auerbach, S. Moeller, A. T. Vu, J. M. Duarte-Carvajalino, C. Lenglet, X. Wu, S. Schmitter, P. F. Van de Moortele *et al.*, "Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project," *NeuroImage*, vol. 80, pp. 80–104, 2013.

[122] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni *et al.*, "The mini-

mal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[123] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith, "Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers," *NeuroImage*, vol. 90, pp. 449–468, 2014.

[124] L. Griffanti, G. Salimi-Khorshidi, C. F. Beckmann, E. J. Auerbach, G. Douaud, C. E. Sexton, E. Zsoldos, K. P. Ebmeier, N. Filippini, C. E. Mackay *et al.*, "ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging," *NeuroImage*, vol. 95, pp. 232–247, 2014.

[125] C. F. Beckmann and S. M. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 2, pp. 137–152, 2004.

CURRICULUM VITAE

Amanda F. Mejia

Ph.D. Candidate
Department of Biostatistics
Johns Hopkins School of Public Health
615 N. Wolfe Street, E3038
Baltimore, MD 21205

Contact Information
Phone: (225) 978-7873
Email: amejia4@jhu.edu
Twitter: @mandyfmejia
http://mandymejia.wordpress.com

EDUCATION

2011 – 2016:	Ph.D. in Biostatistics Johns Hopkins School of Public Health, Baltimore, MD Advisors: Martin Lindquist, Ph.D. and Brian Caffo, Ph.D. Degree expected May 2016
2009 – 2011:	M.S. in Industrial and Systems Engineering Georgia Institute of Technology, Atlanta, GA Advisor: Eva K. Lee, Ph.D.
2002 – 2007:	B.S. in Industrial and Systems Engineering Georgia Institute of Technology, Atlanta, GA *Highest Honors*

GRANTS AND FELLOWSHIPS

2010-2015	PI, Graduate Research Fellowship, National Science Foundation

HONORS AND AWARDS

2016	Margaret Merrell Award for outstanding research by a Biostatistics doctoral student, The Johns Hopkins University Bloomberg School of Public Health
2015	JSM Best Student Paper Award, ASA Statistics in Imaging Section
2015	Student Paper Award Winner, Statistical Methods in Imaging Workshop
2014	Gertrude Cox Scholarship, Honorable Mention
2014	ENAR Best Poster Award

PUBLICATIONS

Webb-Vargas Y, Chen S, Fisher AJ, **Mejia AF**, Xu Y, Crainiceanu CM, Caffo B and Lindquist MA. Big data and neuroimaging. In press at *Statistics in Biosciences* (special issue on big data analytics).

Mejia AF, Sweeney EM, Dewey B, Nair G, Sati P, Shea C, Reich D and Shinohara RT. Statistical estimation of T1 relaxation time using conventional magnetic resonance imaging. In press at *NeuroImage*.

Paik J, Wigley FM, **Mejia AF**, Hummers LK. Severity of muscle weakness independently associates with disability as measured by the Health Assessment Questionnaire-Disability Index (HAQ-DI) in scleroderma. In press at *Arthritis Care and Research*.

Mejia AF, Nebel MB, Shou H, Crainiceanu CM, Pekar JJ, Mostofsky SH, Caffo B and Lindquist MA. Improving reliability of subject-level resting-state fMRI parcellations with shrinkage. *NeuroImage* 112 (2015): 14-29.

Lindquist MA and **Mejia AF**. Zen and the Art of Multiple Comparisons. *Psychosomatic Medicine* 77.2 (2015): 114-125.

Shou H, Eloyan A, Nebel MB, **Mejia AF**, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA and Crainiceanu, CM. Shrinkage Prediction of Seed-voxel Brain Connectivity using Resting-state fMRI. *NeuroImage* (2014).

Peterson D, Rajneesh M, Crocetti D, **Mejia AF**, Mostofsky SH. Left-Hemispheric Microstructural Abnormalities in Children With High Functioning Autism Spectrum Disorder. *Autism Research* (2014).

Ament K, **Mejia A**, Buhlman R, Erklin S, Caffo B, Mostofsky S, & Wodka E. Evidence for specificity of motor impairments in catching and balance in children with autism. *Journal of autism and developmental disorders* (2014): 1-10.

Lee EK, **Mejia AF**, Senior T and Jose J. Improving patient safety through medical alert management: an automated decision tool to reduce alert fatigue. *AMIA Annual Symposium Proceedings*. Vol. 2010. American Medical Informatics Association (2010).

Under Review

Mejia AF, Nebel MB, Eloyan A, Caffo B and Lindquist MA. PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data. Undergoing revisions for *Biostatistics*.

Pomann GM, Staicu AM, Lobaton E, **Mejia AF**, Dewey B, Reich DS, Sweeney E and Shinohara R. A lag functional linear model for prediction of magnetization transfer ratio in multiple sclerosis lesions. Undergoing revisions for *Annals of Applied Statistics*.

INVITED ORAL PRESENTATIONS

May 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Transitional Workshop, SAMSI Program on Challenges in Computational Neuroscience, Research Triangle Park, NC

April 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." 2016 New England Statistics Symposium, New Haven, CT

February 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Department of Statistics, Indiana University, Bloomington, IN

February 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Quantitative Sciences Unit, Stanford University, Stanford, CA

February 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Department of Biostatistics, University of North Carolina, Chapel Hill, NC

February 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Department of Biostatistics, Boston University, Boston, MA

February 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Department of Statistics and Data Science, University of Texas, Austin, TX

January 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Department of Biostatistics, University of Rochester, Rochester, NY

January 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Department of Biostatistics, Emory University, Atlanta, GA

January 2016 "PCA Leverage: Outlier Detection for High-Dimensional fMRI Data." Department of Statistics, University of Virginia, Charlottesville, VA

October 2015 "Improving Reliability of Resting-State Functional Connectivity Using Shrinkage Estimators, with an Application to the Human Connectome Project." Center for Biomedical Imaging Statistics, Department of Biostatistics, Emory University, Atlanta, GA

CONTRIBUTED ORAL PRESENTATIONS

March 2016 "Scan Length, Shrinkage and Reliability of Resting-State Functional Connectivity in the Human Connectome Project." 2016 Whistler Workshop on Brain Functional Organization, Connectivity, and Behavior, Whistler, BC, Canada

May 2015 "Automated Outlier Detection for fMRI." (Winner of Best Student Paper Award). Statistical Methods in Imaging Conference, Ann Arbor, MI

August 2014 "Improving Reliability of Subject-Level Resting-State Parcellation with Empirical Bayes Shrinkage." Joint Statistical Meetings 2014, Boston, MA

November 2010 "Optimizing Electronic Medical Records: Medical Alert Management." American Medical Informatics Association (AMIA) 2010 Annual Symposium, Washington, DC

November 2009 "Optimizing Electronic Medical Records: Medical Alert Management." Institute for Operations Research and the Management Sciences 2009 Annual Meeting, San Diego, CA

CONTRIBUTED POSTER PRESENTATIONS

August 2015 "Automated Outlier Detection for fMRI." Opening Workshop, SAMSI Program on Challenges in Computational Neuroscience, Research Triangle Park, NC

June 2015 "Automated Outlier Detection for fMRI." 2015 Annual Meeting of the Organization for Human Brain Mapping, Honolulu, HI

March 2015 "Automated PCA-Based Outlier Detection for fMRI." 2015 Annual Meeting of the Eastern North America Region of the International Biometric Society, Miami, FL

September 2014 "Improving Reliability of Subject-Level Resting-State fMRI Parcellation with Empirical Bayes." 2014 Conference on Resting State and Brain Connectivity, Boston, MA

September 2014 "Statistical Estimation of Quantitative T1 Using Standard Clinical Modalities." Joint Meeting of the American and European Committees for Treatment and Research in Multiple Sclerosis, Boston, MA

August 2014 "Improving Reliability of Subject-Level Resting-State Parcellation with Empirical Bayes Shrinkage." Joint Statistical Meetings 2014, Boston, MA

June 2014 "Improving Reliability of Subject-Level Resting-State fMRI Parcellation with Empirical Bayes." 2016 Annual Meeting of the Organization for Human Brain Mapping, Hamburg, Germany

March 2014 "Improving Reliability of Subject-Level Resting-State fMRI Parcellations with Shrinkage" (poster). 2014 Annual Meeting of the Eastern North America Region of the International Biometric Society, Baltimore, MD

November 2013 "Estimating T1 maps using clinical MRI." 2013 Imaging Conference, Johns Hopkins University, Baltimore, MD

RESEARCH EXPERIENCE

2013 – present: Research Intern
Center for Neurocognitive and Imaging Research, Kennedy Krieger Institute

2013 – present: Research Associate, Courtesy Appointment
Department of Biostatistics and Epidemiology, University of Pennsylvania

2011 – present: Graduate Research Assistant
Department of Biostatistics, Johns Hopkins School of Public Health

Summer 2012: Visiting Researcher
Department of Statistics, Ludwig Maximilian University of Munich

2009 – 2011: Graduate Research Assistant
Department of Industrial and Systems Engineering, Georgia Institute of Technology

WORKING GROUP MEMBERSHIP

2012 – present:	Statistical Methods and Applications for Research in Technology (SMART) Department of Biostatistics, Johns Hopkins School of Public Health
2013 – present:	Penn Statistical Imaging and Visualization Endeavor (PennSIVE) Department of Biostatistics and Epidemiology, University of Pennsylvania

TEACHING EXPERIENCE

Johns Hopkins

Spring 2016	Principles & Methods of Functional Neuroimaging I-II (Prof: Dr. Martin Lindquist) *Teaching Assistant*
Fall 2015	Statistical Reasoning in Public Health I-II, 100 students (Prof: Dr. John McGready) *Teaching Assistant*
Spring 2015	Methods in Biostatistics III-IV, 50 ScM students (Prof: Dr. Scott Zeger) *Lab Instructor. Taught computing skills, including R and reproducibility tools. Responsible for developing labs and solutions to homework assignments.*
Fall 2014	Methods in Biostatistics I-II, 50 ScM students (Prof: Dr. Ciprian Crainiceanu) *Guest Lecturer (2 lectures), Lab Instructor and Teaching Assistant*
Spring 2014	Advanced Statistical Methods in Biostatistics III-IV, 10 PhD students (Prof: Dr. Jeffrey Leek) *Guest Lecturer (1 lecture), Lab Instructor and Teaching Assistant*
Fall 2013	Statistical Methods in Public Health I-II, 500 students (Prof: Dr. Marie Diener-West) *Teaching Assistant*
Spring 2013	Advanced Statistical Theory I-II, 10 PhD students (Prof: Dr. Constantine Frangakis) *Guest Lecturer (1 lecture), Teaching Assistant*
	Masters in Public Health Capstone Project *Statistical Consultant*
Fall 2012	Statistical Reasoning in Public Health I-II, 100 students (Prof: Dr. John McGready) *Teaching Assistant*