STATISTICAL ANALYSIS OF FUNCTIONAL CONNECTIVITY IN BRAIN IMAGING: MEASUREMENT RELIABILITY AND CLINICAL APPLICATIONS

by

Zeyi Wang

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

Baltimore, Maryland
April, 2020

© 2020 Zeyi Wang
All rights reserved

Abstract

Measurement reliability is crucial for the research of functional connectivity data in the context of pursuing more reproducible research. Unfortunately, the utility of traditional reliability measures, such as the intraclass correlation coefficient, is limited given the size and complexity of functional connectivity data. In recent work, novel reliability measures have been introduced in the context where a set of subjects are measured twice or more, including: fingerprinting, rank sums, and generalizations of the intraclass correlation coefficient. However, the relationships between, and the best practices among these measures remains largely unknown. In this thesis, we consider a novel reliability measure, discriminability. We show that it is deterministically linked with the correlation coefficient under univariate random effect models, and has desired property of optimal accuracy for inferential tasks using multivariate measurements. Additionally, we propose a universal framework of reliability test based on permutations of the statistics.The power of permutation tests derived from these measures are compared numerically under Gaussian and non-Gaussian settings, with and without simulated batch effects. Motivated by both theoretical and empirical results, we provide methodological recommendations for each benchmark setting to serve as a resource for future analyses. We investigate the Poisson and Gaussian approximations of the tests so that the computational cost is reduced. We demonstrate possible follow-up research using reliability tests via applications on the Human Connectome Project functional connectivity data. We believe these results will play an important role towards improving reproducibility not only for functional connectivity, but also in fields such as functional magnetic resonance imaging in general, genomics, pharmacology, and more. Lastly, we illustrate the potential of functional connectivity as a source of causal biomarkers with an example of analyzing the trial data for an aphasia treatment.

Thesis Committee

Primary Readers

Brian Caffo (Primary Advisor)
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Martin Lindquist
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Justin Lessler
Associate Professor
Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health

Kyrana Tsapkini
Assistant Professor
Department of Neurology
Johns Hopkins School of Medicine

Joshua Vogelstein
Assistant Professor
Department of Biomedical Engineering
Johns Hopkins University

Alternate Readers

Vadim Zipunnikov
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Donna Strobino
Professor
Department of Population, Family and Reproductive Health
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

I hope to express my deepest gratitude to my advisor, Dr. Caffo, for the invaluable insight into the field of neuroimaging, guidance, encouragement, and relentless support. My gratitude extends to Dr. Vogelstein and Dr. Tsapkini. It would have not been possible for me to overcome the difficulty and continue the research without the greatest support from you.

I wish to express my appreciation to Dr. Lindquist, Dr. Frangakis, members at Dr. Tsapkini's lab and all my collaborators who have been extremely generous in sharing your time and thoughts. Without the advice and inspirations from the discussions with you, the projects would have not be realized.

I appreciate the understanding and help from Dr. Lessler and all the defense committee referees to join the virtual defense at this special difficult time.

Special thanks to the Johns Hopkins School of Public Health and the School of Medicine who have provided funding support for the dissertation research.

Finally, I would like to acknowledge the support and encouragement I have been receiving throughout my study from my family and my friends.

Table of Contents

Abstract ii

Acknowledgments vi

Table of Contents vii

List of Tables x

List of Figures xii

1 Introduction 1

2 Measurement Reliability Measures 11

- 2.1 Review of Existing Measurement Reliability Measures 11
  - 2.1.1 Intraclass Correlation Coefficients 11
  - 2.1.2 Fingerprinting 15
  - 2.1.3 Rank Sums 16
- 2.2 Discriminability as a Reliability Measure 17
  - 2.2.1 General Model of Repeated Measurements 18

2.2.2 Definition of Discriminability 19

2.2.3 Discriminability is Deterministically Linked with ICC . 25

2.2.4 Relation with Other Reliability Measures 28

2.2.4.1 Fingerprinting 28

2.2.4.2 Rank Sums 29

2.2.4.3 I2C2 30

3 Permutation Tests **39**

3.1 General Framework of Reliability Tests 39

3.2 Simulation on Hypothesis Testing Power 41

3.2.1 Univariate ANOVA Simulations 41

3.2.2 MANOVA Simulations 44

3.2.3 Batch Effects 45

3.2.4 Discussion 49

3.3 Approximations 51

3.3.1 Fingerprinting 51

3.3.2 Rank Sums 64

4 Covariate Analysis for Fingerprinting on HCP Data **70**

4.1 Covariates Associated with Matching 72

4.2 Matching for Comparing Connectome Similarities Between Twins or Non-Twin Siblings 73

4.3 Brain Maps of Identifying Pairs of Nodes by Network 77

4.4 Discussion 78

5 Clinical Applications of Functional Connectivity 83

5.1 Causal Biomarkers Discovery in a PPA Trial 83

5.1.1 Data 84

5.1.1.1 Participants and Overall Design 84

5.1.1.2 Methods of tDCS 87

5.1.1.3 Language Intervention 87

5.1.1.4 Language and Cognitive Assessment 89

5.1.1.5 Imaging Acquisition and Preprocessing . . . 89

5.1.2 Unbiased Transformations and Causal Biomarkers . . 91

5.1.3 Predictor Selection 93

5.1.4 Results 94

5.2 Discussion 97

6 Conclusions 101

Curriculum Vitae 105

List of Tables

3.1 Example resampling matrix from matching with replacement. Here the statistic value is 3. 52

3.2 Example resampling matrix from matching without replacement. Here the statistic value is 2. 53

4.1 Numbers of the identifying pairs between the five combined networks on the HCP dataset (see Section 4.3). The pairs of nodes were selected by the Poisson approximated permutation test on the total matches from matching with replacement using only the z transformed correlations between each single pair. . 78

4.2 Identification rates (in %) from matching with replacement using only the z transformed correlations of the pairs between the five combined networks on the HCP dataset (see Section 4.3). 78

5.1 Means (standard deviations) of demographics grouped by first-phase condition (n=36). *Fisher's exact test used. **F(1, 26) reported. FTD-CDR, Frontotemporal Dementia Clinical Dementia Rating Scale sum of boxes (Knopman et al., 2008). F, female; M, male. L, logopenic; N, nonfluent; S semantic. . . . 86
5.2 Means (standard deviations) of demographics grouped by PPA variant (n=36). *Fisher's exact test used. **F(2, 25) reported. FTD-CDR, Frontotemporal Dementia Clinical Rating Scale sum of boxes (Knopman et al., 2008). F, female; M, male. s, sham; t, tDCS. 86
5.3 Imaging factors for individual tDCS effect prediction. Predictiveness was evaluated by the LOOCV (predictive) R^2. 94
5.4 The linear prediction model for individual tDCS effects with the two selected imaging factors. 94
5.5 Non-imaging factors for individual tDCS effect prediction. . . 97

List of Figures

2.1 The relation between discriminability and ICC under the ANOVA model with Gaussian random effects. See Section 2.2.3. 27

2.2 Non-decreasing bounds of the discriminability approximation (2.19) using functions of I2C2 under the MANOVA model with random Gaussian effects. The dispersion measures, defined as V_1^2/W_1 and V_2^2/W_2, are fixed at 10 or 30. The upper (red) and lower (blue) bounds are color coded, respectively. The dispersion 10 scenario is plotted with solid lines whereas the dispersion 30 scenario is plotted with dashed lines. V_1, W_1 (or V_2, W_2) are the sum and the sum of squares of the positive (or negative) eigenvalues from the distributional decomposition (2.18). See Section 2.2.4.3. 32

3.1 ANOVA simulations when the Gaussian assumption is satisfied (left) or violated with logarithm transformations (right). Simulated distributions of estimators are plotted on the top, including the discriminability estimation (using the estimator $\tilde{D}$ or the rank sum version $\hat{D}_{rs}$), the fingerprint index estimation, and the ICC estimation. Simulated permutation test powers are plotted on the bottom, where solid lines and dotted lines represent nonparametric and parametric statistics, respectively. $\sigma^2 = 5$. $\sigma_\mu^2 = 3$. n ranges from 5 to 40. $1,000$ iterations in total. See Section 3.2.1. 43

3.2 MANOVA simulations when the Gaussian assumption is satisfied (left) or violated with element-wise logarithm transformations (right). Simulated distributions of estimators are plotted on the top, including the discriminability estimation (using the estimator $\tilde{D}$ or the rank sum version $\hat{D}_{rs}$), the fingerprint index, and the I2C2. Simulated permutation test powers are plotted on the bottom, where solid lines and dotted lines represent nonparametric and parametric statistics, respectively. $\sigma^2 = 5$. $\sigma_\mu^2 = 3, \rho = 0.5, l = 10$. n ranges from 5 to 40. $1,000$ iterations in total. See Section 3.2.2. 46

3.3 Simulations for mean shifts (left), scaling (middle), and no batch effect (right). Simulated distributions of the discriminability estimators ($\tilde{D}$ and $\hat{D}_{rs}$) are plotted on the top, including six estimation strategies (Section 3.2.3). Simulated permutation test powers are plotted on the bottom, where solid lines and dotted lines represent the discriminability ($\tilde{D}$) based and the rank sum ($\hat{D}_{rs}$) based strategies, respectively. $s = 15, \sigma^2 = 3$, $\sigma_\mu^2 = 5$. The number of subjects is $n = 20$. $1,000$ iterations are conducted for each scenario. 48

3.4 Relation between discriminability and I2C2 with smaller ($\rho = 0.1$, left) or larger ($\rho = 0.5$, right) within measurement correlation. The Gaussian MANOVA model in Section 3.2.2 is assumed with $l = 10, n = 20, s = 2$. Covariance matrices, $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_\mu$, are proportional to a matrix with diagonals being 1 and off-diagonals being ρ. Small circles are the simulated ($1,000$ iterations) true discriminability with $\sigma_\mu^2 = 100$ and σ^2 ranging from 3 to 300. This shows error of the approximation (2.19) is within 0.1 and the non-decreasing relation holds true even with larger ρ value. 50

4.1 The simulated distribution of the total number of matches when matching two groups of distinct people (each of size 20) who were randomly selected from different families and were matched in age and race in the HCP dataset (see Section 4.1). Matching with replacement was conducted. The empirical distribution of a Poisson(1) random variable after 1,000 iterations was also plotted as comparison. 74

4.2 The simulated distributions of the total number of matches when matching two groups of distinct people (each of size 20) from the HCP dataset. For each person selected in the first group, there was another monozygotic twin/dizygotic twin/non-twin sibling/random person in the second group for the MZ/DZ/NotTwin/Random scenarios, respectively (see Section 4.2). Matching with replacement was conducted. The empirical distribution of a Poisson(1) random variable after 1,000 iterations is also plotted as comparison. 76

4.3 The 20 identifying pairs of nodes within the FP network visualized on the ICBM 152 brain template (see Section 4.3). Nodes were labeled by their orders on the atlas and were plotted at the center. Pairs of nodes were colored from blue to red depending on the number of matches when matching with replacement was conducted with only the z transformed correlations between each single pair. 79

5.1 Participants recruited and randomization to tDCS or sham. . 85

5.2 Predictive imaging factors and the increases of predictive R-squared for individual tDCS effects. Solid lines represent the selected factors in each round, whereas dotted lines represent the factors that were not selected but also provided over 0.1 increase of predictive R-squared in the first round. 95

5.3 Visualization of the selected predictive imaging factors for individual tDCS effects. The positions of the nodes are the average centers of each ROI from the cohort. ROI pairs are plotted and connected if predictiveness of the baseline connectivity is confirmed. Thickness of the edge and the scale of the edge color represent on average how much extra potential semantic fluency score increase one would expect for tDCS compared to sham, with 0.01 higher baseline Fisher z-transformed connectivity. 96

Chapter 1

Introduction

The reproducibility crisis is a concern in many scientific domains (Baker, 2016; Open Science Collaboration and others, 2015), including and perhaps especially the field of functional neuroimaging (Button et al., 2013), where noise, an absence of replicability, site variation, and inter- and intra-scanner variation are known issues. Behind the crises, measurement reliability plays a crucial role. In addition to data consistency being conceptually fundamental for results (Bennett and Miller, 2010), reliability is used as a key tool to detect likely irreproducible findings and statistical errors. For example, a recent outcry over issues in repeated use of data in the field of cognitive neuroscience (Vul et al., 2009) relied on absence of the required reliability as proof of the issue. Some have also argued that the misinterpretation of reliability can result in false confidence in a study's reproducibility and subsequently lead to the neglect of important design issues (Turner et al., 2018). A thorough investigation and accurate interpretation of measurement reliability is crucial for a better understanding of existing issues of reproducibility and working towards better future practices.

Functional connectivity (FC) is a recently developed data type derived from resting-state functional magnetic resonance imaging (rs-fMRI) (Van Den Heuvel and Pol, 2010). Functional MRI records blood-oxygen-level-dependent (BOLD) time series from small regions of the brain. In contrast, rs-fMRI does so while a subject is at rest in the scanner considering simultaneous synchronous behavior. Despite the promising prospect of non-invasively measuring functional brain connectivity in vivo, FC raises questions of data quality by its nature. Since the synchronous fluctuations are evaluated by second order statistics (usually correlations or network-based graph metrics), FC is potentially noisier than other fMRI data, which already involves complex acquisition and processing choices. Resting state correlations are particularly sensitive to biological confounds, in contrast to task based fMRI, where the confound is often not correlated with the task. Variability can be induced by changes in physiological and cognitive status of a subject, within a single scan session or between two sessions that are hour, days, or months apart. In addition, common practices in the field can raise questions in data quality too (Zuo and Xing, 2014; Jiang et al., 2015). For example, auto-correlations in the BOLD time series might violate independence and parametric assumptions in correlation analyses. Averaging the time series over a large region may involve voxels with low functional homogeneity and introduce spurious variability. It is also a concern when, as is typical, a number of reasonable preprocessing options are available that produce highly variable measurement outcomes. Processing choices can be particularly difficult to generalize across studies, since target measurements can be on different scales or formed with a different data reduction strategy (seed-to-voxels, voxel-by-voxel, region-by-region, etc.).

In all of these scenarios, understanding measurement reliability of FC is a prerequisite for any meaningful scientific discovery or clinical application.

The evaluation of reliability is crucial, not only because of the varying quality of correlation-based FC measurements (Noble, Scheinost, and Constable, 2019), but also for its broader implications. Some examples include: *i*. selecting best practices for data acquisition and preprocessing (Pervaiz et al., 2019), *ii*. identifying FC biomarkers (Gabrieli, Ghosh, and Whitfield-Gabrieli, 2015; Castellanos et al., 2013; Kelly et al., 2012), *iii*. optimizing FC-based prediction models (Svaldi et al., 2019), and *iv*. evaluating the accuracy of multi-class prediction algorithms (Zheng, Achanta, and Benjamini, 2018).

To avoid ambiguity, measurement reliability is defined as consistency or similarity across technical replicates of a measurement. We restrict the use of the term without assuming one of the replicates is the correct, true measurement. The same definition is often referred to as test-retest reliability, reliability, repeatability, or the reproducibility of a measurement procedure, where the consistency of repeated measurements is being emphasized (Müller and Büttner, 1994). However, caution should be taken that the general concepts of reliability and reproducibility are often applied beyond the definition of repeated measurements' consistency, depending on the actual context. General reviews of the concept of research reproducibility, with comparison to replicability can be found: in Goodman, Fanelli, and Ioannidis, 2016 and Patil, Peng, and Leek, 2016. Reviews of the general reliability of fMRI can be found in Bennett and Miller, 2010 and Fröhner et al., 2017, where emphasis was put on the reliability of results, not necessarily restricted to measurement. For example, popular cluster-overlap-based reliability measures, such as the Dice coefficient and Jaccard index, are designed specifically for the consistency of the inferential results. Moreover, a rich literature exists for other related, but distinct, types of reliability, such as inter-rater reliability (an overview can be found Gwet, 2014). A similar issue in fMRI is in inter-site, inter-scanner or inter-technologist reliability, which is not discussed in detail herein, but the measures discussed in this thesis can be applied. In summary, we selectively focus on the evaluation of measurement reliability, as a crucial starting point for evaluating measurement validity.

FC raises new challenges for reliability evaluation. For example, the intraclass correlation coefficient (ICC) is a commonly used metric for test-retest reliability. However, the ICC is limited in several ways when applied for FC data. First, it was developed for univariate data, and there is no consensus on how one should synthesize multiple ICC's over each dimension of the measurement, or for measurements with different dimensions. The definition and inference of ICC is based on a relatively strict parametric analysis of variance (ANOVA) model assuming separability and additivity. Often, Gaussian assumptions are applied for inference, an assumption that is suspect in fMRI studies. Ideally, an objective reliability measure, preferably non-parametric and able to accommodate varying data dimensions, is needed.

Recently, several novel reliability measures have been proposed, including fingerprinting, which is based on the idea of subject identification (Finn et al., 2015; Finn et al., 2017; Wang et al., 2018), rank sums (Airan et al., 2016), and the image intraclass correlation coefficient (I2C2) (Shou et al., 2013), which is a generalization of the classical univariate ICC. Unlike univariate methods, such as ICC, these newly proposed methods can handle high-dimensional imaging data and computationally scale. By building the measures on ranks transformations, the nonparametric methods (fingerprinting, rank sums) are robust to model violations.

However, the relations between, and the best practices among, these methods remains largely unknown. Furthermore, clear relationships in interpretations and performance are lacking. Thus, often less effective or robust measures of FC measurement quality are being used, potentially leading to worse study practices, worse processing pipelines and sub-optimal application of FC-based prediction algorithms.

In Chapter 2, we particularly focus on discriminability (Bridgeford et al., 2019), a new data quality measure. We argue that discriminability is in fact a measure of reliability. It is defined upon a general repeated measurement model that is free of parametric assumptions, yet remains deterministically linked to ICC for univariate measurement, when ANOVA assumptions are met. The flexibility of the general repeated measurement model also allows us to investigate the mathematical relationships of discriminability with all other multivariate reliability measures. These analytical results give the first insights into the relationships between the recently proposed reliability measures.

In Chapter 3, we propose a framework of permutation testing specifically designed for discovering the evidence for the existence of measurement reliability. Thus, the aforementioned reliability measures can be numerically compared in the terms of their ability to detect significance in such permutation tests. To summarize, our results illustrate the general power advantages of discriminability when compared to other nonparametric methods, and its robustness advantages against the violation of Gaussian assumptions, when compared to parametric methods. Of course, parametric methods may be more powerful when distributional assumptions are satisfied. In addition,the rank sum method shows additional robustness against mean shift batch effects compared to discriminability. Moreover, we give Poisson or normal approximations for the permutation tests of fingerprinting or rank-based statistics, respectively. This allows power analysis for large samples and reduces the computational cost for the scenarios when the tests are performed in large batches. However, in Chapter 4, we note that evidence beyond the test result is desirable for assessing reliability. We focus on fingerprinting in this example, where both the individual score and approximation of the permutation distribution are available. The results highlight that covariates can be associated with the individual estimates of reliability and require further investigation.

Although not being one of the common targets of interventions, FC has potential in both investigating treatment mechanisms and promoting personalized medicines in clinical trials. In Chapter 5, we will demonstrate the opportunities of application and methodological development with an example from a randomized trial of primary progressive aphasia (PPA) patients and transcranial direct current stimulation (tDCS) treatments. However, while this and other applications of resting state fMRI being suggest its potential as a biomarker (Finn et al., 2015; Rosenberg et al., 2016), precaution should be taken based on the lessons learned from the study of its reliability.

References

Baker, M. (2016). "1,500 scientists lift the lid on reproducibility". In: *Nature* 533.7604, pp. 452–454.

Open Science Collaboration and others (2015). "Estimating the reproducibility of psychological science". In: *Science* 349.6251, aac4716.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò (2013). "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature Reviews Neuroscience* 14.5, pp. 365–376.

Bennett, Craig M and Michael B Miller (2010). "How reliable are the results from functional magnetic resonance imaging?" In: *Annals of the New York Academy of Sciences* 1191.1, pp. 133–155.

Vul, Edward, Christine Harris, Piotr Winkielman, and Harold Pashler (2009). "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition". In: *Perspectives on psychological science* 4.3, pp. 274–290.

Turner, Benjamin O, Erick J Paul, Michael B Miller, and Aron K Barbey (2018). "Small sample sizes reduce the replicability of task-based fMRI studies". In: *Communications Biology* 1.1, pp. 1–10.

Van Den Heuvel, Martijn P and Hilleke E Hulshoff Pol (2010). "Exploring the brain network: a review on resting-state fMRI functional connectivity". In: *European neuropsychopharmacology* 20.8, pp. 519–534.

Zuo, Xi-Nian and Xiu-Xia Xing (2014). "Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective". In: *Neuroscience & Biobehavioral Reviews* 45, pp. 100–118.

Jiang, Lili, Ting Xu, Ye He, Xiao-Hui Hou, Jinhui Wang, Xiao-Yan Cao, Gao-Xia Wei, Zhi Yang, Yong He, and Xi-Nian Zuo (2015). "Toward neurobiological characterization of functional homogeneity in the human cortex: regional variation, morphological association and functional covariance network organization". In: *Brain Structure and Function* 220.5, pp. 2485–2507.

Noble, Stephanie, Dustin Scheinost, and R Todd Constable (2019). “A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis”. In: *NeuroImage*, p. 116157.

Pervaiz, Usama, Diego Vidaurre, Mark W Woolrich, and Stephen M Smith (2019). “Optimising network modelling methods for fMRI”. In: *bioRxiv*, p. 741595.

Gabrieli, John DE, Satrajit S Ghosh, and Susan Whitfield-Gabrieli (2015). “Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience”. In: *Neuron* 85.1, pp. 11–26.

Castellanos, F Xavier, Adriana Di Martino, R Cameron Craddock, Ashesh D Mehta, and Michael P Milham (2013). “Clinical applications of the functional connectome”. In: *Neuroimage* 80, pp. 527–540.

Kelly, Clare, Bharat B Biswal, R Cameron Craddock, F Xavier Castellanos, and Michael P Milham (2012). “Characterizing variation in the functional connectome: promise and pitfalls”. In: *Trends in cognitive sciences* 16.3, pp. 181–188.

Svaldi, Diana O, Joaquín Goñi, Kausar Abbas, Enrico Amico, David G Clark, Charanya Muralidharan, Mario Dzemidzic, John D West, Shannon L Risacher, Andrew J Saykin, et al. (2019). “Optimizing differential identifiability improves connectome predictive modeling of cognitive deficits in Alzheimer’s disease”. In: *arXiv preprint arXiv:1908.06197*.

Zheng, Charles, Rakesh Achanta, and Yuval Benjamini (2018). “Extrapolating expected accuracies for large multi-class problems”. In: *The Journal of Machine Learning Research* 19.1, pp. 2609–2638.

Müller, Reinhold and Petra Büttner (1994). “A critical discussion of intraclass correlation coefficients”. In: *Statistics in medicine* 13.23-24, pp. 2465–2476.

Goodman, Steven N, Daniele Fanelli, and John PA Ioannidis (2016). “What does research reproducibility mean?” In: *Science Translational Medicine* 8.341, 341ps12–341ps12.

Patil, Prasad, Roger D Peng, and Jeffrey Leek (2016). “A statistical definition for reproducibility and replicability”. In: *BioRxiv*, p. 066803.

Fröhner, Juliane H, Vanessa Teckentrup, Michael N Smolka, and Nils B Kroemer (2017). “Addressing the reliability fallacy: Similar group effects may arise from unreliable individual effects”. In: *BioRxiv*, p. 215053.

Gwet, Kilem L (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Finn, Emily S, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable

(2015). “Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity”. In: *Nature Neuroscience* 18.11, p. 1664.

Finn, Emily S, Dustin Scheinost, Daniel M Finn, Xilin Shen, Xenophon Papademetris, and R Todd Constable (2017). “Can brain state be manipulated to emphasize individual differences in functional connectivity?” In: *NeuroImage* 160, pp. 140–151.

Wang, Zeyi, Haris Sair, Ciprian Crainiceanu, Martin Lindquist, Bennett A Landman, Susan Resnick, Joshua T Vogelstein, and Brian Scott Caffo (2018). “On statistical tests of functional connectome fingerprinting”. In: *bioRxiv*, p. 443556.

Airan, Raag D, Joshua T Vogelstein, Jay J Pillai, Brian Caffo, James J Pekar, and Haris I Sair (2016). “Factors affecting characterization and localization of interindividual differences in functional connectivity using MRI”. In: *Human Brain Mapping* 37.5, pp. 1986–1997.

Shou, H, A Eloyan, S Lee, Vadim Zipunnikov, AN Crainiceanu, MB Nebel, B Caffo, MA Lindquist, and Ciprian M Crainiceanu (2013). “Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2)”. In: *Cognitive, Affective, & Behavioral Neuroscience* 13.4, pp. 714–724.

Bridgeford, Eric W, Shangsi Wang, Zhi Yang, Zeyi Wang, Ting Xu, Cameron Craddock, Gregory Kiar, William Gray-Roncal, Carey E Priebe, Brian Caffo, et al. (2019). “Optimal experimental design for big data: applications in brain imaging”. In: *bioRxiv*, p. 802629.

Rosenberg, Monica D, Emily S Finn, Dustin Scheinost, Xenophon Papademetris, Xilin Shen, R Todd Constable, and Marvin M Chun (2016). “A neuromarker of sustained attention from whole-brain functional connectivity”. In: *Nature Neuroscience* 19.1, p. 165.

Chapter 2

Measurement Reliability Measures

2.1 Review of Existing Measurement Reliability Measures

In this section, we define and investigate several measures of data reliability under associated statistical models. Notably, we define the measures as population quantities for statistical inference. We subsequently give the natural sample estimators for each.

2.1.1 Intraclass Correlation Coefficients

We consider two types of intraclass correlations, ICC and I2C2 (Shou et al., 2013). Without modifications, ICC is designed for evaluating the reliability for one dimensional measurements, such as expert ratings or composite mental health scores. It can also be utilized in various ways for multivariate measurements, for example, by averaging ICCs over each of the dimensions or by counting percentage of dimensions that pass a threshold on ICC. However, for the latter scenario there is no consensus on the best practice, and the interpretation is subjective based on the researcher's choices. ICC can be generalized to higher dimensions, provided a multivariate model that decomposes variation into a sum of intra- and inter-subject levels and a definition of the fraction of variation that is inter-subject. I2C2, is one such generalization of ICC for multivariate settings that was designed for high dimensional settings.

Other generalizations of ICC are outside the setting of interest for this thesis. For example, intraclass correlations can also be defined under various two-way ANOVA models (Shrout and Fleiss, 1979), which are suitable for the evaluation of inter-rater reliability or internal consistency. However, these measures are not relevant for the evaluation of test-retest reliability (Rousson, Gasser, and Seifert, 2002; Bruton, Conway, and Holgate, 2000). Other popular reliability measures, such as variations on the Alpha and Kappa statistics are not covered, for the same reason of being less relevant to the study of test-retest reliability.

To elaborate on models, for ICC, suppose that we have n subjects, each with s measurements. A univariate Analysis of Variance (ANOVA) model with Gaussian random effects is specified as:

$$x_{it} = \mu + \mu_i + e_{it}, \tag{2.1}$$

where $\mu_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right)$ and $e_{it} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ are mutually independent.

For l-dimensional measurements, (2.1) is generated as Multivariate Analysis of Variance (MANOVA) with Gaussian random effects:

$$\boldsymbol{x}_{it} = \boldsymbol{\mu} + \boldsymbol{\mu}_i + \boldsymbol{e}_{it}, \tag{2.2}$$

where $\boldsymbol{\mu}_i \overset{iid}{\sim} \mathcal{N}_l(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}})$, $\boldsymbol{e}_{it} \overset{iid}{\sim} \mathcal{N}_l(\mathbf{0}, \boldsymbol{\Sigma})$, independently. All the vectors are l-dimensional.

In the univariate case (2.1), ICC is defined as:

$$\lambda = \mathrm{corr}(x_{it}, x_{it'}) = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2},$$

for all $t' \neq t$. Assuming the measurements of a same subject form a class, then x_{it} and $x_{it'}$ are both from the i-th class, hence the name "intra-class".

For the multivariate case (2.2), a popular generalization of ICC using matrix determinants is

$$\Lambda = \frac{\det(\boldsymbol{\Sigma}_\mu)}{\det(\boldsymbol{\Sigma}) + \det(\boldsymbol{\Sigma}_\mu)},$$

commonly known as Wilks' lambda (Λ). Using matrix traces, the generalization becomes

$$\Lambda_{tr} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\mu)}{\mathrm{tr}(\boldsymbol{\Sigma}) + \mathrm{tr}(\boldsymbol{\Sigma}_\mu)}.$$

This reliability measure is particularly useful for high-dimensional imaging settings and was utilized in the the image intraclass correlation coefficient (I2C2) (Shou et al., 2013). Recall that the trace of the covariance matrix captures the total variability of the random quantity of interest. Then, Λ_{tr} intuitively represents the fraction of the variability in the observed data $\boldsymbol{x}_{it}$ due to the subject effect $\boldsymbol{\mu}_i$.

It is well known that an estimator for ICC is

$$\hat{\lambda} = \frac{MS_B - MS_W}{MS_B + (s-1) \cdot MS_W}, \tag{2.3}$$

where the means square between and within subjects are

$$MS_B = \left(\sum_{i=1}^{n} s(\bar{x}_{i.} - \bar{x}_{..})^2 \right) / \, (n-1) \, ,$$

$$MS_W = \left(\sum_{i=1}^{n} \sum_{t=1}^{s} (x_{it} - \bar{x}_{..})^2 - \sum_{i=1}^{n} s(\bar{x}_{i.} - \bar{x}_{..})^2 \right) / \, (ns-n) \, .$$

In addition, the F statistic is defined as

$$F = \frac{MS_B}{MS_W}.$$

It follows that $\hat{\lambda} = (F-1)(F-1+s)$, which is a non-decreasing function of the F statistic given $s \geq 2$.

I2C2 was estimated using a hierarchical generalization on principal components called multilevel functional principal components analysis (MFPCA) (Di et al., 2009). The MFPCA algorithm utilizes a moment based approach to separate variability into inter- and intra-subject components in a method similar to Henderson's equations in mixed models (Henderson et al., 1959). Singular value decomposition tricks can be used to make calculations tractable in higher dimensions (Zipunnikov et al., 2011). In principle, other multivariate approaches can be used to estimate Λ_{tr} and Λ. For example, it would be a straightforward change in I2C2 to estimate Λ instead of Λ_{tr}. In addition, latent Gaussian models (Chib and Greenberg, 1998) can extend these approaches to binary data and graphs (Yue et al., 2015).

One of the commonly discussed properties of ICC is its relation with the optimal correlation between two univariate outcomes (Vul et al., 2009; Bennett and Miller, 2010; Zuo and Xing, 2014). It states:

$$\mathrm{corr}\left(x_{it}^{1}, x_{it}^{2}\right) = \mathrm{corr}\left(\mu_i^1, \mu_i^2\right)\sqrt{\mathrm{ICC}(x_{it}^1)\cdot \mathrm{ICC}(x_{it}^2)},$$

where x_{it}^1 and x_{it}^2 follow the ANOVA model respectively, without the requirement of Gaussian distributions.

2.1.2 Fingerprinting

As its name suggests, fingerprinting is the idea of matching subjects to themselves in repeated measurements where errors could potential occur by mismatches with other subjects (Wang et al., 2018). The count or proportion of matches for a matching scheme represents an intuitive summary of data reliability. This measure has become especially popular in neuroimaging due to a few highly visible articles (Anderson et al., 2011; Finn et al., 2015; Xu et al., 2016).

We first formalize the idea of a population-level fingerprinting measure for repeated measurements. It is assumed that each subject is measured twice, and that the measurement is possibly multivariate. Then each subject, i, at time point, t, has measurement, $\boldsymbol{x}_{it}$, $i = 1, \ldots, n$, $t = 1, 2$. Suppose there exists a distance metric, $\delta(\cdot, \cdot)$, defined between measurements, $\delta_{i,1,2} = \delta(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, and $\delta_{i,i',1,2} = \delta(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i'2})$. Define the population level fingerprint index as:

$$F_{index} = \mathbb{P}\left(\delta_{i,1,2} < \delta_{i,i',1,2};\ \forall\, i' \neq i\right), \tag{2.4}$$

where the probability is calculated over a random sample of n subjects. This is the population probability that a random subject matches themselves over any other in the sample.

Implicitly, such a measure is defined under a much more flexible model. For (2.4) to be a meaningful population quantity, it is only required that the resulted F_{index} is equal for all i's, which covers the (M)ANOVA models (2.1) and (2.2) with Gaussian random effects as special cases. However, the relationship between ICC and the fingerprinting index is unknown.

The natural estimate of (2.4) is the proportion of correct matches in a group of subjects. This requires assuming a matching strategy, such as whether matching is done with or without replacement (Wang et al., 2018). Almost all fingerprint index studies use matching with replacements as follows. The total number of correct matches (with replacement) is:

$$T_n = \sum_{i=1}^{n} \mathbb{I}_{\left\{\delta_{i,1,2} < \delta_{i,i',1,2};\, \forall i' \neq i\right\}}, \tag{2.5}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. Then, the fingerprint index estimator is simply the proportion of correct matches:

$$\hat{F}_{index} = \frac{T_n}{n}. \tag{2.6}$$

2.1.3 Rank Sums

In the test-retest setting with $s = 2$, the fingerprint statistic can be generalized as a Mann-Whitney style statistic. Instead of counting the events where $\boldsymbol{x}_{i2}$ is the closest to $\boldsymbol{x}_{i1}$ among all other $\boldsymbol{x}_{i'2}$ with $i' \neq i$, consider calculating the rank. Formally, the rank sum statistic is defined by summing up r_{ii}'s, the rank of $\delta_{i,1,2}$ among all $\delta_{i,i',1,2}$ with $i' \neq i$. Assuming that there are no ties (or the max ranks are assigned) then the rank sum statistic is defined as:

$$R_n = \sum_{i=1}^{n} r_{ii} = \sum_{i=1}^{n} \sum_{i' \neq i} \mathbb{I}_{\left\{\delta_{i,1,2} < \delta_{i,i',1,2}\right\}}. \tag{2.7}$$

Notice that $\mathbb{I}_{\left\{\delta_{i,1,2} < \delta_{i,i',1,2};\, \forall\, i' \neq i\right\}} = \mathbb{I}_{\{r_{ii}=1\}}$; thus the ranks are sufficient for determining the fingerprint index. Of course, the fingerprinting statistic ignores the information contained in ranks, other than the number of the ranks equal to 1 within subjects. Thus, it may seem obvious that the rank sum statistic is superior to the fingerprint statistic in some sense. However, it should also be noted that the rank sum statistic lacks an intuitive relationship with a population quantity, like the fingerprint statistic does with the fingerprint index. In addition, both the fingerprint and rank sum statistics lack an obvious generalization for repeated measurements, as they were developed on compared paired measurements.

2.2 Discriminability as a Reliability Measure

In this section, we will formally define the concept of discriminability under a flexible model of repeated measurements. We will then prove that discriminability is indeed a reliability measure, as it is deterministically related to ICC when the Gaussian ANOVA assumptions are met. Notably, an optimal accuracy property of discriminability in the Bayes error rate is applicable for multivariate measurements (Bridgeford et al., 2019), whereas this property has only been shown under univariate measurements for ICC. We will also investigate the relation between discriminability and the other aforementioned measures with the goal of increasing interpretability across studies when using different reliability measures.

2.2.1 General Model of Repeated Measurements

Let $\boldsymbol{v}_i \in \boldsymbol{\mathcal{V}}$ be a true physical property of interest for subject i. Without the ability to directly observe $\boldsymbol{v}_i$, we instead observe $\boldsymbol{w}_{it} = f_\phi(\boldsymbol{v}_i, t)$, for some random measurement process $f_\phi : \boldsymbol{\mathcal{V}} \times T \to \boldsymbol{\mathcal{W}}$, where $\phi \in \boldsymbol{\Phi}$ characterizes the measurement process, and $\boldsymbol{w}_{it} \in \boldsymbol{\mathcal{W}}$ is the observed measurement of property $\boldsymbol{v}_i$. As f_ϕ is a random process, the index, $t \in T$, is used to emphasize that the observation $\boldsymbol{v}_i$ using process f_ϕ may differ across repeated trials, typically performed sequentially in time.

In many settings, the measurement process may suffer from known or unknown confounds created in the process of measurement. For example, when taking a magnetic resonance image (MRI) of a brain, the MRI may be corrupted by motion (movement) or signal intensity artifacts. The observed data, $\boldsymbol{w}_{it}$, may therefore be unsuitable for direct inference, and instead is pre-processed via the random process $g_\psi : \boldsymbol{\mathcal{W}} \to \boldsymbol{\mathcal{X}}$ to reduce measurement confounds. Here, $\psi \in \boldsymbol{\Psi}$ characterizes the pre-processing procedure chosen, such as motion or other artifact correction in our MRI example. We define $\boldsymbol{x}_{it} = g_\psi \circ f_\phi(\boldsymbol{v}_i, t)$ as the pre-processed measurement of $\boldsymbol{v}_i$ for subject i from measurement index t. Let $\delta : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ be a distance metric. Simplified notations such as $\delta_{i,t,t'} = \delta(\boldsymbol{x}_{it}, \boldsymbol{x}_{it'})$ and $\delta_{i,i',t,t''} = \delta(\boldsymbol{x}_{it}, \boldsymbol{x}_{i't''})$ can be used in characterizing the reliability.

Measurement reliability can be considered as a function of the combination of an acquisition procedure, ϕ, and a chosen pre-processing procedure, ψ. Of course, it can be defined exclusively for a subset of the data generating procedure. For instance, when the data has already been collected, the researchers may only be able to manipulate pre-processing, ψ and not acquisition, ϕ, procedures. Then, one intended use of the reliability measure is to optimize over those aspects of the measurement process the researcher is able to manipulate: $\psi_* = \arg\max_{\psi \in \boldsymbol{\Psi}} u(\psi)$, where u is an unspecified reliability measure.

Throughout the rest of the chapter, we may analyze the nonparametric measures under the following additive noise model in order to maintain tractability:

$$\boldsymbol{x}_{it} = \boldsymbol{v}_i + \boldsymbol{\epsilon}_{it} \tag{2.8}$$

where $\boldsymbol{\epsilon}_{it} \overset{ind}{\sim} f_\epsilon$, and $\text{var}(\boldsymbol{\epsilon}_{it}) < \infty$ with $\mathbb{E}[\boldsymbol{\epsilon}_{it}] = \boldsymbol{c}$. Such modeling still contains (M)ANOVA scenarios as special cases and is free of parametric assumptions, where the fingerprinting index and the discriminability are both well-defined.

2.2.2 Definition of Discriminability

If the measurement procedure is effective, we would anticipate that our physical property of interest for any subject i, $\boldsymbol{v}_i$, would differ from that of another subject i', $\boldsymbol{v}_{i'}$. Thus, an intuitive notion of reliability would expect that subjects would be more similar to themselves than to other subjects. Specifically, we would expect in a good measurement that $\boldsymbol{x}_{it}$ is more similar to $\boldsymbol{x}_{it'}$ (a repeated measurement on subject i) than to $\boldsymbol{x}_{i't''}$ (a measurement on subject i' at time t'').

Discriminability is defined as:

$$D(\psi, \phi) = \mathbb{P}(\delta_{i,t,t'} < \delta_{i,i',t,t''}). \tag{2.9}$$

Similar to the fingerprinting index, discriminability is well defined as long as $D(\psi, \phi)$ is equal for all i, i', t, t', t'' (such that $i \neq i', t \neq t'$). That is, this definition assumes that discriminability does not depend on the specific subjects and measurements being considered. This can be considered a form of exchangeability. Subsequently, we consider models that are consistent with this definition in the Gaussian (M)ANOVA models (2.1), (2.2). One could consider a form of population averaged discriminability if D does depend on subjects. However, this is outside of the scope of this thesis.

To estimate discriminability, assume that for each individual, i, we have s repeated measurements. Sample discriminability is then defined as:

$$\hat{D} = \frac{\sum\limits_{i=1}^{n} \sum\limits_{t=1}^{s} \sum\limits_{t' \neq t} \sum\limits_{i' \neq i} \sum\limits_{t''=1}^{s} \mathbb{I}_{\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\}}}{n \cdot s \cdot (s-1) \cdot (n-1) \cdot s}, \tag{2.10}$$

where n is the total number of subjects. Then $\hat{D}$ represents the fraction of observations where $\boldsymbol{x}_{it}$ is more similar to $\boldsymbol{x}_{it'}$ than to the measurement $\boldsymbol{x}_{i't''}$ of another subject i', for all pair of subjects $i \neq i'$ and all pairs of time points $t \neq t'$.

Under the additive noise model (2.8), it can be proven that $\hat{D}$ is unbiased and consistent for discriminability.

Unbiasedness and Consistency of $\hat{D}$

Assume that for each individual i, we have s repeated measurements. We define the local discriminability:

$$\hat{D}^{n}_{i,t,t'} = \frac{\sum\limits_{i' \neq i} \sum\limits_{t''=1}^{s} \mathbb{I}_{\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\}}}{s \cdot (n-1)} \tag{2.11}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function, and n is the total number of subjects. Then $\hat{D}_{i,t,t'}$ represents the fraction of observations from other subjects that are more distant from $\boldsymbol{x}_{it}$ than $\boldsymbol{x}_{it'}$, or a local estimate of the discriminability for individual i between measurements t and t'. The sample discriminability estimator is:

$$\hat{D}_{n} = \frac{\sum\limits_{i=1}^{n} \sum\limits_{t=1}^{s} \sum\limits_{t' \neq t} \hat{D}_{i,t,t'}}{n \cdot s \cdot (s-1)}, \tag{2.12}$$

where $D_{i,t,t'}$ is the local discriminability. We establish first the unbiasedness for the local discriminability, under the additive noise setting (2.8):

$$\boldsymbol{x}_{it} = \boldsymbol{v}_{i} + \boldsymbol{\epsilon}_{it},$$

where $\boldsymbol{\epsilon}_{it} f_{\epsilon}$, and $\text{var}(\boldsymbol{\epsilon}_{it}) < \infty$ with $\mathbb{E}[\epsilon_{it}] = c$. That is, our additive noise can be characterized by bounded variance and fixed expectation, and our noise is independent across subjects.

Lemma 2.2.2.1 (local discriminability is unbiased for discriminability). *For fixed n:*

$$\mathbb{E}\left[\hat{D}^{n}_{i,t,t'}\right] = D; \tag{2.13}$$

that is, the local discriminability is unbiased for the true discriminability.

Proof.

$$\mathbb{E}\left[\hat{D}^n_{i,t,t'}\right] = \mathbb{E}\left[\frac{\sum\limits_{i' \neq i}\sum\limits_{t''=1}^{s} \mathbb{I}_{\left\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\right\}}}{s \cdot (n-1)}\right]$$

$$= \frac{\sum\limits_{i' \neq i}\sum\limits_{t''=1}^{s} \mathbb{E}\left[\mathbb{I}_{\left\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\right\}}\right]}{s \cdot (n-1)}$$

$$= \frac{\sum\limits_{i' \neq i}\sum\limits_{t''=1}^{s} \mathbb{P}(\delta_{i,t,t'} < \delta_{i,i',t,t''})}{s \cdot (n-1)}$$

$$= \frac{\sum\limits_{i' \neq i}\sum\limits_{t''=1}^{s} D}{s \cdot (n-1)}$$

$$= \frac{s \cdot (n-1) \cdot D}{s \cdot (n-1)}$$

$$= D$$

□

Without knowledge of the distribution of $\boldsymbol{x}_{it}$, we can instead estimate the discriminability via $\hat{D}(\phi, \psi)$, the observed sample discriminability. Consider the additive noise case. Recall that $\hat{D}_n \equiv \hat{D}_n(\phi, \psi)$, the sample discriminability for a fixed number of individuals n. We consider the following two lemmas:

Lemma 2.2.2.2 (Unbiasedness of Sample Discriminability). *For fixed n:*

$$\mathbb{E}\left[\hat{D}_n\right] = D,$$

that is, the sample discriminability is an unbiased estimate of discriminability.

Proof. The proof of this lemma is a rather trivial application of the result in Lemma (2.2.2.1).

Recall that sample discriminability is as-defined in Equation (2.12). Then:

$$\begin{aligned}
\mathbb{E}\left[\hat{D}_n\right] &= \mathbb{E}\left[\frac{\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{s}\sum\limits_{t'\neq t}\hat{D}_{i,t,t'}}{n\cdot s\cdot(s-1)}\right] \\
&= \frac{\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{s}\sum\limits_{t'\neq t}\mathbb{E}\left[\hat{D}^n_{i,t,t'}\right]}{n\cdot s\cdot(s-1)} \\
&= \frac{\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{s}\sum\limits_{t'\neq t}D}{n\cdot s\cdot(s-1)} \qquad \text{Lemma (2.2.2.1)} \\
&= \frac{n\cdot s\cdot(s-1)\cdot D}{n\cdot s\cdot(s-1)} \\
&= D
\end{aligned}$$

□

Lemma 2.2.2.3 (Consistency of Sample Discriminability). *As $n \to \infty$:*

$$\hat{D}_n \xrightarrow[n\to\infty]{\mathcal{P}} D.$$

That is, the sample discriminability is a consistent estimate of discriminability.

Proof. Recall that Chebyshev's inequality gives:

$$\mathbb{P}\left(\left|\hat{D}_n - \mathbb{E}\left[\hat{D}_n\right]\right| \geq \epsilon\right) = \mathbb{P}\left(\left|\hat{D}_n - D\right| \geq \epsilon\right) \qquad \hat{D}^n_{i,t,t'} \text{ is unbiased}$$
$$\leq \frac{\operatorname{var}\left(\hat{D}_n\right)}{\epsilon^2}$$

To show convergence in probability, it suffices to show that $\operatorname{var}\left(\hat{D}_n\right) \xrightarrow[n\to\infty]{} 0$. Then:

$$\operatorname{var}\left(\hat{D}_n\right) = \operatorname{var}\left(\frac{\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{s}\sum\limits_{t'\neq t}\hat{D}^n_{i,t,t'}}{n\cdot s\cdot(s-1)}\right)$$
$$= \frac{1}{m_*^2}\operatorname{var}\left(\sum_{i=1}^{n}\sum_{t=1}^{s}\sum_{t'\neq t}\sum_{i'\neq i}\sum_{t''=1}^{s}\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,i',t,t''}\right\}}\right)$$
$$= \frac{1}{m_*^2}\sum_{i,i',t,t',t''}\sum_{j,j',r,r',r''}\operatorname{cov}\left(\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,i',t,t''}\right\}},\mathbb{I}_{\left\{\delta_{j,r,r'}<\delta_{j,j',r,r''}\right\}}\right),$$

where $m_* = n\cdot s\cdot(s-1)\cdot(n-1)\cdot s$.

Note that there are, in total, m_*^2 covariance terms in the sums. For each term, by Cauchy-Schwarz:

$$\left|\operatorname{cov}\left(\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,i',t,t''}\right\}},\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,j',t,r''}\right\}}\right)\right| \leq \sqrt{\operatorname{var}\left(\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,i',t,t''}\right\}}\right)\cdot\operatorname{var}\left(\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,j',t,r''}\right\}}\right)}$$
$$\leq \sqrt{\frac{1}{4}\cdot\frac{1}{4}} = \frac{1}{4}.$$

Furthermore, note that $\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,i',t,t''}\right\}} = f(\boldsymbol{x}_{i,t},\boldsymbol{x}_{i,t'},\boldsymbol{x}_{i',t''})$. Under the assumption of between-subject independence, then $\mathbb{I}_{\left\{\delta_{i,t,t'}<\delta_{i,i',t,t''}\right\}} \perp g\left(\boldsymbol{x}_{i'',q} : i'' \neq i, i'\right)$, as it will be independent of any function $g(\cdot)$ of subjects other than i and i'. Then as long as $\{i,i'\} \cap \{j,j'\} = \varnothing$, we have $\mathbb{I}_{\left\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\right\}} \perp\!\!\!\perp \mathbb{I}_{\left\{\delta_{j,r,r'} < \delta_{j,j',r,r''}\right\}}$. Under the assumption that $\forall\, i, n_i = s$, we have $m_* = ns^2(s-1)(n-1)$. Then there are $(n-2)s^2(s-1)(n-3)$ combinations of j, j', r, r', r'' that will produce covariances taking values of 0, and $m_* - (n-2)s(s-1)(n-3)s = (4n-6)\cdot s^2 \cdot (s-1)$ combinations that may be non-zero. Then:

$$\begin{aligned}
\mathrm{var}(\hat{D}_n) &= \frac{1}{m_*^2} \sum_{i,i',t,t',t''} \sum_{j,j',r,r',r''} \mathrm{cov}\left(\mathbb{I}_{\left\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\right\}}, \mathbb{I}_{\left\{\delta_{j,r,r'} < \delta_{j,j',r,r''}\right\}}\right) \\
&\leq \frac{\sum_{i,i',t,t',t''} (4n-6)s^2(s-1)}{4m_*^2} \\
&= \frac{(4n-6)s^2(s-1)}{4ns^2(s-1)(n-1)} \\
&= \frac{4n-6}{4n(n-1)} \\
&< \frac{1}{n} \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

□

2.2.3 Discriminability is Deterministically Linked with ICC

Interestingly, under the ANOVA model (2.1), discriminability is deterministically linked to ICC. It is relatively easy to argue and instructive on the relationship between these constructs, and therefore we present the argument here. Considering a Euclidean distance as the metric, discriminability (D) is:

$$D = \mathbb{P}(|x_{it} - x_{it'}| < |x_{it} - x_{i't''}|)$$

$$= \mathbb{P}(|e_{it} - e_{it'}| < |\mu_i - \mu_{i'} + e_{it} - e_{it''}|)$$

$$\overset{def}{=} \mathbb{P}(|A| < |B|)$$

for $i \neq i'$, $t \neq t'$. Then $(A, B)^t$ follows a joint normal distribution, with mean vector $\mathbf{0}$ and covariance matrix $\begin{pmatrix} 2\sigma^2 & \sigma^2 \\ \sigma^2 & 2\sigma_\mu^2 + 2\sigma^2 \end{pmatrix}$. Hence:

$$D = 1 - \frac{\arctan\left(\frac{\sqrt{\sigma^2(3\sigma^2 + 4\sigma_\mu^2)}}{\sigma_\mu^2}\right)}{\pi} = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{\text{ICC}}{\sqrt{(1 - \text{ICC})(\text{ICC} + 3)}}\right). \tag{2.14}$$

Therefore, D and ICC are deterministically linked with a non-decreasing transformation under the ANOVA model with Gaussian random effects. Figure 2.1 shows a plot of the non-linear relationship. For an ICC of roughly 0.68, the two measures are equal, with discriminability being smaller for ICCs larger than 0.68 and larger for ICCs lower. It is perhaps useful to let $D^* = 2D - 1$ to transform discriminability to range between 0 to 1, similar to ICC.

Recall, the optimal correlation between two univariate measurements is equal to a non-decreasing function of the ICC of each of the measurement. Since discriminability is deterministically linked to ICC via a strictly increasing function, this property also holds for discriminability.

Another scenario where the reliability measure may become critical is in the prediction problem with multivariate predictors. Under this scenario, the optimal prediction error in terms of the Bayes error rate of a classification task can be bounded by a decreasing function of discriminability of the multivariate predictors (Bridgeford et al., 2019). Thus, it is interesting to note that ICC inherits this property exactly, as it holds for any one-to-one transformation of

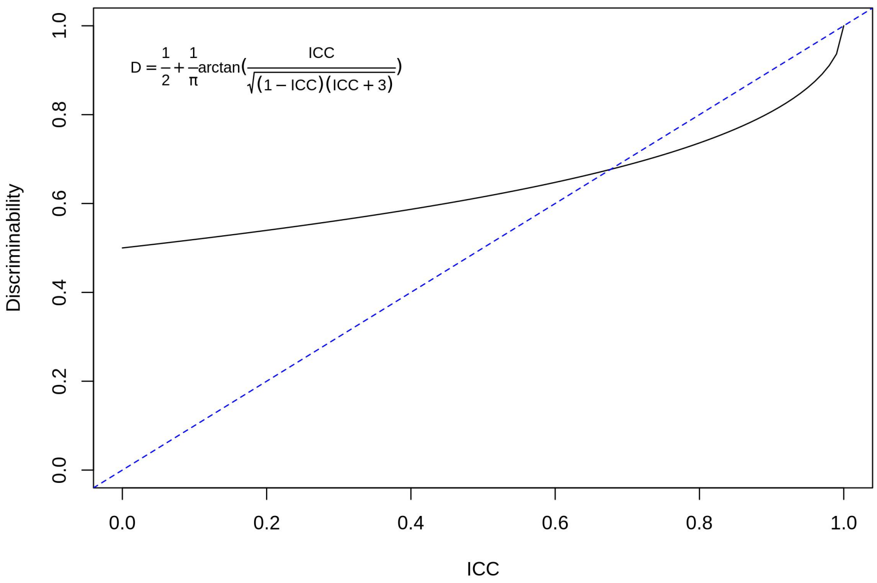

Figure 2.1: The relation between discriminability and ICC under the ANOVA model with Gaussian random effects. See Section 2.2.3.

discriminability.

2.2.4 Relation with Other Reliability Measures

2.2.4.1 Fingerprinting

In a test-retest setting, where the fingerprint index is defined, it can ve proved that the fingerprint index has the following relationship with the discriminability, D,

$$F_{index} = \rho D + (1-\rho) D^{n-1},$$

as long as the correlation, $\rho \overset{def}{=} corr(\mathbb{I}_{\{\delta_{i,1,2} < \delta_{i,i',1,2}\}}, \mathbb{I}_{\{\delta_{i,1,2} < \delta_{i,i'',1,2}\}})$, is non-negative for $i', i'' \neq i$.

The non-negativity condition can be checked with simulation or numerical integrals when a parametric model is posited. For example, it holds under the Gaussian ANOVA model, (2.1), where the correlation, ρ, is positive for all of the simulated values of σ^2 and σ_μ^2 between 0 and 100.

Assuming non-negativity of ρ, the fingerprint index decreases to a limit of ρD, as the sample size, n, increases. However, the diminishing term, $(1-\rho)D^{n-1}$, may not be negligible with large enough D and small enough n. This illustrates the fact that the fingerprint index may not be invariant for different sample sizes that are below 10 to 15, even when discriminability is constant.

2.2.4.2 Rank Sums

Discriminability has no direct relationship with fingerprinting, which is a function of the distance rank matrix. However, interestingly, sample discriminability can be rewritten as a function of a form of rank sums. This suggest that discriminability retains the rank information that the fingerprint index discards. Below we demonstrate this relationship.

Denote the n by n inter-measurement distance sub-matrix as $\boldsymbol{D}^{t,t'} = (\delta_{i,i',t,t'})_{i,i'=1,\dots,n}$. Let the combined $n \cdot s$ by $n \cdot s$ distance matrix be $\boldsymbol{D} = \left(\boldsymbol{D}^{t,t'}\right)_{t,t'=1,\dots,s}$, which consists of s by s blocks where the (t,t') block is $\boldsymbol{D}^{t,t'}$. Let $r_{i,i'}^{t,t'}$ denote the ranking within rows in the combined distance matrix $\boldsymbol{D} = (\delta_{i,i',t,t'})$. We assign the maximum ranks for ties.

Another consistent estimator of discriminability in the rank form is

$$\tilde{D} = \frac{n^2 s^2 (s-1) - \sum_{t=1}^{s} \sum_{t' \neq t} \sum_{i=1}^{n} r_{ii}^{tt'}}{ns(s-1)(n-1)s} \tag{2.15}$$

or $\tilde{D} - \frac{s-2}{2(n-1)s}$, where

$$0 \leq \tilde{D} - \hat{D} \leq \frac{s-2}{2(n-1)s}. \tag{2.16}$$

Equality is taken in (2.16) when no tie exists between $\delta_{i,i}^{t,t'}$ and $\delta_{i,i}^{t,t''}$ for all $i \in \{1,\dots,n\}$, $t \in \{1,\dots s\}$, $t' \notin \{t\}$, $t'' \notin \{t,t'\}$. Therefore we have that $\tilde{D}$ and $\tilde{D} - \frac{s-2}{2(n-1)s}$ are also consistent estimators for discriminability. In fact $\hat{D} = \tilde{D} - \frac{s-2}{2(n-1)s}$ when assuming continuous measurements with no ties in distance ranking. This representation highlights the close relationship between discriminability and rank sums.

In fact, the specific form of the rank sum statistic, (2.7), can be transformed to another estimator of discriminability. In a test-retest setting with $s = 2$, instead of ranking the combined distance matrix, $\boldsymbol{D}$, let r_{ij} be the rank of $\delta_{i,j}^{1,2}$ among $\delta_{i,1}^{1,2}, \ldots, \delta_{i,n}^{1,2}$, which ranks the row of the inter-measurement distance sub-matrix $\boldsymbol{D}^{1,2}$. If ties occur, the max ranks are assigned.

This transformation of the rank sum statistic, R_n, forms an unbiased and consistent estimator of D:

$$\hat{D}_{rs} = \frac{\sum_{i=1}^{n}(n - r_{ii})}{n(n-1)} = \frac{n^2 - R_n}{n(n-1)}. \tag{2.17}$$

If there exist multiple measurements for each subject, for all the pairs of distinct t_1 and t_2, the rank sum statistic and estimation can be calculated between the t_1-th measurements and the t_2-th measurement. Comparing to $\hat{D}$ and $\tilde{D}$, the rank sum statistic does not involve any ranking information from the diagonal blocks in the combined distance matrix, $\boldsymbol{D}^{t,t}, t = 1, \ldots, s,$. This may result in a larger standard error for estimation and a lower power for inference using the rank sums. However, it provides some robustness against mean shift batch effects, as demonstrated in Section 3.2.3.

2.2.4.3 I2C2

Under the l-dimensional MANOVA model specified in (2.2), again considering a Euclidean distance metric, discriminability becomes:

$$D = P(||\boldsymbol{x}_{it} - \boldsymbol{x}_{it'}|| - ||\boldsymbol{x}_{it} - \boldsymbol{x}_{i't''}|| < 0)$$

$$= P(||\boldsymbol{e}_{it} - \boldsymbol{e}_{it'}|| - ||\boldsymbol{e}_{it} - \boldsymbol{e}_{i't''} + \boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}|| < 0)$$

$$\overset{def}{=} P(||\boldsymbol{A}|| - ||\boldsymbol{B}|| < 0),$$

where A and B are jointly multivariate normal with means 0, variances $2\boldsymbol{\Sigma}$ and $2\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$, respectively, and covariance, $\boldsymbol{\Sigma}$. Note that $Z \overset{def}{=} \boldsymbol{A}^t\boldsymbol{A} - \boldsymbol{B}^t\boldsymbol{B}$ is an indefinite quadratic form of the vector $(\boldsymbol{A}^t\ \boldsymbol{B}^t)^t$ (around a matrix whose block diagonal entries are an identity matrix and the negative of an identity matrix). Thus, Z can be decomposed as a linear combination of independent χ^2 variables (Provost and Rudiuk, 1996):

$$Z \overset{D}{=} \sum_{u=1}^{r} \lambda_u U_u - \sum_{u=r+1}^{r+w} \lambda'_u U_u, \tag{2.18}$$

where $\lambda_1, \ldots, \lambda_r$ are the positive eigenvalues of $\begin{pmatrix} 2\boldsymbol{\Sigma} & -\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & -2\boldsymbol{\Sigma} - 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \end{pmatrix}$, $\lambda'_{r+1}, \ldots, \lambda'_{r+w}$ are the absolute values of the negative eigenvalues of $\begin{pmatrix} 2\boldsymbol{\Sigma} & -\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & -2\boldsymbol{\Sigma} - 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \end{pmatrix}$, $U_1, \ldots, U_{r+w}$ are IID χ^2 variables with degrees of freedom being 1.

Although this does not result in a deterministic link between D and I2C2, it can be shown that there exist approximations matching the first two moments of $\sum_{u=1}^{r} \lambda_u U_u$ and $\sum_{u=r+1}^{r+w} \lambda'_u U_u$. Furthermore, the approximation of D can be bounded by two non-decreasing functions of I2C2 (Appendix 2.2.4.3). Specifically, the resulting discriminability approximation has the form of a CDF value of an F-distribution,

$$D = P(Z \leq 0) \approx F_{F\left(\frac{V_1^2}{W_1}, \frac{V_2^2}{W_2}\right)}\left(\frac{V_2}{V_1}\right), \tag{2.19}$$

where V_1, W_1 (or V_2, W_2) are the sum and the sum of squares of the absolute values of the positive (or negative) eigenvalues. Moreover, when V_1, W_1, V_2, W_2

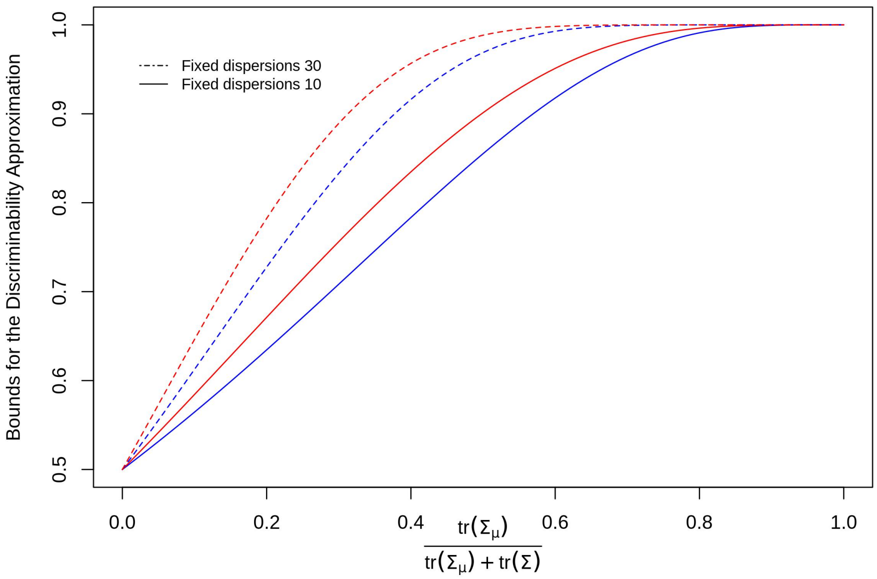

Figure 2.2: Non-decreasing bounds of the discriminability approximation (2.19) using functions of I2C2 under the MANOVA model with random Gaussian effects. The dispersion measures, defined as V_1^2/W_1 and V_2^2/W_2, are fixed at 10 or 30. The upper (red) and lower (blue) bounds are color coded, respectively. The dispersion 10 scenario is plotted with solid lines whereas the dispersion 30 scenario is plotted with dashed lines. V_1, W_1 (or V_2, W_2) are the sum and the sum of squares of the positive (or negative) eigenvalues from the distributional decomposition (2.18). See Section 2.2.4.3.

are constant, the approximation is bounded by a non-decreasing interval of I2C2 (Figure 2.2):

$$F_{F\left(\frac{V_1^2}{W_1},\frac{V_2^2}{W_2}\right)}\left(f_1(\Lambda_{tr})\right) \leq F_{F\left(\frac{V_1^2}{W_1},\frac{V_2^2}{W_2}\right)}\left(\frac{V_2}{V_1}\right) \leq F_{F\left(\frac{V_1^2}{W_1},\frac{V_2^2}{W_2}\right)}\left(f_2(\Lambda_{tr})\right),$$

where $f_1(v) = 1 + v/(1-v)$ and $f_2(v) = 1 + (4/3)\cdot v/(1-v)$ are both non-decreasing functions.

Appendix: Discriminability and I2C2

We will give the approximation and then prove the non-decreasing bounds in Section 2.2.4.3.

Applying the Satterthwaite approximation that matches the first two moments (Yuan and Bentler, 2010), we have $\sum_{u=1}^{r} \lambda_u U_u \overset{D}{\approx} g_1\chi^2_{h_1}$ and $\sum_{u=r+1}^{r+w} \lambda'_u U_u \overset{D}{\approx} g_2\chi^2_{h_2}$, where $g_1 = \left(\sum_{u=1}^{r} \lambda_u^2\right) / \left(\sum_{u=1}^{r} \lambda_u\right)$, $h_1 = \left(\sum_{u=1}^{r} \lambda_u\right)^2 / \left(\sum_{u=1}^{r} \lambda_u^2\right)$, $g_2 = \left(\sum_{u=r+1}^{r+w} \lambda_u'^2\right) / \left(\sum_{u=r+1}^{r+w} \lambda'_u\right)$, $h_2 = \left(\sum_{u=r+1}^{r+w} \lambda'_u\right)^2 / \left(\sum_{u=r+1}^{r+w} \lambda_u'^2\right)$. Let $V_1 = \sum_{u=1}^{r} \lambda_u = h_1 g_1$, $W_1 = \sum_{u=1}^{r} \lambda_u^2$, $V_2 = \sum_{u=r+1}^{r+w} \lambda'_u = h_2 g_2$, $W_2 = \sum_{u=r+1}^{r+w} \lambda_u'^2$. Thus:

$$\begin{aligned} D = P(Z \leq 0) &\approx P\left(\frac{g_1\chi^2_{h_1}}{g_2\chi^2_{h_2}} \leq 1\right) \\ &= P\left(\frac{\chi^2_{h_1}/h_1}{\chi^2_{h_2}/h_2} \leq \frac{h_2 g_2}{h_1 g_1}\right) \\ &= F_{F\left(\frac{V_1^2}{W_1},\frac{V_2^2}{W_2}\right)}\left(\frac{V_2}{V_1}\right). \end{aligned}$$

Here, $\frac{\chi^2_{h_1}/h_1}{\chi^2_{h_2}/h_2}$ follows F distribution with degrees of freedom $h_1 = \frac{V_1^2}{W_1}, h_2 = \frac{V_2^2}{W_2}$.

Now we derive the non-decreasing bounds. Note that $\boldsymbol{H} \overset{def}{=} \begin{pmatrix} 2\boldsymbol{\Sigma} & -\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & -2\boldsymbol{\Sigma} - 2\boldsymbol{\Sigma_\mu} \end{pmatrix} = \begin{pmatrix} 2\boldsymbol{\Sigma} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & 2\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma_\mu} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I} \end{pmatrix} \overset{def}{=} \boldsymbol{PM}$ is congruent to $\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I} \end{pmatrix}$ since $\boldsymbol{P}$ is symmetric and positive definite. By Sylvester's law of inertia (Sylvester, 1852) we have $r = w = l$, i.e. the numbers of positive and negative eigenvalues of $\boldsymbol{H}$ are both l.

Denote the sums of positive or negative eigenvalues of the matrix $\boldsymbol{H}$ as

$\sigma_+(\boldsymbol{H})$ or $\sigma_-(\boldsymbol{H})$, respectively. We will apply the monotonicity of $\sigma_\pm(\boldsymbol{H}) = \sigma_\pm(\boldsymbol{MP})$ (Lieb and Siedentop, 1991) for the following statements (Appendix 2.2.4.3):

Lemma 2.2.4.1 (Monotonicity of Sums of Positive or Negative Eigenvalues)**.**

$$tr(\frac{3}{2}\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}}) \leq |\sigma_-(\boldsymbol{H})| = V_2 \leq tr(2\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}}) \tag{2.20}$$

$$tr(\frac{3}{2}\boldsymbol{\Sigma}) \leq \sigma_+(\boldsymbol{H}) = V_1 \leq tr(2\boldsymbol{\Sigma}). \tag{2.21}$$

Proof. For (2.20), note that $\boldsymbol{P} - \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + v\boldsymbol{\Sigma} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2-v \end{pmatrix} \otimes \boldsymbol{\Sigma}$ is positive definite for all $v \in (0, 3/2)$. Therefore

$$|\sigma_-(\boldsymbol{MP})| \geq |\sigma_-(\boldsymbol{M} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + v\boldsymbol{\Sigma} \end{pmatrix})|$$

for all $v \in (0, 3/2)$. Finally,

$$|\sigma_-(\boldsymbol{MP})| \geq \lim_{v \to \frac{3}{2}} |\sigma_-(\boldsymbol{M} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + v\boldsymbol{\Sigma} \end{pmatrix})| = \text{tr}\, \frac{3}{2}\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}}.$$

Meanwhile $\begin{pmatrix} v\boldsymbol{\Sigma} & \boldsymbol{\Sigma} \\ \boldsymbol{0} & 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + v\boldsymbol{\Sigma} \end{pmatrix} - \boldsymbol{P} = \begin{pmatrix} v-2 & 0 \\ -1 & v-2 \end{pmatrix} \otimes \boldsymbol{\Sigma}$ is positive definite for all $v > 2$. Similarly,

$$tr(2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + 2\boldsymbol{\Sigma}) = \lim_{v \to 2} |\sigma_-(\boldsymbol{M} \begin{pmatrix} v\boldsymbol{\Sigma} & \boldsymbol{\Sigma} \\ \boldsymbol{0} & 2\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + v\boldsymbol{\Sigma} \end{pmatrix})| \geq |\sigma_-(\boldsymbol{MP})|.$$

To get (2.21) from (2.20), note $V_1 - V_2 = tr(\boldsymbol{H}) = tr(-2\boldsymbol{\Sigma}_{\boldsymbol{\mu}})$. □

Therefore,

$$\frac{V_2}{V_1} = 1 + \frac{2tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{V_1} \in \left(1 + \frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma})}, 1 + \frac{4}{3} \cdot \frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma})}\right)$$

$$= \left(f_1\left(\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}\right), f_2\left(\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}\right) \right),$$

where $f_1(v) = 1 + v/(1-v)$ and $f_2(v) = 1 + (4/3) \cdot v/(1-v)$ are both non-decreasing functions.

If $l = 2$, by the monotonicity of F distribution (Ghosh, 1973) we have bounds for the approximation (2.19):

$$F_{F(2,1)}\left(f_1\left(\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}\right) \right) \le F_{F(2,1)}\left(\frac{V_2}{V_1}\right) \le F_{F\left(\frac{V_1^2}{W_1}, \frac{V_2^2}{W_2}\right)}\left(\frac{V_2}{V_1}\right)$$
$$\le F_{F(1,2)}\left(f_2\left(\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}\right) \right),$$

where f_1, f_2, $F_{F(2,1)}$, $F_{F(1,2)}$ are all non-decreasing functions.

For $l \ge 3$, when the dispersion measures V_1^2/W_1 and V_2^2/W_2 remain constants (in fact $1 \le V_j^2/W_j \le l$ for $j = 1, 2$ by the property of l_1 and l_2 norms), the approximation of D in (2.19) is bounded by a non-decreasing interval of I2C2 (Figure 2.2):

$$F_{F\left(\frac{V_1^2}{W_1}, \frac{V_2^2}{W_2}\right)}\left(f_1\left(\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}\right) \right) \le F_{F\left(\frac{V_1^2}{W_1}, \frac{V_2^2}{W_2}\right)}\left(\frac{V_2}{V_1}\right)$$
$$\le F_{F\left(\frac{V_1^2}{W_1}, \frac{V_2^2}{W_2}\right)}\left(f_2\left(\frac{tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}{tr(\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma}_{\boldsymbol{\mu}})}\right) \right).$$

References

Shou, H, A Eloyan, S Lee, Vadim Zipunnikov, AN Crainiceanu, MB Nebel, B Caffo, MA Lindquist, and Ciprian M Crainiceanu (2013). “Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2)”. In: *Cognitive, Affective, & Behavioral Neuroscience* 13.4, pp. 714–724.

Shrout, Patrick E and Joseph L Fleiss (1979). “Intraclass correlations: uses in assessing rater reliability.” In: *Psychological Bulletin* 86.2, p. 420.

Rousson, Valentin, Theo Gasser, and Burkhardt Seifert (2002). “Assessing intrarater, interrater and test–retest reliability of continuous measurements”. In: *Statistics in medicine* 21.22, pp. 3431–3446.

Bruton, Anne, Joy H Conway, and Stephen T Holgate (2000). “Reliability: what is it, and how is it measured?” In: *Physiotherapy* 86.2, pp. 94–99.

Di, Chong-Zhi, Ciprian M Crainiceanu, Brian S Caffo, and Naresh M Punjabi (2009). “Multilevel functional principal component analysis”. In: *The Annals of Applied Statistics* 3.1, p. 458.

Henderson, Charles R, Oscar Kempthorne, Shayle R Searle, and CM Von Krosigk (1959). “The estimation of environmental and genetic trends from records subject to culling”. In: *Biometrics* 15.2, pp. 192–218.

Zipunnikov, Vadim, Brian Caffo, David M Yousem, Christos Davatzikos, Brian S Schwartz, and Ciprian Crainiceanu (2011). “Multilevel functional principal component analysis for high-dimensional data”. In: *Journal of Computational and Graphical Statistics* 20.4, pp. 852–873.

Chib, Siddhartha and Edward Greenberg (1998). “Analysis of multivariate probit models”. In: *Biometrika* 85.2, pp. 347–361.

Yue, Chen, Shaojie Chen, Haris I Sair, Raag Airan, and Brian S Caffo (2015). “Estimating a graphical intra-class correlation coefficient (GICC) using multivariate probit-linear mixed models”. In: *Computational Statistics & Data Analysis* 89, pp. 126–133.

Vul, Edward, Christine Harris, Piotr Winkielman, and Harold Pashler (2009). “Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition”. In: *Perspectives on Psychological Science* 4.3, pp. 274–290.

Bennett, Craig M and Michael B Miller (2010). “How reliable are the results from functional magnetic resonance imaging?” In: *Annals of the New York Academy of Sciences* 1191.1, pp. 133–155.

Zuo, Xi-Nian and Xiu-Xia Xing (2014). “Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective”. In: *Neuroscience & Biobehavioral Reviews* 45, pp. 100–118.

Wang, Zeyi, Haris Sair, Ciprian Crainiceanu, Martin Lindquist, Bennett A Landman, Susan Resnick, Joshua T Vogelstein, and Brian Scott Caffo (2018). “On statistical tests of functional connectome fingerprinting”. In: *bioRxiv*, p. 443556.

Anderson, Jeffrey S, Michael A Ferguson, Melissa Lopez-Larson, and Deborah Yurgelun-Todd (2011). “Reproducibility of single-subject functional connectivity measurements”. In: *American Journal of Neuroradiology* 32.3, pp. 548–555.

Finn, Emily S, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable (2015). “Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity”. In: *Nature Neuroscience* 18.11, p. 1664.

Xu, Ting, Alexander Opitz, R Cameron Craddock, Margaret J Wright, Xi-Nian Zuo, and Michael P Milham (2016). “Assessing variations in areal organization for the intrinsic brain: from fingerprints to reliability”. In: *Cerebral Cortex* 26.11, pp. 4192–4211.

Bridgeford, Eric W, Shangsi Wang, Zhi Yang, Zeyi Wang, Ting Xu, Cameron Craddock, Gregory Kiar, William Gray-Roncal, Carey E Priebe, Brian Caffo, et al. (2019). “Optimal experimental design for big data: applications in brain imaging”. In: *bioRxiv*, p. 802629.

Provost, Serge B and Edmund M Rudiuk (1996). “The exact distribution of indefinite quadratic forms in noncentral normal vectors”. In: *Annals of the Institute of Statistical Mathematics* 48.2, pp. 381–394.

Yuan, Ke-Hai and Peter M Bentler (2010). “Two simple approximations to the distributions of quadratic forms”. In: *British Journal of Mathematical and Statistical Psychology* 63.2, pp. 273–291.

Sylvester, James Joseph (1852). “XIX. A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 4.23, pp. 138–142.

Lieb, Elliott H and Heinz Siedentop (1991). “Convexity and concavity of eigenvalue sums”. In: *Journal of Statistical Physics* 63.5-6, pp. 811–816.

Ghosh, BK (1973). “Some monotonicity theorems for χ2, F and t distributions with applications”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 35.3, pp. 480–492.

Chapter 3

Permutation Tests

3.1 General Framework of Reliability Tests

In fingerprinting studies, one attempts to match a subject's first session image with their second, in a blinded fashion, in a group of twice measured subjects. The number or percentage of correct matches is reported as the statistic. Under the hypothesis of exchangeability of the subject labels, the number or percent of matches can be analyzed relative to a reference permutation distribution to establish evidence of reliability, or lack thereof. Such practice is bolstered by novel applications (Finn et al., 2015; Airan et al., 2016) and large scale replication studies (Zuo et al., 2014; Van Essen et al., 2013) as well as general interest in fMRI reproducibility (Choe et al., 2017; Poldrack and Poline, 2015; Choe et al., 2015; Landman et al., 2011; Griffanti et al., 2016; Shou et al., 2013; Aron, Gluck, and Poldrack, 2006). Despite the simplicity and increasing popularity of such matching and inference procedures, the soundness of the statistical tests, the power, and the factors impacting the test are unstudied.

Essentially, similar permutations tests for extreme large values can be conducted using any one of the reliability statistics described in Section 2.1. Such permutation tests are constructed based on a distributional exchangeability null hypothesis on the permuted statistics. That is, under the null, the distribution of the reliability statistic is assumed to be invariant against some permutation of the subject labels. For repeated measurements with multiple time points, the subject labels are permuted within each of the time points.

In practice, non-parametric approximations of the test statistic distribution under the null can be achieved by actually permuting the observed sample. In fact, to perform the test, Monte Carlo resampling (Good, 2013) is used to reduce the computational burden of looping over each of the possible permutations, which can be up to $(n!)^s$ scenarios for n subjects measured at s time points. Exploiting the approximated null distribution, the test rejects the null when the observed value of the reliability statistic is more extreme than one would have expected under the null given significance level.

The additive noise setting (2.8) for the general model of repeated measurements, $\boldsymbol{x}_{it} \perp \boldsymbol{v}_i$ implies $\boldsymbol{x}_{it} = \boldsymbol{\epsilon}_{it}$, guarantees exchangeability of any reliability statistics defined in the previous sections. Thus, if the associated model is correctly specified, rejection in the permutation test using any of the aforementioned statistics implies the existence of dependence between a subject's unobserved true subject-specific effect, $\boldsymbol{v}_i$, and its observed measurement, $\boldsymbol{x}_{it}$. Therefore, permutation tests with the weaker null of exchangeability are conducted for the purpose of confirming reliability. The resulting test significance provides evidence against no reliability, where the measurement reveals no information on differences in subject specific effects.

The general properties of these reliability statistics under different model settings other than the ANOVA model may be less mathematically clear. In Section 3.2 we present numerical results, including deviations from the ANOVA model. In Section 3.3, we discuss the potentials of parametric approximations for fingerprinting and rank sums.

3.2 Simulation on Hypothesis Testing Power

3.2.1 Univariate ANOVA Simulations

We first evaluate the estimation and testing power performance under the ANOVA model (2.1) or when its Gaussian assumptions are violated. $t = 1, 2$. $\sigma^2 = 5, \sigma_\mu^2 = 3$. The number of subjects, n, ranges from 5 to 40.

In addition to the correct Gaussian model, consider the following lognormal misspecification:

$$\mu_i \overset{d}{\sim} \text{log-}\mathcal{N}\left(0, \sigma_\mu^2\right);\ \log(\mu_i) \overset{d}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right),$$

$$e_{it} \overset{d}{\sim} \text{log-}\mathcal{N}\left(0, \sigma^2\right);\ \log(e_{it}) \overset{d}{\sim} \mathcal{N}\left(0, \sigma^2\right),$$

where we still define $ICC = \text{var}(\mu_i)/(\text{var}(\mu_i) + \text{var}(e_{it}))$, but now $\text{var}(e_{it}) = (\exp(\sigma^2) - 1) \cdot \exp(\sigma^2)$. Note that the relation between discriminability and ICC does not hold in this setting.

For $1,000$ iterations, estimates of discriminability (using $\tilde{D}$ in the Equation 2.15), the rank sum estimator ($\hat{D}_{rs}$ in Equation 2.17), estimations of ICC using one-way ANOVA, estimations of the fingerprint index (using $\hat{F}_{index}$ in the

Equation 2.6) were recorded and compared to their theoretical true values (for discriminability and ICC) or its simulated average value (for the fingerprint index, with $10,000$ simulations).

Within each iteration, we also conducted permutation tests against exchangeability, each with 200 Monte Carlo simulations, using the previously mentioned estimators. F-tests using the ICC F-statistics were also conducted. The proportion of rejections (power curves) by iterations were plotted.

When the parametric assumption is satisfied, all estimators are distributed around their true values (Figure 3.1). Note that the distribution of the fingerprint index is skewed. In addition, a higher fingerprinting index estimation with fewer subjects does not imply better reliability, compared the lower estimation with more subjects. Of note, the true ICC and discriminability remain constant as sample size increases in the simulation setup. Thus, insofar as these measures summarize reliability, this emphasizes that the fingerprint index is not directly comparable across sample sizes. In terms of the testing power, as we expected, tests using statistics associated with the ICC produce higher power, as the Gaussian model is correctly specified. The discriminability estimator using the whole combined ranking matrix shows slight advantage in power compared to the rank sum estimator, which only uses rank sums within a submatrix of the combined distance matrix. Lastly, switching to fingerprinting results in a loss in testing power.

We repeated the simulation in an otherwise similar setting where normality does not hold: $\text{var}(\mu_i) = (\exp(\sigma_\mu^2) - 1) \cdot \exp(\sigma_\mu^2) \approx 383$, $\text{var}(e_{ij}) = (\exp(\sigma^2) - 1) \cdot \exp(\sigma^2) \approx 21878$, and ICC is around 0.017. Because of model

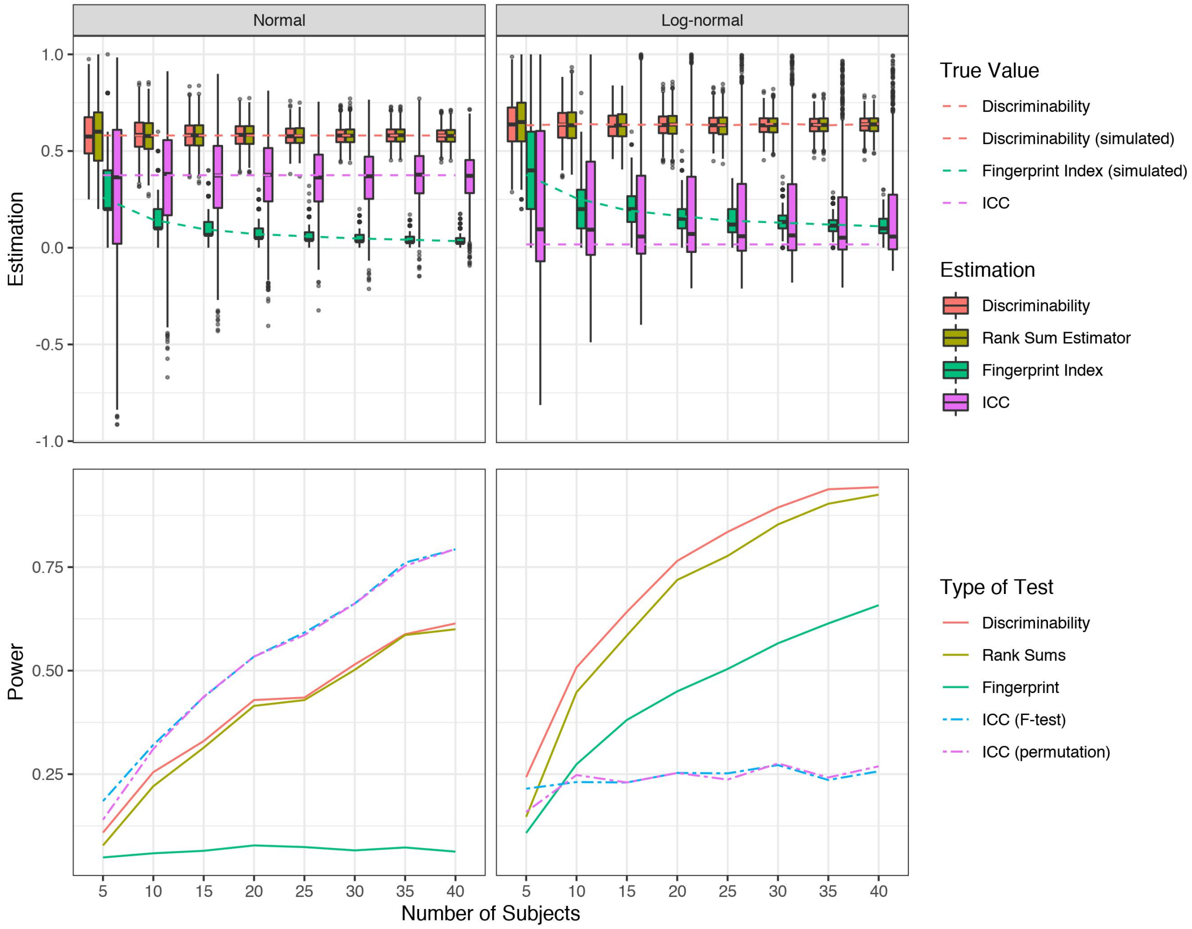

Figure 3.1: ANOVA simulations when the Gaussian assumption is satisfied (left) or violated with logarithm transformations (right). Simulated distributions of estimators are plotted on the top, including the discriminability estimation (using the estimator $\tilde{D}$ or the rank sum version $\hat{D}_{rs}$), the fingerprint index estimation, and the ICC estimation. Simulated permutation test powers are plotted on the bottom, where solid lines and dotted lines represent nonparametric and parametric statistics, respectively. $\sigma^2 = 5$. $\sigma_\mu^2 = 3$. n ranges from 5 to 40. $1,000$ iterations in total. See Section 3.2.1.

misspecification, ICC is overestimated with relatively large variation. As for testing power, the discriminability estimator, rank sum and the fingerprint index estimator outperform, due to their nonparametric framework, which does not rely on Gaussian assumptions. $\tilde{D}$ again has higher power than $\hat{D}_{rs}$ for including more ranking information. $\hat{F}_{index}$ has a loss in power over disciminability or rank sums, but is now better than the tests using parametric estimations of ICC or F-statistics.

3.2.2 MANOVA Simulations

Next, we consider the MANOVA model (2.2) and a similar misspecification with element-wise log-transformations on the subject mean vectors, $\boldsymbol{\mu}_i$, and the noise vectors, $\boldsymbol{e}_{it}$. $t = 1, 2$. n ranges from 5 to 40.

We simulated data with $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{Q}$, $\boldsymbol{\Sigma}_{\boldsymbol{\mu}} = \sigma_\mu^2 \boldsymbol{Q}$, and $\boldsymbol{Q} = \boldsymbol{I}(1-\rho) + \mathbf{1}\mathbf{1}^t \rho$ (an $l \times l$ exchangeable correlation matrix, with off diagonals ρ). Then I2C2 becomes:

$$\frac{\operatorname{tr} \boldsymbol{\Sigma}_\mu}{\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{\mu}} + \operatorname{tr} \boldsymbol{\Sigma}} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2}.$$

Let $\sigma^2 = 5, \sigma_\mu^2 = 3, \rho = 0.5, l = 10$. For $1,000$ iterations, the estimations and the permutation test (each performed with 200 Monte Carlo simulations) power were compared for discriminability, the rank sum estimator of discriminability, the fingerprint index, the sample ICC, $\hat{\lambda}$, calculated with the first principal components from the measurements, and I2C2.

When the Gaussin assumption is satisfied, I2C2 outperforms other statistics, and most statistics produce higher testing power compared to the fingerprint index (by a large margin Figure 3.2). Note that the strategy of conducting

PCA before ICC also shows advantage over discriminability in power when the sample size is as small as 5, but power converges with larger sample sizes.

When normality is violated, the nonparametric statistics (discriminability, rank sums, and fingerprinting) outperform the parametric methods in power with any sample sizes greater than 10. The discriminability estimator provides the best power under the multivariate lognormal assumptions.

3.2.3 Batch Effects

Consider the ANOVA model (2.1) where each subject is remeasured for s times, $s > 2$. We evaluate two types of batch effects, mean shifts and scaling factors (Johnson, Li, and Rabinovic, 2007).

For the mean shifts, we replace the subject means, μ_i's, with the batch specific means μ_{it}'s defined as:

$$\mu_{i1} \overset{d}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right),$$

$$\mu_{it} = \mu_{i1} + t, t = 2, \ldots, s.$$

Without loss of generality, consider the first batch as a reference batch, where μ_{i1}'s follow the same distribution as the previous μ_i's. For the t-th batch, there exists a mean shift, t, from the reference batch for all subjects. The scaling effects are applied on the noise variances as:

$$e_{i1} \overset{d}{\sim} \mathcal{N}\left(0, \sigma^2\right),$$

$$e_{it} \overset{d}{\sim} \mathcal{N}\left(0, t\sigma^2\right), t = 2, \ldots, s.$$

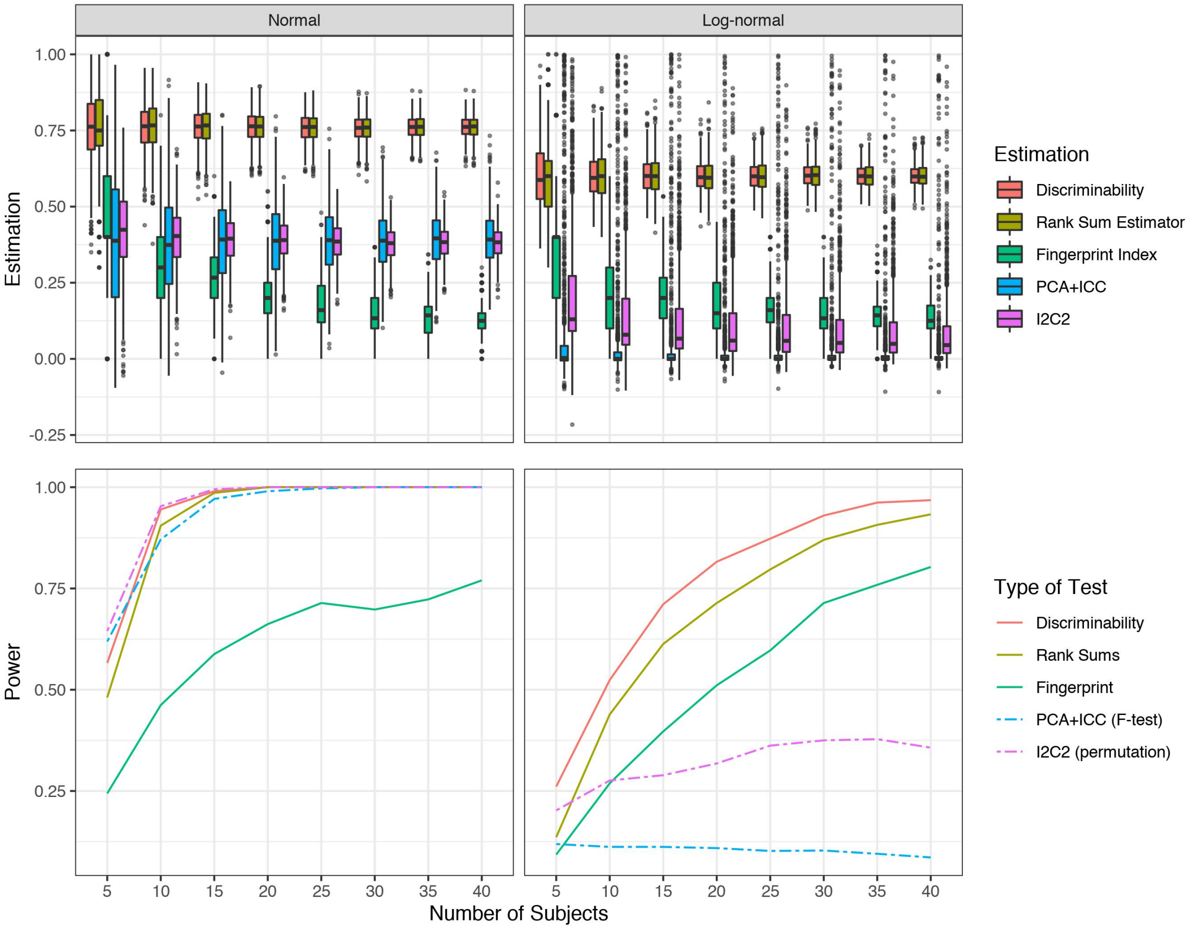

Figure 3.2: MANOVA simulations when the Gaussian assumption is satisfied (left) or violated with element-wise logarithm transformations (right). Simulated distributions of estimators are plotted on the top, including the discriminability estimation (using the estimator $\tilde{D}$ or the rank sum version $\hat{D}_{rs}$), the fingerprint index, and the I2C2. Simulated permutation test powers are plotted on the bottom, where solid lines and dotted lines represent nonparametric and parametric statistics, respectively. $\sigma^2 = 5$. $\sigma_\mu^2 = 3, \rho = 0.5, l = 10$. n ranges from 5 to 40. $1,000$ iterations in total. See Section 3.2.2.

At each time point $t = 2, \ldots, s$, two strategies of evaluating the reliability were considered. First, we first considered the first and the t-th batches (*two stages*) and secondly we used all measurements up to the t-th time point.

For the first strategy, the discriminability estimator and its rank sum alternative are used without modification. For the second strategy, we calculated $\tilde{D}$ as usual (*all stages*) but averaged out the $\hat{D}_{rs}$'s over all the pairs of distinct time points (*all pairs*), as described in Section (2.2.4). In addition, we also averaged $\tilde{D}$ and $\hat{D}_{rs}$, respectively, for all the pairs of time points between the first and the rest, up to the t-th (*all pairs from initial*). In total, six types of estimators were considered.

We simulated $s = 15$ batches in total with $\sigma^2 = 3, \sigma_\mu^2 = 5$ and let the number of subjects, n, range from 5 to 40. For $1,000$ iterations, the estimations and the permutation test (each with 200 Monte Carlo iterations) power of the six estimators described above are plotted.

For the mean shift only batch effects, the rank sum estimator outperforms discriminability in power with the highest power achieved using all time point pairs (Figure 3.3). The estimation from rank sums is also closer to the batch-effect-free true discriminability, 0.625. The rank sum method may benefit from the fact that, whenever $t = t''$, it avoids averaging over indicators

$$\mathbb{I}_{\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\}} = \mathbb{I}_{\{|(t-t')+(e_{it}-e_{it'})| < |(\mu_{it}-\mu_{i't})+(e_{it}-e_{i't})|\}},$$

where the batch difference, $(t - t')$, if larger enough, may force the indicator to be 0 with high probability, regardless of the true batch-effect-free discriminability level. For example, for the *all pairs from initial* scenario, rank sums

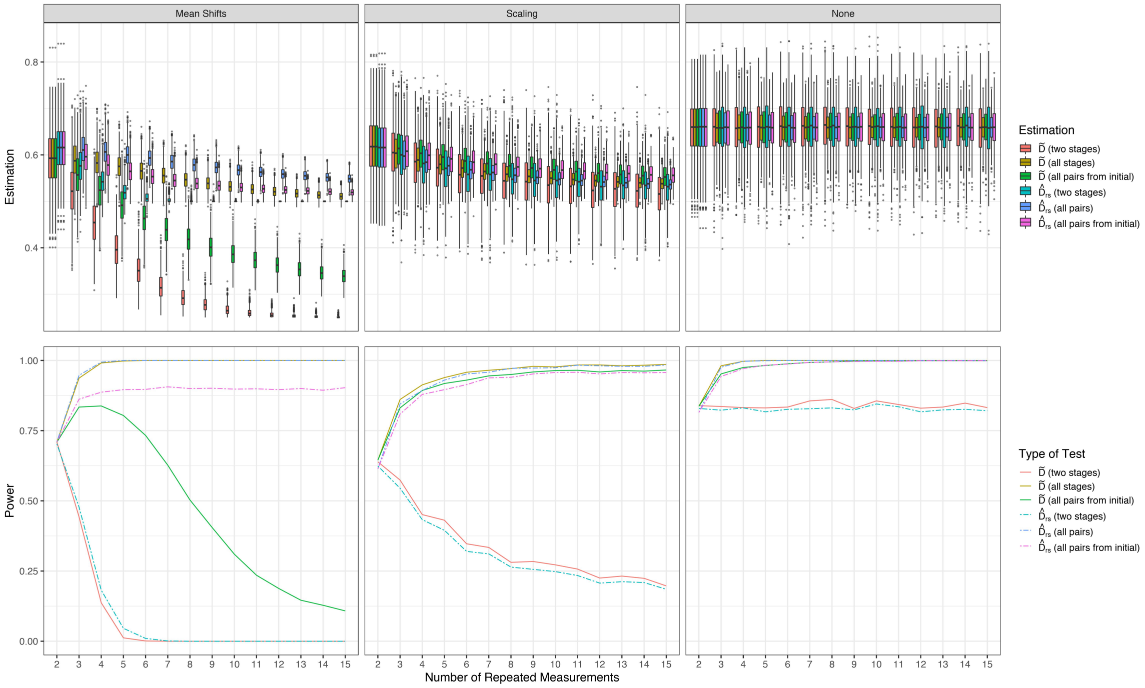

Figure 3.3: Simulations for mean shifts (left), scaling (middle), and no batch effect (right). Simulated distributions of the discriminability estimators ($\tilde{D}$ and $\hat{D}_{rs}$) are plotted on the top, including six estimation strategies (Section 3.2.3). Simulated permutation test powers are plotted on the bottom, where solid lines and dotted lines represent the discriminability ($\tilde{D}$) based and the rank sum ($\hat{D}_{rs}$) based strategies, respectively. $s = 15, \sigma^2 = 3, \sigma_\mu^2 = 5$. The number of subjects is $n = 20$. $1,000$ iterations are conducted for each scenario.

outperform discriminability by a huge margin, since batch differences become larger when later batches are compared to the reference batch.

For the scaling only batch effects, discriminability now outperforms rank sums, regardless of the strategy used. (Using all time points produces the highest power.) This is similar to the case with no batch effects, where having more repeated measurements increases testing power, and the advantage of discriminability over rank sums and the advantage of using all time points are attained.

3.2.4 Discussion

One of our major findings is the relationship between discriminability, ICC or I2C2 on the population level. Note this is different from the non-decreasing relation between ICC estimation and the F statistic, which guarantees the same ordering and power in the permutation test. The fact that ICC and I2C2 may still have higher power when parametric assumptions are satisfied hints the potential of improving the current discriminability estimation. Another potential improvement is the approximation (2.19) of the weighted sum of χ^2's, as it tends to underestimate D with larger within measurement correlations (Figure 3.4). But, even with the current approximation the error is within 0.1 and the non-decreasing relation holds true in the simulations with larger ρ values. Other limitation includes the lack of analysis for the fixed effect, while we focus on the random effect models for cleaner illustration. Lastly, in practice dissimilarity (pseudo)distances such as one minus Pearson correlation may be applied instead of the Euclidean distance; this does not impact testing results if measurements are standardized with mean 0 and variance 1, and if measurements are non-negatively correlated.

On the other hand, the relation we found with rank sums and fingerprinting is between the testing statistics; based on the simulations we argue that the discriminability should be preferred in practice unless there exist concerns about mean shift batch effects.

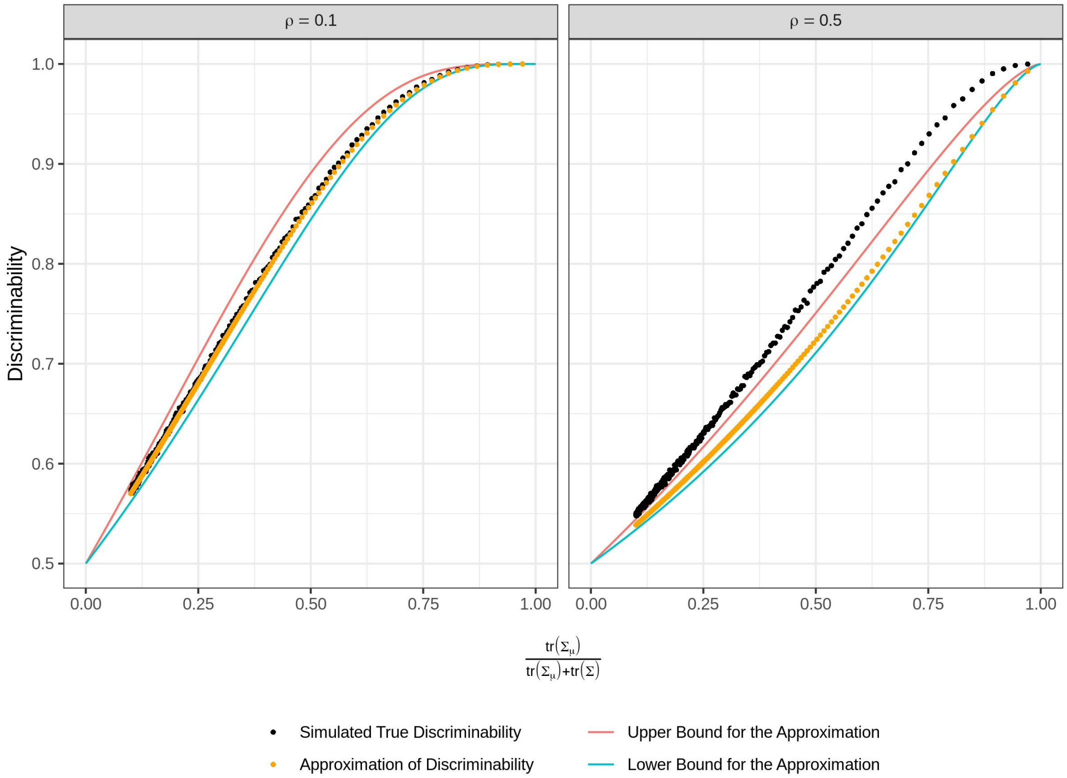

Figure 3.4: Relation between discriminability and I2C2 with smaller ($\rho = 0.1$, left) or larger ($\rho = 0.5$, right) within measurement correlation. The Gaussian MANOVA model in Section 3.2.2 is assumed with $l = 10, n = 20, s = 2$. Covariance matrices, $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_\mu$, are proportional to a matrix with diagonals being 1 and off-diagonals being ρ. Small circles are the simulated (1, 000 iterations) true discriminability with $\sigma_\mu^2 = 100$ and σ^2 ranging from 3 to 300. This shows error of the approximation (2.19) is within 0.1 and the non-decreasing relation holds true even with larger ρ value.

3.3 Approximations

3.3.1 Fingerprinting

We will demonstrate that, including a slight deviation (matching with replacement) from the type of fingerprinting described in Section 3.1 (matching without replacement), the generated permutation distribution can be well approximated by a Poisson(1). Throughout this section, we will focus on the number of total correct matches (2.5), T_n, as the test statistic.

Before getting to the details of approximating the permutation distributions, we need to define the matching procedure and the exchangeability hypothesis explicitly.

Matching Mechanics

The most common form of matching tries to match one measurement, say the second, to the first. Let $\delta_{ij} = \delta(\boldsymbol{x}_{i1}, \boldsymbol{x}_{j2})$ be the distance between subject i on occasion 1 and subject j on occasion 2 given observation $\boldsymbol{x}_{it}$, $i = 1, \cdots, n$, $t = 1, 2$. Let m_i be the subject label of the best match for subject i. Of course, the term "best" is in reference to a matching strategy and we will use m_i generically regardless of which strategy was used. As an example strategy, consider, $m_i = \operatorname{argmin}_j \delta_{ij}$. Under this scheme, subjects on occasion 2 can be matched multiple times if they are the best match for more than one subject. Because of this, we call this strategy **matching with replacement** (or MWR).

A matrix form is an often preferable method to represent the data. Let $\boldsymbol{B}$ be a matrix with a 1 in position i, j if subject i on sampling occasion 1

Table 3.1: Example resampling matrix from matching with replacement. Here the statistic value is 3.

	Time 2				
Time 1	1	2	3	4	Total
1	0	1	0	0	1
2	0	1	0	0	1
3	0	0	1	0	1
4	0	0	0	1	1
Total	0	2	1	1	4

is best matched with subject j on occasion 2. That is, $\boldsymbol{B} = [b_{ij}]_{i,j}$ where $b_{ij} = I\{m_i = j\}$ where $I\{a = j\}$ is an indicator that returns 1 if $a = j$ and 0 otherwise. It is interesting to note that matrices of these forms are exactly bootstrap resampling matrices. Table 3.1 gives an example for $n = 4$. Recall that the first row, $(0, 1, 0, 0)$, implies that among the occasion 2 measurements, subject 2's is the best match for the occasion 1 measurement of subject 1. The second row, $(0, 1, 0, 0)$, implies subject 2's occasion 2 measurement is correctly matched to the subject's occasion 1 measurement. Thus, in this case, subject 2's occasion 2 measurements are matched twice, for both subject 1 and subject 2 on occasion 1. The standard statistic measurement agreement is the number of correct matches (the trace of $\boldsymbol{B}$, $tr(\boldsymbol{B})$). In our example, the statistic value would be 3.

Alternatively, one could **match without replacement** (or MWOR). That is, find the best permutation of subjects on the second occasion to match up with the first. As an example, let $\boldsymbol{\Gamma}$ be the collection of $n \times 1$ vectors of permutations of the integers $1, \ldots, n$. Then consider

$$\boldsymbol{M} = (m_1 \ldots m_n)' = \text{argmin}_{\boldsymbol{\pi} \in \boldsymbol{\Gamma}} \sum_{i=1}^{n} \delta_{i\pi_i}.$$

Table 3.2: Example resampling matrix from matching without replacement. Here the statistic value is 2.

	Time 2				
Time 1	1	2	3	4	Total
1	0	1	0	0	1
2	1	0	0	0	1
3	0	0	1	0	1
4	0	0	0	1	1
Total	1	1	1	1	4

The Hungarian algorithm allows that this optimization can be performed in polynomial time (Pentico, 2007). This is a harder optimization problem, because the optimization is conducted simultaneously and not sequentially, as in the matching with replacement. It is possible to have a non-unique best match. However, given the size and noise of neuroimaging, data the best match is usually unique for the best permutation. If this result is put into a matrix with $b_{ij} = I\{m_i = j\}$, then $\boldsymbol{B}$ is a permutation matrix (a 0,1 matrix with row and column totals all equal to one). Again, the relevant statistic is the trace. Table 3.2 shows an example with $n = 4$ that has statistic value equal to 2.

Inference

Permutation-based inference is the norm in this area. One typically repeatedly permutes the subject labels at occasion 2 and re-performs the matching at each iteration to obtain a null distribution. Given the dimension of the characteristics being matched on, it is typical for no ties to exist in the δ_{ij}, so that the best matches are all unique at each iteration.

This permutation test is motivated by an implicit exchangeability assumption. That is, the underlying null distribution of the statistic is the same for any permutation. Alternatively, the null hypotheses can be developed under stronger iid sampling assumptions.

One of our main results in this section is to show that under nearly all sampling strategies the null distribution of the test statistic is well approximated by a Poisson with a mean of one. The implication of this result is both simple and widespread: the use of the permutation test is unnecessary, as the null hypothesis will be rejected under the same conditions, when $tr(\boldsymbol{B})$ is larger than 3 or 4, say, depending on the desired Type I error rate. Thus, computation time and costs can be systematically reduced using this simple, slightly unexpected, but powerful statistical result. Below we provide details on the implicit assumptions associated with the permutation test and the interpretation given these results.

Exchangeability and the Null Hypothesis

A difficult task in permutation tests is strictly defining the null hypothesis under consideration. We focus on exchangeability as perhaps the most general and useful form of the null hypothesis in this setting. This hypothesis is defined as irrelevance of the labels in the form of an identical distribution being obtained under permutations. We formalize the concepts below.

Recall that $\boldsymbol{x}_{it}$ is the l dimensional feature vector of subject i on occasion t where $i = 1, \cdots, n$ and $t = 1, 2$. Denote $\boldsymbol{X}_{(t)}$ as the $l \times n$ data matrix for occasion $t = 1, 2$ with columns $\boldsymbol{x}_{1t}, \cdots, \boldsymbol{x}_{nt}$. Let $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}]$ be the

$l \times 2n$ combined data matrix with columns $\boldsymbol{x}_{11}, \boldsymbol{x}_{21}, \cdots, \boldsymbol{x}_{n1}, \boldsymbol{x}_{12}, \boldsymbol{x}_{22}, \cdots, \boldsymbol{x}_{n2}$. Let $\boldsymbol{X} = \boldsymbol{x}$ be the observed data. Recall also, in MWR, the best match for subject i's occasion 1 image is $m_i = \operatorname{argmin}_j \delta(\boldsymbol{x}_{i1}, \boldsymbol{x}_{j2})$. In MWOR, the best match for subject i's occasion 1 image is m_i, where $\boldsymbol{M} = (m_1, \cdots, m_n)' = \operatorname{argmin}_{\boldsymbol{\pi} \in \boldsymbol{\Gamma}} \sum_{i=1}^{n} \delta_{i\pi_i} = \operatorname{argmin}_{\boldsymbol{\pi} \in \boldsymbol{\Gamma}} \sum_{i=1}^{n} d(\boldsymbol{x}_{i1}, \boldsymbol{x}_{\pi_i 2})$, $\boldsymbol{\Gamma}$ is the collection of permutation vectors of $(1, \cdots, n)'$. In both scenarios, the test statistic is defined as $T(\boldsymbol{x}) = \sum_{i=1}^{n} I\{m_i = i\}$, the number of correct matches.

The exchangeable null hypothesis, H_E, is defined as the invariant distribution of test statistic when permuting the labels of occasion 2 images. That is,

$$P\{T(\boldsymbol{X}) = t\} = P\{T(\boldsymbol{X}_{\boldsymbol{P}}) = t\}$$

for all $t \in \{0, \cdots, n\}$, $\boldsymbol{P} \in \mathcal{P}$, where $\mathcal{P}$ is the collection of $n \times n$ permutation matrices, $\boldsymbol{X}_{\boldsymbol{P}} = \{\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}\boldsymbol{P}\}$ is the $n \times 2$ data matrix obtained after permuting occasion 2 labels.

Exact Permutation Tests

Following Hoeffding, 1952, under H_E, the permutation test can be executed to have an exact α type I error rate if a randomized test function is defined as:

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) > T^{(k)}(\boldsymbol{x}) \\ a(\boldsymbol{x}), & T(\boldsymbol{x}) = T^{(k)}(\boldsymbol{x}) \\ 0, & T(\boldsymbol{x}) < T^{(k)}(\boldsymbol{x}) \end{cases}.$$

Here, $\phi(\boldsymbol{x})$ is the probability of rejecting the null given observation $\boldsymbol{X} = \boldsymbol{x}$. The variables, $T^{(k)}(\boldsymbol{x})$, for $k = 1, \ldots, n!$ is the ordered list of all permuted test statistics. The index k determines the closest quantile less than or equal to

α of the permuted test statistics level, i.e. $k = n! - \lfloor n!\alpha \rfloor$ where $\lfloor \cdot \rfloor$ is the floor function. This is equivalently, the inverse, $\hat{F}^{-1}(1-\alpha)$, of the distribution function of the permuted test statistics:

$$\hat{F}(t) = \frac{1}{n!} \sum_{\boldsymbol{P} \in \mathcal{P}} I\{T(\boldsymbol{x}_{\boldsymbol{P}}) \leq t\}.$$

A randomized test with exact level α occurs if one rejects H_E when $\phi(\boldsymbol{x})$ is 1, i.e. the test statistic lies strictly in the upper α area of the permutation distribution, fails to reject when $\phi(\boldsymbol{x})$ is 0, and rejects with probability $a(\boldsymbol{x})$ otherwise. In the latter case, a uniform random variable is simulated and the test is rejected if it is less than $a(\boldsymbol{x})$.

Hoeffding, 1952 showed that $a(\boldsymbol{x})$ defined as $\{n!\alpha - M^{+}(\boldsymbol{x})\}/M^{0}(\boldsymbol{x})$ yields an α level randomized test. Here, $M^{+}(\boldsymbol{x})$ and $M^{0}(\boldsymbol{x})$ are the counts of permuted statistics larger than or equal to $T^{(k)}$, respectively. These are formally defined as: $M^{+}(\boldsymbol{x}) = |\{j \in \{1, \cdots, n!\} : T^{(j)}(\boldsymbol{x}) > T^{(k)}(\boldsymbol{w})\}|$ and $M^{0}(\boldsymbol{x}) = |\{j \in \{1, \cdots, n!\} : T^{(j)}(\boldsymbol{x}) = T^{(k)}(\boldsymbol{x})\}|$.

Since having an ancillary coin flip determine rejection is not desirable, the more conservative non-randomized test simply uses the non-randomized test function:

$$\phi'(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) > T^{(k)}(\boldsymbol{x}) \\ 0, & T(\boldsymbol{x}) \leq T^{(k)}(\boldsymbol{x}) \end{cases}.$$

This yields a test with a type I error rate guaranteed to be less than α, though cannot yield an exact α level test, except in rare cases, such as when $n!\alpha$ is an integer.

Note that with the matrix representation we have $T(\boldsymbol{x}) = tr(\boldsymbol{B})$ as the total number of correct matches and hence $T(\boldsymbol{x}_{\boldsymbol{P}}) = tr(\boldsymbol{BP}) = tr(\boldsymbol{PB})$ is the total number of correct matches after permuting occasion 2 labels according to some $\boldsymbol{P} \in \mathcal{P}$. Therefore an alternative expression for permutation distribution function is:

$$\hat{F}(t) = \frac{1}{n!} \sum_{\boldsymbol{P} \in \mathcal{P}} I\{tr(\boldsymbol{PB}) \leq t\},$$

the CDF from the traces of all the row permutations of $\boldsymbol{B}$.

Thus, the CDF arising from placing equal (discrete uniform) probability on all permutations is derived equivalently from permuting either the occasion 1 or occasion 2 labels. Suppose $\boldsymbol{\Pi}$ is uniformly distributed over $\mathcal{P}$. Then $\hat{F}(t) = P\{tr(\boldsymbol{\Pi B}) = t\}$.

Poisson Approximation for MWOR

In matching without replacement, each occasion 1 image is matched to a distinct occasion 2 image. This implies each column and row of $\boldsymbol{B}$ sums to 1, as $\boldsymbol{B}$ is a permutation matrix, since the vector of matches is a permuted version of $(1, \cdots, n)'$. In this case, permuting occasion 1 labels and then calculating $tr(\boldsymbol{\Pi B})$ is equivalent to shuffling a batch of ordered cards and counting the number of cards still in its original order, which follows Montmort's matching distribution (Barton, 1958). Hence

$$P\{tr(\boldsymbol{\Pi B}) = t\} = \frac{1}{t!} \sum_{j=0}^{n-t} \frac{(-1)^j}{j!}.$$

As n goes to infinity, for any fixed t, $P\{tr(\boldsymbol{\Pi B}) = t\} \rightarrow 1/(t!) \cdot \sum_{j=0}^{\infty} (-1)^j/(j!) = \exp(-1)/(t!)$ and $tr(\boldsymbol{\Pi B})$ converges to a Poisson(1) distribution.

This is the distribution of correct matches under permutations, famously originally derived in a letter between Montmort and Nicolaus Bernoulli. This distribution and matching setting is often used in probability courses to illustrate the law of total probability. It is interesting to note that the Poisson approximation has an upper 95^{th} percentile of 3, 99^{th} percentile of 4 and 99.9^{th} percentile of 5. Therefore, relatively few matches need be made to reject this null hypothesis and that number is fairly static with n, since convergence occurs quite quickly. The reason the p-value is robust to large changes in n is because although the number of possible matches increases with n, the probability of a match decreases in a balanced way.

Poisson Approximation for MWR

Suppose we observed combined data matrix $\boldsymbol{X} = \boldsymbol{x}$ and its representation matrix $\boldsymbol{B}$ in a matching with replacement process. Each occasion 1 image will be matched to exactly one occasion 2 image whereas some occasion 2 images may get matched multiple times and some occasion 2 images may not get matched at all. In this case the sum of any row of $\boldsymbol{B}$ will still be 1 but column sums of $\boldsymbol{B}$ can vary.

Without loss of generality, suppose only the column sums of first k columns of $\boldsymbol{B}$ are nonzero. Denote the column sums as $c_1, \cdots, c_k$. Then $\sum_{i=1}^{k} c_i = n$. For $h \subset \{1, \cdots, k\}$, denote the size of h as $|h|$. By the inclusion-exclusion formula we have (see Appendix in Section 3.3.1)

$$P\{T(\boldsymbol{\Pi B}) = t\} = \sum_{u \in \{h \subset \{1,\cdots,k\}:|h|=t\}} \sum_{s=0}^{k-t} (-1)^s \sum_{v \in \{J \subset \{1,\cdots,k\} \setminus u:|J|=s\}} \left(\prod_{i \in u} c_i \right) \left(\prod_{j \in v} c_j \right) \frac{(n-t-s)!}{n!}.$$

When $k = n$ and $c_1 = \cdots = c_n = 1$, the distribution coincides with the matching without replacement distribution:

$$P\{T(\mathbf{\Pi B}) = t\} = \binom{n}{t}\sum_{s=0}^{n-t}(-1)^s\binom{n-t}{s}\frac{(n-t-s)!}{n!},$$

$$= \sum_{s=0}^{n-t}(-1)^s\frac{1}{s!t!}.$$

Via Stein-Chen's method (see Appendix in Section 3.3.1), the total variation between $T(\Pi B)$ and a Poisson(1) for matching with replacement is:

$$d_{TV}\left\{T(\mathbf{\Pi B}), \text{Poisson}(1)\right\} = \frac{1}{n-1} + \frac{(n-2)}{n^2(n-1)}\sum_{i=1}^{k} c_i^2,$$

$$\leq \frac{1}{n-1} + \frac{n-2}{n-1}\frac{\sum_{i=1}^{k} Cc_i}{n^2},$$

$$= \frac{1}{n-1} + \frac{n-2}{n-1}\frac{C}{n},$$

where C is the number of matches of the occasion 2 image with the most matches, that is, $C = \max_{i\in\{1,\cdots,k\}}\ c_i$. Thus the permutation distribution will be approximated by a Poisson(1) if C is small and n is large. Specifically, $C/n \to 0$ as $n \to \infty$ is sufficient for the distribution of $T(\mathbf{\Pi B})$ to converge to a Poisson(1).

Appendix: Inclusion-Exclusion Formula for MWR

For matching with replacement, the matching matrix **B** will have one 1 on each row but $rank(\mathbf{B})$ could be smaller than n. For example, if the first two occasion 1 images are all matched to the first occasion 2 image. Than by permuting the occasion 1 labels, it is impossible to have $b_{22} = 1$.

Assume $rank(B) = k$, which means k of the occasion 2 images are matched, each with one or more occasion 1 images. We calculate $p_1, p_2, ..., p_k$ the proportion of occasion 1 images that are matched to those occasion 2 images. Then $\sum_{i=1}^{k} p_i = 1$. Note that matching without replacement becomes a special case with $k = n$ and $p_i = 1/n$ for all i.

Without loss of generality, suppose only the column sums of first k columns of B are nonzero. Denote the column sums as $c_1, \cdots, c_k$. Then $\sum_{i=1}^{k} c_i = n$.

Recall that $\mathbf{\Pi}$ has a discrete uniform distribution over $\mathcal{P}$. Let $A_i = \{c_i = i\}$ be the event that subject i gets the correct match after permuting the occasion 1 labels. For $h \subset \{1, \cdots, k\}$, let $B_h = \cap_{i \in h} A_i = \{\forall\, i \in h, c_i = i\}$ be the event that all the subjects within h get correct matches and denote the size of h as $|h|$. By the inclusion-exclusion formula we have

$$\begin{aligned}
P(T(\mathbf{\Pi B}) = 0) &= 1 - P(\cup_{i=1}^{k} A_i) \\
&= 1 - \sum_{l=1}^{k} (-1)^{l-1} \sum_{v \in \{J \subset \{1, \cdots, k\} : |J| = l\}} P(\cap_{j \in v} A_j) \\
&= 1 - \sum_{l=1}^{k} (-1)^{l-1} \sum_{v \in \{J \subset \{1, \cdots, k\} : |J| = l\}} (\prod_{j \in v} c_j) \frac{(n-l)!}{n!} \\
&= \sum_{l=0}^{k} (-1)^{l} \sum_{v \in \{J \subset \{1, \cdots, k\} : |J| = l\}} (\prod_{j \in v} c_j) \frac{(n-l)!}{n!}
\end{aligned}$$

Furthermore, let $\mathbf{B}^{(-h)}$ be a copy of the matching matrix $\mathbf{B}$ with the rows and columns having their orders in h deleted and denote $\mathbf{\Pi}_l$ as a random matrix with discrete uniform distribution over the collection of $l \times l$ permutation matrices. We have

$$\begin{aligned}
P(T(\mathbf{\Pi B}) = t) &= \sum_{u \in \{h \subset \{1,\cdots,k\}:|h|=t\}} P(B_u) P(T(\mathbf{\Pi B}) = t | B_u) \\
&= \sum_{u \in \{h \subset \{1,\cdots,k\}:|h|=t\}} P(B_u) P(T(\mathbf{\Pi}_{n-t}\mathbf{B}^{(-u)}) = 0) \\
&= \sum_{u \in \{h \subset \{1,\cdots,k\}:|h|=t\}} \left\{ (\prod_{i \in u} c_i) \frac{(n-t)!}{n!} \right\} \\
&\quad \left\{ \sum_{s=0}^{k-t} (-1)^s \sum_{v \in \{J \subset \{1,\cdots,k\} \setminus u : |J|=s\}} (\prod_{j \in v} c_j) \frac{(n-t-s)!}{(n-t)!} \right\} \\
&= \sum_{u \in \{h \subset \{1,\cdots,k\}:|h|=t\}} \sum_{s=0}^{k-t} (-1)^s \sum_{v \in \{J \subset \{1,\cdots,k\} \setminus u : |J|=s\}} \\
&\quad (\prod_{i \in u} c_i)(\prod_{j \in v} c_j) \frac{(n-t-s)!}{n!}.
\end{aligned}$$

When $k = n$ and $c_1 = \cdots = c_n = 1$, the distribution coincides with the matching without replacement distribution

$$\begin{aligned}
P(T(\mathbf{\Pi B}) = t) &= \binom{n}{t} \sum_{s=0}^{n-t} (-1)^s \binom{n-t}{s} \frac{(n-t-s)!}{n!} \\
&= \sum_{s=0}^{n-t} (-1)^s \frac{1}{s!t!}
\end{aligned}$$

The Poisson(1) approximation could be achieved heuristically for small t and many $c_i = 1$ if the rank of B keeps increasing with a certain rate as $n \to \infty$:

$$
\begin{aligned}
P(T(\mathbf{\Pi B}) = t) &= \sum_{u\in\{h\subset\{1,\cdots,k\}:|h|=t\}} \sum_{s=0}^{k-t}(-1)^s \sum_{v\in\{J\subset\{1,\cdots,k\}\backslash u:|J|=s\}} (\prod_{i\in u} c_i)(\prod_{j\in v} c_j)\frac{(n-t-s)!}{n!} \\
&\approx \sum_{u\in\{h\subset\{1,\cdots,k\}:|h|=t\}} \sum_{s=0}^{k-t}(-1)^s \sum_{v\in\{J\subset\{1,\cdots,k\}\backslash u:|J|=s\}} \frac{(n-t-s)!}{n!} \\
&= \binom{k}{t}\sum_{s=0}^{k-t}\binom{k-t}{s}(-1)^s\frac{(n-t-s)!}{n!} \\
&= \sum_{s=0}^{k-t}(-1)^s\frac{k!}{(k-t)!t!}\frac{(k-t)!}{(k-t-s)!s!}\frac{(n-t-s)!}{n!} \\
&= \sum_{s=0}^{k-t}\frac{(-1)^s}{t!s!}\frac{k!}{n!}\frac{(n-t-s)!}{(k-t-s)!} \\
&\approx \sum_{s=0}^{k-t}\frac{(-1)^s}{t!s!} \\
&\to \frac{e^{-1}}{t!} \text{ as } k \to \infty \\
&= F_U(l)
\end{aligned}
$$

where $U \sim$ Poisson(1).

Stein-Chen Method for Poisson(1) Approximation

For matching without replacement, again we assume the first $k \leq n$ occasion 2 images get paired after matching with replacement (where some occasion 2 images are paired with more than one occasion 1 images; and only those k occasion 2 images are possible to get paired again after a row permutation).

We have c_i the number of occasion 1 images paired to the i-th occasion 2 image. Then $c_1 + ... + c_k = n$.

Let X_i be the order of the occasion 2 image paired to the i-th occasion 1 image after a row permutation of B. $I_i = \mathbf{1}_{X_i=i}$. Then $T(\mathbf{\Pi B}) = \sum_{i=1}^{k} I_i$. $E[I_i] = c_i/n$, $var(I_i) = c_i(n-c_i)/n^2$, $cov(I_i, I_j) = E[I_iI_j] - E[I_i]E[I_j] = P(I_i = 1)P(I_j = 1|I_i = 1) - c_ic_j/n^2 = c_i/n \cdot c_j/(n-1) - c_ic_j/n^2 = c_ic_j/(n^2(n-1))$.

Let $S = \sum_{i=1}^{k} I_i = T(\mathbf{\Pi B})$. By Stein-Chen's method (see Theorem 8.1 in DEY, 2014) we have:

$$\begin{aligned}
d_{TV}(S, Poisson(1)) &\leq var(S) - 1 + 2\sum_{i=1}^{k}(E[I_i])^2 \\
&= \sum_{i\neq j} cov(I_i, I_j) + \sum_{i=1}^{k} var(I_i) - 1 + 2\sum_{i=1}^{k}(E[I_i])^2 \\
&= \sum_{i\neq j} \frac{c_ic_j}{n^2(n-1)} + \sum_{i=1}^{k}\frac{c_i(n-c_i)}{n^2} - 1 + 2\sum_{i=1}^{k}\frac{c_i^2}{n^2} \\
&= \frac{(c_1 + ... + c_k)^2}{n^2(n-1)} - \frac{\sum_{i=1}^{k} c_i^2}{n^2(n-1)} + 1 - \sum_{i=1}^{k}\frac{c_i^2}{n^2} - 1 + 2\sum_{i=1}^{k}\frac{c_i^2}{n^2} \\
&= \frac{1}{n-1} - \frac{\sum_{i=1}^{k} c_i^2}{n^2(n-1)} + \frac{\sum_{i=1}^{k} c_i^2}{n^2} \\
&= \frac{1}{n-1} + \frac{(n-2)}{n^2(n-1)}\sum_{i=1}^{k} c_i^2
\end{aligned}$$

To check the result:

Let $c_1 = ... = c_{13} = 4$ then we get bound $21/221$ as the Montmort's Preize

Problem (see Example 8.3 in DEY, 2014).

Let $k = n$ and $c_1 = ... = c_n = 1$ (or equivalently we assume matching without replacement). Then the above just becomes $2/n$, another known bound for the Montmort's hat matching problem (Chatterjee, Diaconis, and Meckes, 2005).

For our HCP dataset in Chapter 4, with $n = 466$ subjects the calculated bound is 0.0063.

3.3.2 Rank Sums

In this section, we discuss the approximation of permutation tests using rank sum statistics. Following the similar notations as Section 3.3.1, suppose each subject is measured twice and that $\boldsymbol{x}_{it}$ denotes the t-th measurement of subject i, where $i = 1, \ldots, n$ and $t = 1, 2$. For simplicity, we denote $\boldsymbol{X}_t = (\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{nt})$ as the t-th combined measurement from all subjects and $\boldsymbol{X}$ as the collection of all measurements, if there is no ambiguity within the context. Let $\boldsymbol{R} = [r_{ij}]$ be the rank matrix, which records the exact ranks within each row instead of only the matches in the matching matrix, $\boldsymbol{B}$. Again we assume there is no tie. Then the trace of $\boldsymbol{R}$ is the rank sum statistic (2.7).

Denote the diagonals of the permuted rank matrix as $D_1, \ldots, D_n$. Then marginally each D_i follows discrete uniform distribution over $\{1, \ldots, n\}$, but correlation may exist between the diagonals. Specifically, we have that

$$\mathrm{cov}(D_i, D_j) = \frac{(n+1)^2}{4(n-1)} - \frac{\sum_{k=1}^{n} r_{ik} r_{jk}}{n(n-1)}$$

and

$$\mathrm{var}\left(\sum_{i=1}^{n} D_i\right) = \frac{n^2(n+1)(2n+1)}{6(n-1)} - \frac{\sum_{k=1}^{n} s_k^2}{n(n-1)} \overset{def}{=} \sigma^2,$$

where s_k is the sum over the k-th column of the rank matrix, $\boldsymbol{R}$ (see Appendix in Section 3.3.2). Therefore, the following approximation matches the permutation distribution up to the first two moments:

$$\sum_{i=1}^{n} D_i \overset{D}{\approx} N\left(\frac{n(n+1)}{2}, \sigma^2\right). \tag{3.1}$$

Note that the minimum of the permutation distribution variance, σ^2, is 0, which is achieved when rows of $\boldsymbol{R}$ are repeated. Approximation matching higher moments is also possible.

Denote $W = \sum_{i=1}^{n} D_i$. Under regularity conditions, we may have another normal approximation

$$\frac{W - \frac{n(n+1)}{2}}{\sqrt{\frac{n^2(n+1)}{12}}} \xrightarrow{D} N(0,1) \tag{3.2}$$

assuming s_k converges to $n(n+1)/2$ for all k with high probabilities. However, the general sufficient conditions for (3.2) are not pursued in this thesis.

Appendix: Rank Sums Covariances

Denote the diagonals of the permuted rank matrix as $D_1, \ldots, D_n$. Then marginally each D_i follows discrete uniform distribution over $\{1, \ldots, n\}$, but correlation may exist between the diagonals.

Specifically, we have that

$$var(D_i) = \frac{n^2 - 1}{12},$$

$$
\begin{aligned}
\operatorname{cov}(D_i, D_j) &= E(D_iD_j) - E(D_i)E(D_j) \\
&= \frac{1}{n(n-1)}\left(\sum_{k=1}^{n}\sum_{l\neq k} r_{ik}r_{jl}\right) - \frac{(n+1)^2}{4} \\
&= \frac{1}{n(n-1)}\left(\sum_{k=1}^{n}\sum_{l=1}^{n} r_{ik}r_{jl} - \sum_{k=1}^{n} r_{ik}r_{jk}\right) - \frac{(n+1)^2}{4} \\
&= \frac{1}{n(n-1)}\left((\sum_{k=1}^{n} r_{ik})(\sum_{l=1}^{n} r_{jl}) - \sum_{k=1}^{n} r_{ik}r_{jk}\right) - \frac{(n+1)^2}{4} \\
&= \frac{1}{n(n-1)}\left(\frac{n^2(n+1)^2}{4} - \sum_{k=1}^{n} r_{ik}r_{jk}\right) - \frac{(n+1)^2}{4} \\
&= \frac{(n+1)^2}{4(n-1)} - \frac{\sum_{k=1}^{n} r_{ik}r_{jk}}{n(n-1)}
\end{aligned}
$$

and

$$
\begin{aligned}
\operatorname{var}\left(\sum_{i=1}^{n} D_i\right) &= \sum_{i=1}^{n}\operatorname{var}(D_i) + \sum_{i\neq j}\operatorname{cov}(D_i, D_j) \\
&= \frac{n(n^2-1)}{12} + \frac{n(n+1)^2}{4} - \frac{1}{n(n-1)}\sum_{i=1}\sum_{j\neq i}\sum_{k=1}^{n} r_{ik}r_{jk} \\
&= \frac{n(n+1)(2n+1)}{6} - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j\neq i} r_{ik}r_{jk} \\
&= \frac{n(n+1)(2n+1)}{6} - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{k=1}^{n} r_{ik}(s_k - r_{ik}) \\
&= \frac{n(n+1)(2n+1)}{6} - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{k=1}^{n} r_{ik}s_k + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{k=1}^{n} r_{ik}^2 \\
&= \frac{n(n+1)(2n+1)}{6}\cdot(1 + \frac{1}{n-1}) - \frac{1}{n(n-1)}\sum_{k=1}^{n} s_k^2
\end{aligned}
$$

$$= \frac{n^2(n+1)(2n+1)}{6(n-1)} - \frac{\sum_{k=1}^{n} s_k^2}{n(n-1)},$$

where s_k is the sum over the k-th column of the rank matrix, $\boldsymbol{R}$. Note that the minimum variance is 0, which is achieved when rows of $\boldsymbol{R}$ are repeated.

References

Finn, Emily S, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable (2015). “Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity”. In: *Nature Neuroscience*.

Airan, Raag D, Joshua T Vogelstein, Jay J Pillai, Brian Caffo, James J Pekar, and Haris I Sair (2016). “Factors affecting characterization and localization of interindividual differences in functional connectivity using MRI”. In: *Human Brain Mapping* 37.5, pp. 1986–1997.

Zuo, Xi-Nian, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John CS Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. (2014). “An open science resource for establishing reliability and reproducibility in functional connectomics”. In: *Scientific Data* 1.

Van Essen, David C, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, WU-Minn HCP Consortium, et al. (2013). “The WU-Minn human connectome project: an overview”. In: *NeuroImage* 80, pp. 62–79.

Choe, Ann S, Mary Beth Nebel, Anita D Barber, Jessica R Cohen, Yuting Xu, James J Pekar, Brian Caffo, and Martin A Lindquist (2017). “Comparing test-retest reliability of dynamic functional connectivity methods”. In: *NeuroImage* 158, pp. 155–175.

Poldrack, Russell A and Jean-Baptiste Poline (2015). “The publication and reproducibility challenges of shared data”. In: *Trends in Cognitive Sciences* 19.2, pp. 59–61.

Choe, Ann S, Craig K Jones, Suresh E Joel, John Muschelli, Visar Belegu, Brian S Caffo, Martin A Lindquist, Peter CM Van Zijl, and James J Pekar (2015). “Reproducibility and temporal structure in weekly resting-state fMRI over a period of 3.5 years”. In: *PLoS One* 10.10, e0140134.

Landman, Bennett A, Alan J Huang, Aliya Gifford, Deepti S Vikram, Issel Anne L Lim, Jonathan AD Farrell, John A Bogovic, Jun Hua, Min Chen, Samson Jarso, et al. (2011). "Multi-parametric neuroimaging reproducibility: a 3-T resource study". In: *NeuroImage* 54.4, pp. 2854–2866.

Griffanti, Ludovica, Michal Rolinski, Konrad Szewczyk-Krolikowski, Ricarda A Menke, Nicola Filippini, Giovanna Zamboni, Mark Jenkinson, Michele TM Hu, and Clare E Mackay (2016). "Challenges in the reproducibility of clinical studies with resting state fMRI: An example in early Parkinson's disease". In: *NeuroImage* 124, pp. 704–713.

Shou, H, A Eloyan, S Lee, V Zipunnikov, AN Crainiceanu, MB Nebel, B Caffo, MA Lindquist, and CM Crainiceanu (2013). "Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2)". In: *Cognitive, Affective, & Behavioral Neuroscience* 13.4, pp. 714–724.

Aron, Adam R, Mark A Gluck, and Russell A Poldrack (2006). "Long-term test–retest reliability of functional MRI in a classification learning task". In: *NeuroImage* 29.3, pp. 1000–1006.

Good, Phillip (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Johnson, W Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127.

Pentico, David W (2007). "Assignment problems: A golden anniversary survey". In: *European Journal of Operational Research* 176.2, pp. 774–793.

Hoeffding, Wassily (1952). "The large-sample power of tests based on permutations of observations". In: *The Annals of Mathematical Statistics*, pp. 169–192.

Barton, DE (1958). "The matching distributions: Poisson limiting forms and derived methods of approximation". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 73–92.

DEY, PARTHA (2014). "Stein-Chen Method for Poisson Approximation". In:

Chatterjee, Sourav, Persi Diaconis, Elizabeth Meckes, et al. (2005). "Exchangeable pairs and Poisson approximation". In: *Probability Surveys* 2, pp. 64–106.

Chapter 4

Covariate Analysis for Fingerprinting on HCP Data

It should be noted that the population-level fingerprinting index and discriminability are well defined only if there exists homogeneity among the probabilities in (2.4) and (2.9). Such assumptions are met implicitly under the (M)ANOVA models. However, it has not been studied how other factors may impact the reliability statistics when heterogeneity exists.

In this chapter, we will take fingerprinting as an example to demonstrate that various covariates may impact the test statistic. For fingerprinting, naturally we can use matching as an estimate for the potentially heterogeneous probabilities (2.4) for all i's, so that the covariate analysis can follow. We not only showed that demographic factors such as age may be marginally associated with the individual reliability scores (Section 4.1), but also found that strong covariate structures such as twins or siblings can construct relatively large overall reliability score across completely distinct groups of people (Section 4.2). In the latter case, the reliability score can be understood as a similarity score, which shows the potential of applying reliability measures in heritability analysis.

The Poisson(1) approximation for fingerprinting also allows the voxel-wise fingerprinting tests at a low computational cost. The results illustrate different patterns of reliability across the brain (Section 4.3). It also shows that heterogeneity may exist across different elements from a measurement vector for the reliability score of a high-dimensional measurements.

For the Human Connectome Project (HCP) dataset, 466 participants (273 females, age 22 to 36), each with two separated resting state fMRI sessions on consecutive Day 1 and Day 2, were included from the HCP (Van Essen et al., 2013) S500 release. Preprocessing was conducted following the minimal preprocessing pipelines (Glasser et al., 2013). For each participant, the rs-fMRI scan with the left-to-right phase encoding direction in each session was used so that we had 932 scans in total for matching.

The atlas with 268 nodes partitioned into eight networks defined with the Shen's functional parcellation method on the independent health controls (Shen et al., 2013; Finn et al., 2015) was applied to each rs-fMRI image. The measurement vector $\boldsymbol{x}_{it}$ for subject i on the first ($t = 1$) or the last ($t = 2$) visit was taken as the upper triangular of the Pearson correlation (z transformed) matrix calculated for all the nodes using their time series during the corresponding scan. The distance, $\delta(\cdot, \cdot)$, was defined as one minus the Pearson correlation between the two feature vectors.

4.1 Covariates Associated with Matching

In this section, we will focus on using if a subject is correctly matched (1 for yes, 0 for no) as an example of individual reliability scores. Potentially, similar scores can be calculated for discriminability and rank sums too.

On the HCP dataset, matching with replacement on the 466 participants resulted in 350 people (75.11%) getting matched to themselves. Let 1 represent that a subject got correctly matched and 0 otherwise. Using a logistic regression model, we regressed the matches against demographic covariates, including years of education, age, sex, race (having levels "Asian/Native Hawaiian/Other Pacific Islander", "Black or African American", "White", "More than one" and "Unknown or Not Reported"; "Asian/Native Hawaiian/Other Pacific Islander" as the baseline), income and whether the participant is still in school. Two variables were marginally interesting: age with estimated odds ratio 1.06, 90% CI $[1.01, 1.12]$, Wald z statistic 1.80 and p-value 0.073; the race category for black or African American, having an estimated odds ratio 0.15, 90% CI $[0.02, 0.94]$, Wald z statistic -1.70 and p-value 0.088. Though these variables show weak evidence for associations with matching, recall that the ages, ranging from 22 to 36 on the HCP dataset, were all healthy and younger.

We further investigated if any similarity in terms of resting state connectivity existed among people with the same age and race category. Within each iteration, from each of the 208 families we randomly selected one subject so that no sibling structure existed. We then partitioned the 208 subjects by age and race categories. We randomly chose 20 combinations of age and race categories that contained more than one subject in the 208 samples. From subjects with each of the selected age and race combination, we then randomly chose an ordered pair of subjects. For the first subject of a pair we took the measurement of the first experiment session and for the second subject we took that of the second experiment session. We then conducted matching with replacement on the two groups of 20 measurements, now having totally distinct participants on the two session. After 1,000 iterations the empirical distribution was plotted for the total matches with an empirical distribution from the previous iid Poisson(1) samples as comparison. From Figure 4.1 a slight right shift from the Poisson(1) was observed for the age and race matched simulated samples with a proportion of rejecting the null at level 0.05 in the Poisson tests being 6.3%, which was larger than that in the Poisson(1) samples as 1.4% and probability 1.9% of being greater than 3 for Poisson(1) distribution; these were substantially smaller than those in the dizygotic twins (54.1%) or non-twin siblings (54.2%) (see Section 4.2).

4.2 Matching for Comparing Connectome Similarities Between Twins or Non-Twin Siblings

Our HCP dataset included 53 families with monozygotic (MZ) twins and other 24 families with dizygotic (DZ) twins, all verified by genotyping. There were another 68 families with genotyping data available that had at least two siblings but no twins (NotTwin), which added up to 157 non-twin siblings.

Within each iteration, from each of the three types of families above (MZ, DZ or NotTwin), we randomly selected 20 families. Then from each of the

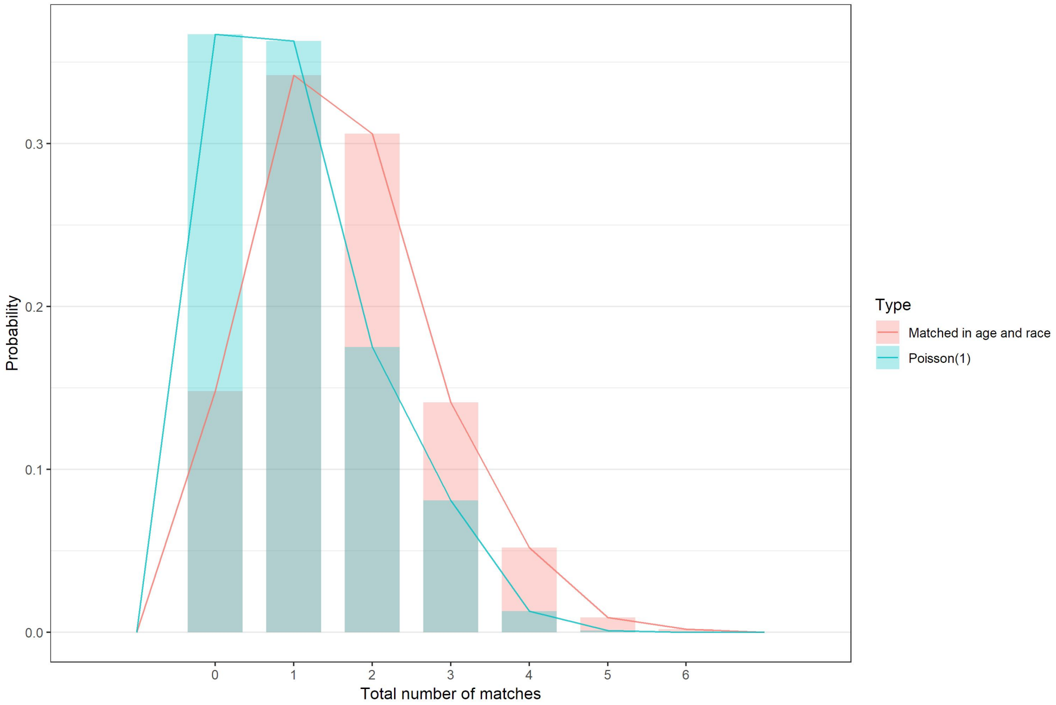

Figure 4.1: The simulated distribution of the total number of matches when matching two groups of distinct people (each of size 20) who were randomly selected from different families and were matched in age and race in the HCP dataset (see Section 4.1). Matching with replacement was conducted. The empirical distribution of a Poisson(1) random variable after 1,000 iterations was also plotted as comparison.

selected families, we randomly chose an ordered pair of twins (for families with MZ and DZ twins) or non-twin siblings (for families with no twins but at least two siblings and with genotyping data available). We also randomly selected 20 ordered pairs of subjects from all the 466 participants (labeled Random).

For each selected ordered pairs, we took the measurement of the first experiment session for the first subject and that of the second experiment session for the second subject. Then for each of the four scenarios (MZ, DZ, NotTwin and Random), we had two groups of 20 measurements from totally distinct subjects.

If different levels of similarities between siblings existed, then the distributions of the total number of matches for siblings could diverge not only from that when siblings were no closer than random people and the exchangeability assumption held, i.e. a Poisson(1) distribution, but between those of different sibling types as well.

After 1,000 iterations the empirical distributions were plotted (Figure 4.2). An empirical distribution of 1,000 iid Poisson(1) samples was also plotted as comparison. We observed similar distributions for DZ twins and non-twin (NotTwin) siblings, with the proportions of rejecting the null at level 5% being 54.1% and 54.2% respectively. But for MZ samples the numbers of matches were greater than 3 in all iterations, meaning the proportion of rejection is 100%. These results could also be seen as supportive evidence in terms of the brain connectivity for the genetic assumption that MZ twins having greater similarity than DZ twins or non-twin siblings, which were all closer than random pairings.

Such matching experiments between distinct subjects demonstrated how the fingerprint test when specially designed can serve as a test for the existence of similarity among people with certain social or genetic relations. According to the experiment results, the power of such a test could be relatively low (around 50% for the level of similarity between DZ twins or non-twin siblings) or very high (close to 100% for the level of similarity between MZ twins) for brain connectivity measurements depending on the (usually unspecified) alternative hypothesis. The empirical distributions of the test statistic demonstrated a way of comparing the levels of brain connectome similarities for different genetic or social relations.

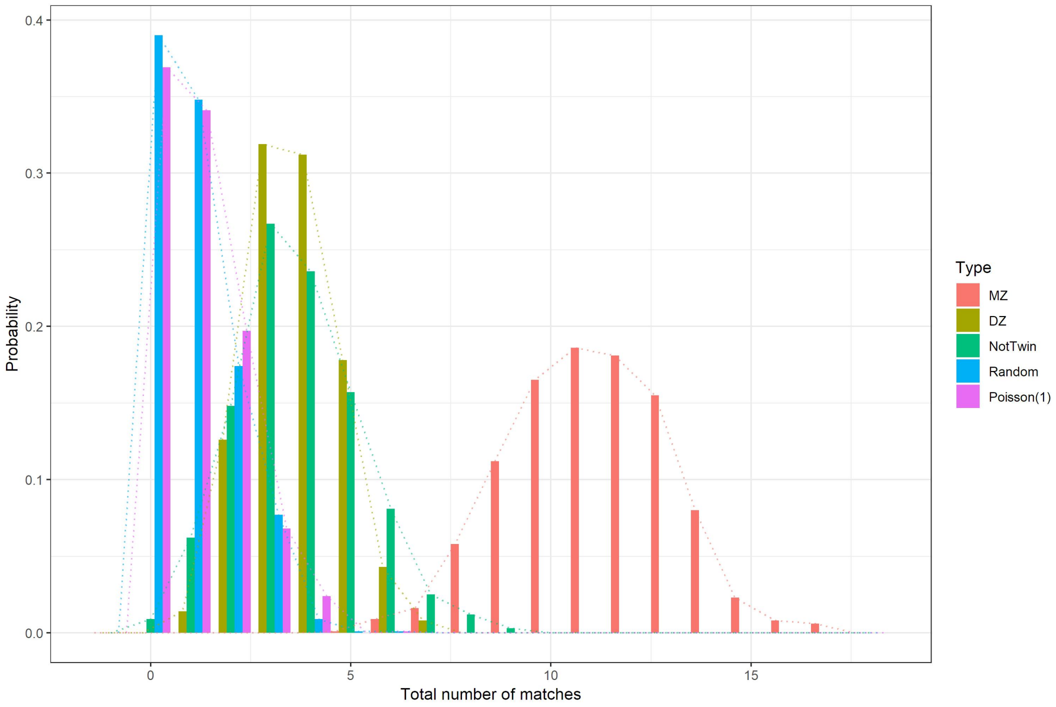

Figure 4.2: The simulated distributions of the total number of matches when matching two groups of distinct people (each of size 20) from the HCP dataset. For each person selected in the first group, there was another monozygotic twin/dizygotic twin/non-twin sibling/random person in the second group for the MZ/DZ/NotTwin/Random scenarios, respectively (see Section 4.2). Matching with replacement was conducted. The empirical distribution of a Poisson(1) random variable after 1,000 iterations is also plotted as comparison.

4.3 Brain Maps of Identifying Pairs of Nodes by Network

Consider evaluating how well a single pair of nodes can identify people by conducting matching with replacement with only the single inter-node z-transformed correlations. Since the measurements are one dimensional we use the absolute difference as distance and randomly choose a match when ties appear. We use the Poisson approximation to the number of matches. An FDR adjustment follows for multiple testing. The Poisson approximation is useful in this setting, as the number of matching experiments grows with the order of the number of nodes squared.

On the HCP dataset using the sample of $466, 106$ identifying pairs of nodes were discovered out of $35,778 = \binom{268}{2}$ pairs (268 nodes). The total matches on those identifying pairs ranged from 7 to 10.

For simplicity, we combined the eight networks into five and then counted the identifying pairs between the following five combined networks: FP (the combination of Medial Frontal and Frontoparietal networks), DMN (Default mode network), SC (Subcortical-cerebellum network), Motor (Motor network) and Visual (the combination of Visual I, Visual II and Visual Association networks). FP was the network with most identifying pairs (20).

We further conducted matching with replacement only using the pairs between any two selected networks. It led to similar results that the identification rate on FP was the highest (90.6%). The 20 identifying pairs within the FP network are visualized (see Figure 4.3) on the ICBM 152 template brain (Mazziotta et al., 2001) with the **rgl** and **misc3d** packages in **R** (Adler,

Table 4.1: Numbers of the identifying pairs between the five combined networks on the HCP dataset (see Section 4.3). The pairs of nodes were selected by the Poisson approximated permutation test on the total matches from matching with replacement using only the z transformed correlations between each single pair.

	FP	DMN	SC	Motor	Visual
FP	20	6	14	11	9
DMN		1	3	3	4
SC			6	6	5
Motor				2	10
Visual					4

Table 4.2: Identification rates (in %) from matching with replacement using only the z transformed correlations of the pairs between the five combined networks on the HCP dataset (see Section 4.3).

	FP	DMN	SC	Motor	Visual
FP	90.6	80.7	70.0	61.2	66.7
DMN		50.0	56.7	33.7	45.5
SC			44.6	45.7	53.6
Motor				42.3	45.7
Visual					58.8

Murdoch, and others, 2018; Feng and Tierney, 2008; Muschelli, Sweeney, and Crainiceanu, 2014).

The matching performance over individual nodes mirrors neuroscientific intuition that frontal networks are more idiosyncratic and personal, while motor and visual networks are more common across individuals.

4.4 Discussion

In this chapter we conducted different covariate analyses for the matching permutation tests with so-called fMRI fingerprinting. We found that, regardless of the matching strategy, the tests results in a Poisson(1) null distribution

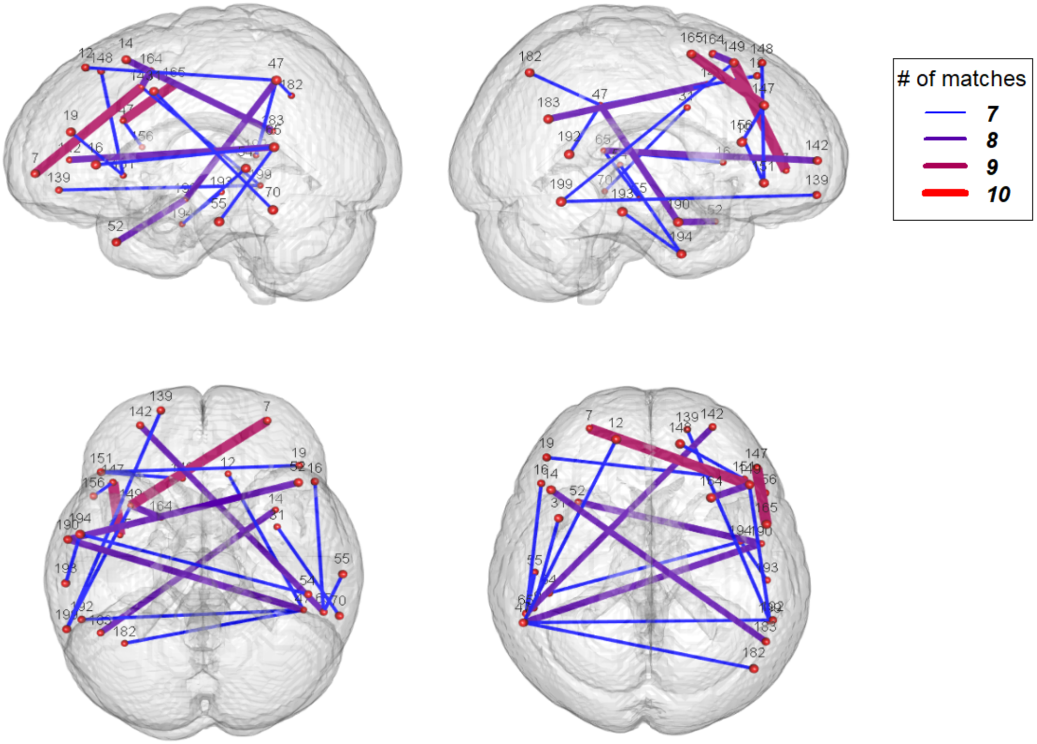

Figure 4.3: The 20 identifying pairs of nodes within the FP network visualized on the ICBM 152 brain template (see Section 4.3). Nodes were labeled by their orders on the atlas and were plotted at the center. Pairs of nodes were colored from blue to red depending on the number of matches when matching with replacement was conducted with only the z transformed correlations between each single pair.

for the number of correct matches (Section 3.3.1). Thus, one can compare the number of matches to the relevant upper quantile of a Poisson(1) without further computing. This is particularly useful for studies of individual brain locations, or pairs of locations. In these settings, the lack of need for calculating a permutation based null distribution dramatically reduces computing time. In addition, the high power of the test mitigates the need for elaborate multiple comparison procedures and simpler more conservative variations would likely suffice.

While nearly any reasonable permutation and matching strategy yields a Poisson(1) null distribution for the number of correct matches, there are differences between the strategies. For example, matching with replacement yields a different answer whether occasion 1 or 2 is used as the reference group. In addition, poor matching without replacement strategies can be dependent on the original subject ordering. Matching with replacement more easily generalizes to multiple measurements per subject.

The exchangeability test was seen to be very highly powered and sensitive to assumptions towards a greater propensity to reject. Most notably, any correlation of the measurement with a demographic or clustering variable will aid in matching. This is intuitive. If one had pairs of outfits from several people and had to match them up in the absence of the owners, the task would be much harder if everyone was the same size, gender, etc. This has implications for the use of fingerprinting as a measure of reproducibility. For example, it is well known that resting state fMRI data changes with age. For the same experimental protocol measures of reproducibility would change depending on the age variation of the study subjects.

When there exist potential covariate or clustering variables, a study of matching performance and its associations is necessary. We suggest the use of logistic regression on whether or not subjects were correctly matched for this task.

Subject identification is also an incomplete measure of the performance of a metric. It is worth remembering that ones actual fingerprint itself is a very good identifier, but is otherwise biologically meaningless, whereas gender, sex, medication usage, etc. are all poor subject identifiers but scientifically useful.

The HCP data included twins and it interesting that matching performance followed the appropriate order (from best performance): self, monozygotic twin, dizygotic twin, non-twin sibling and stranger. Among the basic demographics, age, education and race showed some association with matching performance. Various numeric experiments showed that one can obtain a more significant result by making the distribution of the significant demographics more variable, even when matching to strangers.

The final analysis considered all pairs of regions separately. It was primarily frontal cortical regions that were the most fingerprint-like (i.e. idiosyncratic). This mirrors both intuition and general results in this area. Intuition would suggest, for example, that intra-motor or intra-visual, connections would be similar across a collection of typical subjects simply because of the consistency of motor and visual function.

References

Van Essen, David C, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, WU-Minn HCP Consortium, et al. (2013). "The WU-Minn human connectome project: an overview". In: *NeuroImage* 80, pp. 62–79.

Glasser, Matthew F, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. (2013). "The minimal preprocessing pipelines for the Human Connectome Project". In: *NeuroImage* 80, pp. 105–124.

Shen, Xilin, Fuyuze Tokoglu, Xenios Papademetris, and R Todd Constable (2013). "Groupwise whole-brain parcellation from resting-state fMRI data for network node identification". In: *NeuroImage* 82, pp. 403–415.

Finn, Emily S, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable (2015). "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity". In: *Nature Neuroscience*.

Mazziotta, John, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. (2001). "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 356.1412, pp. 1293–1322.

Adler, Daniel, Duncan Murdoch, and others (2018). *rgl: 3D visualization using OpenGL*. URL: `https://CRAN.R-project.org/package=rgl`.

Feng, Dai, Luke Tierney, et al. (2008). "Computing and displaying isosurfaces in R". In: *Journal of Statistical Software* 28.1, pp. 1–24.

Muschelli, John, Elizabeth Sweeney, and Ciprian Crainiceanu (2014). "brainR: interactive 3 and 4D images of high resolution neuroimage data". In: *The R Journal* 6.1, p. 41.

Chapter 5

Clinical Applications of Functional Connectivity

As the evidence from measurement reliability (Bridgeford et al., 2019) and fingerprinting (Finn et al., 2015; Rosenberg et al., 2016) analyses suggest, FC has potentials of a reliable measurement of brain functional characteristics and a new source of functional biomarkers. In this chapter, we will discuss a real data example where FC is applied in the discovery of the biomarkers associated with a causal quantity of interest. This highlights the prospect of FC applications in various aspects of trial data analysis, such as treatment effect heterogeneity, dynamic treatment regimes, and precision medicine. Future directions, common issues, and precautions are also discussed.

5.1 Causal Biomarkers Discovery in a PPA Trial

Transcranial direct current stimulation (tDCS)—weak electrical current passed over the scalp into the brain—has been shown to benefit language performance in primary progressive aphasia (PPA) (Baker, Rorden, and Fridriksson, 2010;

Chrysikou and Hamilton, 2011; Fiori et al., 2011; Fridriksson et al., 2011; Kang et al., 2011; Marangolo et al., 2011; Monti et al., 2008). In this study, tDCS showed a significant effect on a behavioral outcome called semantic fluency. We are now further interested in if any language related FC baselines are predictive for the potentially heterogeneous tDCS effects. These predictors are biomarkers associated with the individual causal effects and are of importance for the precision health purposes.

5.1.1 Data

5.1.1.1 Participants and Overall Design

Thirty-six individuals with primary progressive aphasia participated in this study (17 female): 14 with logopenic variant PPA (lvPPA), 13 with non-fluent variant PPA (nfvPPA), and 9 with semantic variant PPA (svPPA). All were right-handed, native English speakers, between 50 and 80 years old, and diagnosed based on clinical assessment, neuropsychological and language testing, and MRI. Informed consent was obtained from participants or their spouses (for those with comprehension deficits), and all data were acquired in compliance with the Johns Hopkins Hospital Institutional Review Board. Figure 5.1 shows the participants recruited and their randomization to tDCS or sham. Each PPA variant group was matched by sex, age, education, years post onset of symptoms, and overall Frontotemporal Dementia Clinical Dementia Rating score (FTD-CDR) and language severity measures (Tables 5.1, 5.2).

A within-subjects crossover design with two experimental conditions was

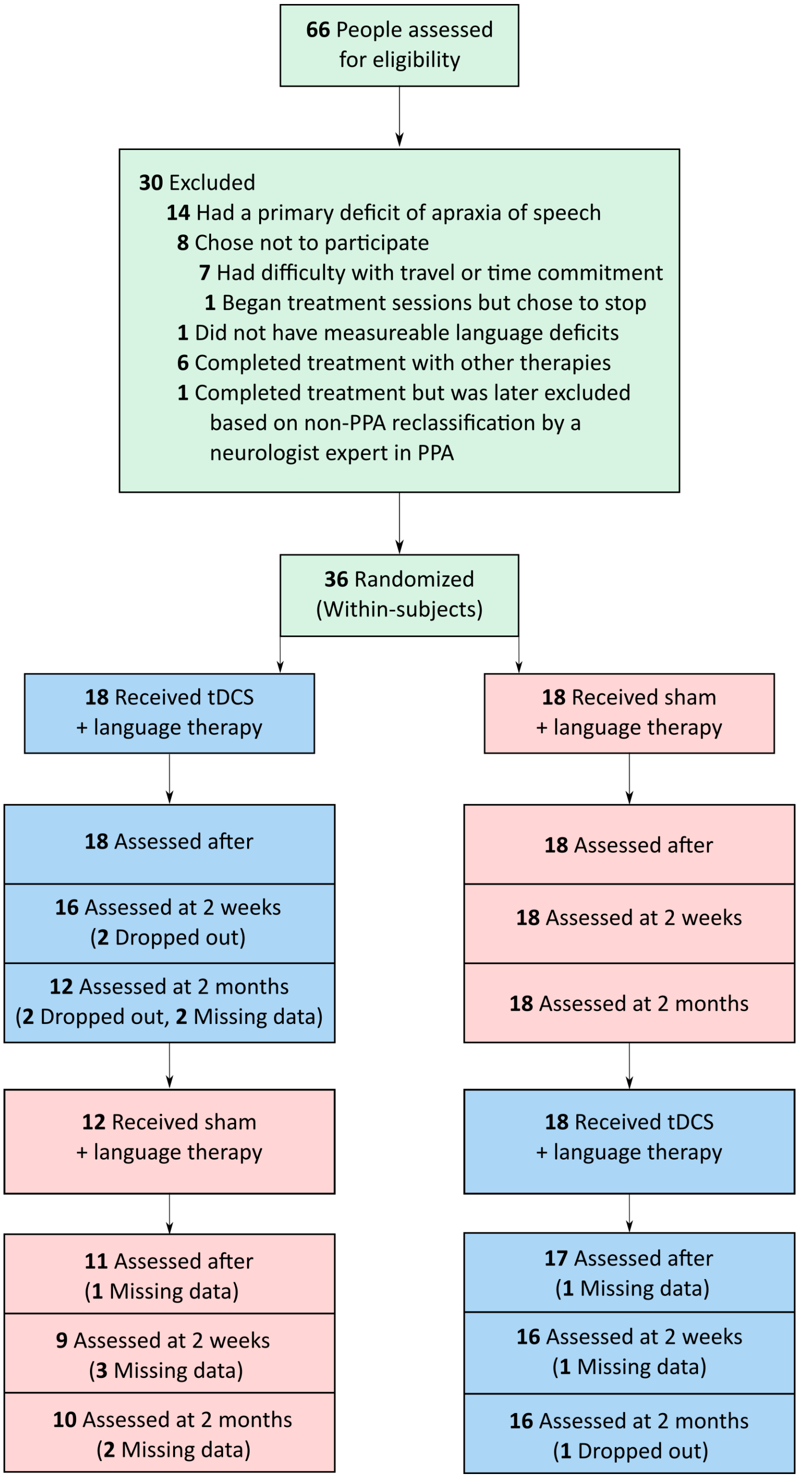
66 People assessed for eligibility
30 Excluded
14 Had a primary deficit of apraxia of speech
8 Chose not to participate
7 Had difficulty with travel or time commitment
1 Began treatment sessions but chose to stop
1 Did not have measureable language deficits
6 Completed treatment with other therapies
1 Completed treatment but was later excluded based on non-PPA reclassification by a neurologist expert in PPA
36 Randomized (Within-subjects)
18 Received tDCS + language therapy
18 Assessed after
16 Assessed at 2 weeks (**2** Dropped out)
12 Assessed at 2 months (**2** Dropped out, **2** Missing data)
12 Received sham + language therapy
11 Assessed after (**1** Missing data)
9 Assessed at 2 weeks (**3** Missing data)
10 Assessed at 2 months (**2** Missing data)
18 Received sham + language therapy
18 Assessed after
18 Assessed at 2 weeks
18 Assessed at 2 months
18 Received tDCS + language therapy
17 Assessed after (**1** Missing data)
16 Assessed at 2 weeks (**1** Missing data)
16 Assessed at 2 months (**1** Dropped out)

Figure 5.1: Participants recruited and randomization to tDCS or sham.

Table 5.1: Means (standard deviations) of demographics grouped by first-phase condition (n=36). *Fisher's exact test used. **F(1, 26) reported. FTD-CDR, Frontotemporal Dementia Clinical Dementia Rating Scale sum of boxes (Knopman et al., 2008). F, female; M, male. L, logopenic; N, nonfluent; S semantic.

	tDCS first	sham first	F(1, 34)	p-value
Sex	9 F, 9 M	8 F, 10 M	*	1.000
Variant	7 L, 6 N, 5 S	7 L, 7 N, 4 S	*	0.500
Age (years)	66.17 (7.49)	69.72 (5.42)	2.66	0.113
Years post symptom onset	5.17 (3.40)	4.72 (2.55)	0.20	0.660
Language severity (FTD-CDR)	1.92 (0.90)	1.83 (0.71)	0.10	0.759
Total severity (FTD-CDR)	6.89 (4.53)	7.53 (4.66)	0.17	0.679
Sessions in phase 1	12.72 (2.11)	11.06 (1.63)	7.05	0.012
Sessions in phase 2	10.64 (3.05)	10.94 (1.35)	0.65**	0.427

Table 5.2: Means (standard deviations) of demographics grouped by PPA variant (n=36). *Fisher's exact test used. **F(2, 25) reported. FTD-CDR, Frontotemporal Dementia Clinical Rating Scale sum of boxes (Knopman et al., 2008). F, female; M, male. s, sham; t, tDCS.

	lvPPA	nfvPPA	svPPA	F(2,33)	p-value
Sex	7 F, 7 M	5 F, 8 M	5 F, 4 M	*	0.800
First-period condition	7 s, 7 t	7 s, 6 t	4 s, 5 t	*	1.000
Age (years)	66.29 (8.11)	69.77 (6.00)	67.89 (4.96)	0.91	0.412
Years post symptom onset	4.82 (3.33)	4.65 (2.66)	5.56 (3.08)	0.25	0.780
Language severity (FTD-CDR)	1.57 (0.83)	2.04 (0.72)	2.11 (0.78)	1.76	0.188
Total FTD-CDR	6.18 (3.76)	7.85 (4.19)	7.89 (6.17)	0.57	0.571
Sessions in phase 1	11.93 (2.02)	11.85 (1.91)	11.89 (2.47)	0.01	0.990
Sessions in phase 2	9.57 (4.35)	9.08 (4.17)	10.44 (3.61)	0.21**	0.812

used: speech-language therapy plus anodal tDCS over the left IFG, and speech-language therapy plus sham tDCS. Each condition lasted approximately 12 daily sessions, consecutive except for weekend breaks; the two phases were separated by a 2-month wash-out period. Evaluations—consisting of a set of treated and untreated items of the same task, as well as extensive neuropsychological and neurolinguistic assessments—occurred immediately before, immediately after, two weeks after, and two months after each treatment phase. Both participants and examiners were blind to the experimental condition.

5.1.1.2 Methods of tDCS

Each daily therapy session lasted one hour. For both tDCS and sham conditions, two 2-inch x 2-inch, non-metallic, conductive, rubber electrodes covered with saline-soaked sponges were placed over the right cheek (cathodal electrode) and the left inferior frontal gyrus centered at F7 of the Homan EEG 10-20 electrode position (anodal electrode). The electrodes were hooked up to a Soterix 1x1 Clinical Trials device, which elicited a tingling sensation on the scalp as it ramped up within 30 seconds, to deliver a weak current at an intensity of 2 mA per minute (estimated current density 0.04 mA/cm2; estimated total charge 0.048 C/cm2). In the tDCS condition, current was delivered for 20 minutes for a daily maximum of 40 mA; in the sham condition, current ramped up to 2 mA and immediately ramped down to elicit the same tingling sensation and thus blind the participant to his treatment condition. Stimulation started at the beginning of each therapy session and lasted for 20 min whereas language therapy continued for a full session, i.e., 20-25 additional minutes. Twice during each session, participants rated their level of pain with the Wong-Baker FACES Pain Rating Scale (www.WongBakerFACES.org).

5.1.1.3 Language Intervention

The written language intervention protocol was based on studies that have successfully treated written language production. We adapted the basic design of a spell-study-spell procedure (Rapp and Glucroft, 2009) to an oral and written naming paradigm (Beeson and Egnor, 2006).

During each treatment session, each participant was shown a picture on the computer, and asked to orally name the object or action and then write it down. Object stimuli sets were chosen from the Philadelphia Naming Test (PNT) and actions from the International Picture Naming Project (Szekely et al., 2004). If the participant could not name the picture, he was asked to provide three properties of the item (what it is, what it does, etc.) to check and reinforce semantic knowledge, as in semantic feature analysis treatment (Beeson and Egnor, 2006). If he made an error, he was given corrective feedback and repeated opportunities to correctly say the object or action name. If he still could not name the word, he was provided with the correct word. If the patient wrote the word incorrectly, the clinician would provide a model of the correct spelling in a spell-study-spell procedure, rehearsing the letters one-by-one in a letter-by-letter manner and reinforcing learning by copying. Repetition has been shown to have synergetic effects for both oral and written naming (Beeson and Egnor, 2006).

Trained and untrained sets (10-30 words depending on each participant's severity level) were matched in length and frequency. Evaluations were administered before, immediately after, 2 weeks after, and 2 months after each treatment period. Percentage of correct letters was determined based on a scoring system evaluating the accuracy of each letter, accounting for deletions, additions, substitutions, and movements of letters. A second judge scored each letter, and any discrepancies were resolved later with discussion to generate a consensus score. Then the sum was divided by the total letters possible. To evaluate whether therapy gains generalized to the naming and spelling of other words, untrained words were presented at all evaluation points.

5.1.1.4 Language and Cognitive Assessment

Participants were also evaluated with a series of standardized language and cognitive assessments. Sham and tDCS groups were matched in language and cognitive scores in each task at baseline. For the semantic fluency task, participants were instructed to name as many fruits, animals, and vegetables as possible, administered separately in the order listed here, in one minute per category (Benton et al., 1994). Scores used in the present analysis were calculated by adding the number of words generated in all three categories. Performance was assessed before, immediately after, two weeks after, and two months after each phase.

5.1.1.5 Imaging Acquisition and Preprocessing

Of the 36 participants, 29 had magnetic resonance imaging (MRI) scans—five were severely claustrophobic and two had pacemakers. MRI scans took place at the Kennedy Krieger Institute at Johns Hopkins University. Magnetization-prepared rapid acquisition gradient echo (MPRAGE) and resting-state functional MRI (rs-fMRI) scans were acquired before treatment on a 3-Tesla Philips Achieva MRI scanner with a 32-channel head coil. T1-weighted MPRAGE sequence acquisition involved the following parameters: a scan time of 6 minutes (150 slices); isotropic 1-mm voxel size; flip angle of 8°; SENSE acceleration factor of 2; TR/TE = 8/3.7 milliseconds (ms). rs-fMRI acquisition involved the following parameters: scan time of 9 minutes (210 time-point acquisitions); slice thickness of 3 mm; in-plane resolution of 3.3x3.3 mm2; flip angle of 75;

SENSE acceleration factor of 2; SPIR for fat suppression; TR/TE = 2500/30 ms.

MPRAGE images were segmented into 238 regions of interest (ROIs) using MRICloud, a multi-atlas based, automated image parcellation approach, using a multi-atlas fusion label algorithm (MALF) and large deformation diffeomorphic metric mapping (LDDMM) (Mori et al., 2016; Tang et al., 2013). Preprocessing involved standard routines from the SPM connectivity toolbox for coregistration, motion, and slice timing correction; physiological nuisance correction using CompCor (Behzadi et al., 2007); and motion and intensity TR outlier rejection using "ART" (https://www.nitrc.org/projects/artifact_detect/). To correct for motion, ART detected outliers and a motion matrix was generated; these were used in combination with the physiological nuisance matrix in the deconvolution regression for the remaining TRs.

rs-fMRI scans were preprocessed using MRICloud and coregistered with MPRAGE scans into the same anatomical space (native space); then 78 of the ROIs were parcellated on the rs-fMRI scans. Average time courses for the voxels in each ROI were normalized, and correlations between ROI pairs were calculated and normalized with the Fisher z-transformation. Of the 78 ROIs, 13 were predefined as language-network ROIs: the left and right pars opercularis, pars orbitalis, and pars triangularis of the inferior frontal gyrus (IFG_opercularis_L, IFG_opercularis_R, IFG_orbitalis_L, IFG_orbitalis_R, IFG_triangularis_L, IFG_triangularis_R), left middle temporal gyrus (MTG_L), left supramarginal gyrus (SMG_L), left superior temporal gyrus (STG_L), left inferior temporal gyrus (ITG_L), left fusiform gyrus (FuG_L), pole of the left middle temporal gyrus (MTG_L_pole) and pole of the left superior temporal gyrus (STG_L_pole). Analyses involved the 78 pairs between these 13 ROIs exclusively.

5.1.2 Unbiased Transformations and Causal Biomarkers

For this analysis, we focus on the first-phase data and the period from before to after the intervention only, in order to avoid any possible impact of carryover and to maximize the number of available samples. The primary behavioral outcome, Y, is the change of semantic fluency scores from the baseline. The assignment of tDCS is denoted as T, which is valued as 1 for tDCS and 0 for sham. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_l)^t$ be a vector of baseline factors, such as the demographics or the FC baselines.

Suppose that $\{Y(t) : t \in \{0,1\}\}$ is the collection of counterfactual outcomes for the two treatment arms, and the observed data is

$$\boldsymbol{O} = (\boldsymbol{X}^t, T, Y(T))^t.$$

It is assumed that $Y = Y(t)$ so long as $T = t$, for $t = 0, 1$. The randomization guarantees that $Y \perp T|\boldsymbol{X}$. We also have that for each possible value, $\boldsymbol{x}$, we have $\mathbb{P}(T = 1|\boldsymbol{X} = \boldsymbol{x}) > 0$. Therefore, the average treatment effect (ATE), $\mathbb{E}[Y(1) - Y(0)]$ is identifiable.

It would be of interest to investigate the following conditional average treatment effect (CATE)

$$\mathbb{E}[Y(1) - Y(0)|\boldsymbol{X}],$$

which captures the potential heterogeneity in the tDCS effect on Y. However, only half of the subjects are observed for $Y(1)$ or $Y(0)$. Therefore, a transformation $Y^*(\cdot) : \boldsymbol{O}_i \mapsto Y^*(\boldsymbol{O}_i)$ that satisfies

$$\mathbb{E}[Y^*(O)|\boldsymbol{X}] = \mathbb{E}[Y_1 - Y_0|\boldsymbol{X}] \overset{def}{=} m(\boldsymbol{X}) \tag{5.1}$$

is desired given an IID sample, $O_1, \ldots, O_n$. Here, the unbiased transformation, $Y^*(\boldsymbol{O}_i)$'s, maintains the correct conditional expectation structure, but is fully observed for each subject.

Jackknife pseudovalues (Equation (11.14) in Chapter 11, Efron and Tibshirani, 1994) result the following expression for ATE,

$$Y_i^{Jackknife} = \frac{Y_i T_i}{e} - \frac{Y_i(1 - T_i)}{1 - e}, \tag{5.2}$$

where e is plugged in as the treatment assignment probability, 0.5. Then $Y_i^{Jackknife}$ satisfies the unbiasedness condition (5.1). The unbiased transformations for data censoring in general have been discussed in Rubin and Laan, 2006 but are not in the scope of this thesis.

In the following section, we will focus on the discovery of linear predictors for the Jackknife pseudovalues, $Y_i^{Jackknife}$. Since the conditional expectations of $Y_i^{Jackknife}$ remain the same as the CATE, i.e. $\mathbb{E}\left[Y_i^{Jackknife}|\boldsymbol{X}_i\right] = \mathbb{E}[Y(1)_i - Y(0)_i|\boldsymbol{X}_i]$ given any set of baseline factors, $\boldsymbol{X}_i$, the selected factors are also the predictors for the unobservable individual tDCS effects, $Y(1)_i - Y(0)_i$. These predictors are the biomarkers of interest that are associated with the potentially heterogeneous causal effects.

5.1.3 Predictor Selection

Jackknife pseudovalues were calculated as (5.2) for each individual. Selection of the linear predictors was conducted based on the leave-one-out cross validated (LOOCV) predictive R-squared. At each step of the forward selection, a threshold of 0.1 on the R-squared increase was applied to stop the selection procedure, otherwise the variable with the largest R-squared increase was selected. The predictive R-squared and the root mean squared error (RMSE) of the current step, as well as the increase in predictive R-squared compared to the last step, are reported for each round of the variable selection.

Predictor selection was conducted for two sets of variables: the non-imaging factors (baseline semantic fluency, PPA variant, number of treatment sessions, sex, age, years post onset of symptoms, and total FTD-CDR severity and language severity measures) and the imaging factors (Fisher-z transformed correlations between the prespecified 13 language ROIs of the baseline rs-fMRI). For the second task, to handle missingness in the baseline resting-state functional connectivity data, an inverse propensity score weighting (IPW) method was applied. Propensity scores were estimated using logistic regression with the imaging missingness and all non-imaging factors. The inverse propensity scores were then used as weights; each least square fitting using all subjects in the aforementioned variable selection procedure was replaced with weighted linear regression on the complete cases. The selection criteria based on LOOCV predictive R-squared increase for predicting pseudovalues remained the same.

5.1.4 Results

Baseline FC of two ROI pairs were confirmed to be predictive with the 0.1 threshold on R-squared increase: the left STG : left MTG pole and left IFG opercularis : left IFG triangularis (Table 5.3; Figure 5.3). These two factors constructed a linear prediction model with LOOCV predictive R-squared being 0.416 and RMSE being 6.78. Coefficients of this model were shown in Table 5.4; higher initial connectivity on these pairs is associated with higher CATE. In addition, we monitored the changes of R-squared increases in each round of variable selection (Figure 5.2). Note that in the first round three other imaging predictors provided R-squared increases greater than 0.1, but they were not selected because the Left STG : Left MTG pole had been selected for providing a larger R-squared increase.

Table 5.3: Imaging factors for individual tDCS effect prediction. Predictiveness was evaluated by the LOOCV (predictive) R^2.

	Accumulated R^2	R^2 increase	RMSE
Null Model	0	0	8.496
Left STG : Left MTG pole	0.307	0.307	7.389
Left IFG opercularis : Left IFG triangularis	0.416	0.109	6.782

Table 5.4: The linear prediction model for individual tDCS effects with the two selected imaging factors.

	Estimate	SE	t(21)	p
Intercept	-10.97	3.79	-2.89	0.009
Left STG : Left MTG pole	30.84	7.67	4.02	0.001
Left IFG opercularis : Left IFG triangularis	14.09	5.02	2.80	0.011

No non-imaging factor was confirmed to be predictive of the individual

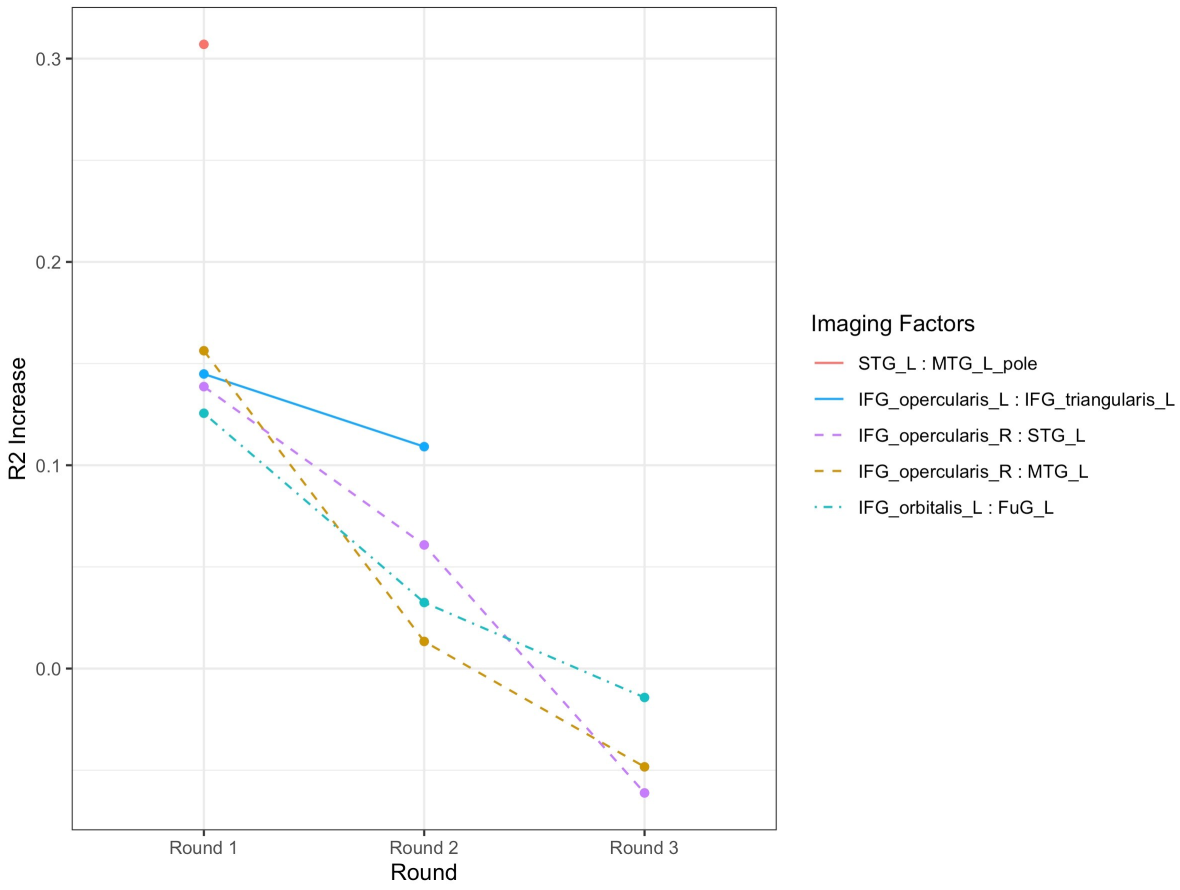

Figure 5.2: Predictive imaging factors and the increases of predictive R-squared for individual tDCS effects. Solid lines represent the selected factors in each round, whereas dotted lines represent the factors that were not selected but also provided over 0.1 increase of predictive R-squared in the first round.

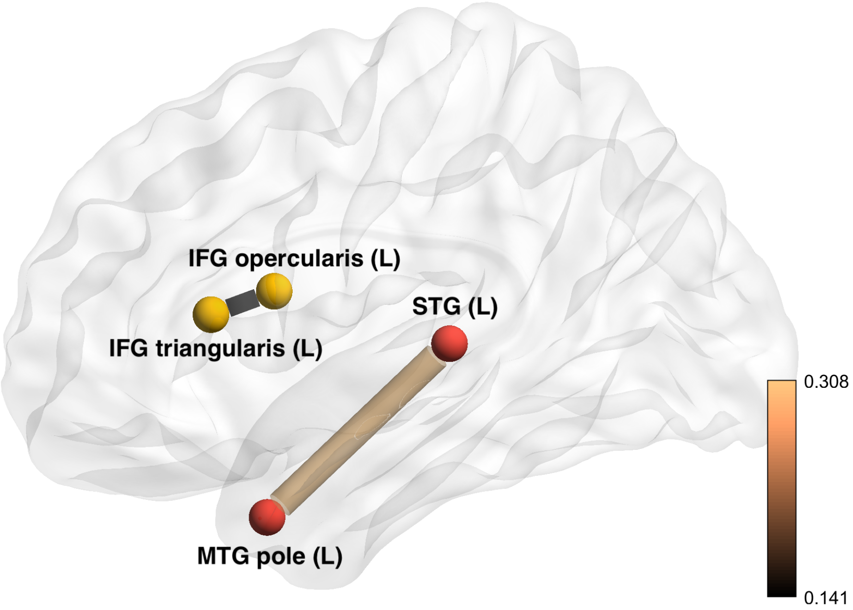

Figure 5.3: Visualization of the selected predictive imaging factors for individual tDCS effects. The positions of the nodes are the average centers of each ROI from the cohort. ROI pairs are plotted and connected if predictiveness of the baseline connectivity is confirmed. Thickness of the edge and the scale of the edge color represent on average how much extra potential semantic fluency score increase one would expect for tDCS compared to sham, with 0.01 higher baseline Fisher z-transformed connectivity.

tDCS effect with the predictive R-squared increment thresholded at 0.1. Overall FTD-CDR and belonging to the nfvPPA group each provided around 0.05 increases in predictive R-squared and resulted in an accumulated predictive R-squared of 0.10, which hints at potential predictiveness for the individual tDCS effect (Table 5.5).

Table 5.5: Non-imaging factors for individual tDCS effect prediction.

	Accumulated R2	R2 increase	RMSE
Null Model	0	0	8.360
Overall FTD-CDR	0.044	0.044	8.407
Having nfvPPA	0.098	0.054	8.166

5.2 Discussion

We conducted screening of FC biomarkers for the potentially heterogeneous causal effects. Essentially, we selected factors based on their predictiveness for the unbiased transformations described in Section 5.1.2, where each transformed outcome maintains the same conditional expectation as the CATE of each individual. Note that the predictiveness of the imaging predictors was evaluated without adjusting for the demographic factors. We did so in order to have a chance to compare the predictiveness of the final models generated by the imaging and non-imaging factors. Such comparison (Table 5.3 and Table 5.5) shows that the accumulated predictive R-squared (0.416 vs 0.098) is higher and the RMSE (6.782 vs 8.166) is lower in the imaging factor model. We argue that such evidence implies that the FC biomarkers provide additional information of individual characteristics compared to the non-imaging factors and have potential importance for precision health purposes.

References

Bridgeford, Eric W, Shangsi Wang, Zhi Yang, Zeyi Wang, Ting Xu, Cameron Craddock, Gregory Kiar, William Gray-Roncal, Carey E Priebe, Brian Caffo, et al. (2019). “Optimal experimental design for big data: applications in brain imaging”. In: *bioRxiv*, p. 802629.

Finn, Emily S, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable (2015). “Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity”. In: *Nature Neuroscience* 18.11, p. 1664.

Rosenberg, Monica D, Emily S Finn, Dustin Scheinost, Xenophon Papademetris, Xilin Shen, R Todd Constable, and Marvin M Chun (2016). “A neuromarker of sustained attention from whole-brain functional connectivity”. In: *Nature Neuroscience* 19.1, p. 165.

Baker, Julie M, Chris Rorden, and Julius Fridriksson (2010). “Using transcranial direct-current stimulation to treat stroke patients with aphasia”. In: *Stroke* 41.6, pp. 1229–1236.

Chrysikou, Evangelia G and Roy H Hamilton (2011). “Noninvasive brain stimulation in the treatment of aphasia: exploring interhemispheric relationships and their implications for neurorehabilitation”. In: *Restorative Neurology and Neuroscience* 29.6, pp. 375–394.

Fiori, Valentina, Michela Coccia, Chiara V Marinelli, Veronica Vecchi, Silvia Bonifazi, M Gabriella Ceravolo, Leandro Provinciali, Francesco Tomaiuolo, and Paola Marangolo (2011). “Transcranial direct current stimulation improves word retrieval in healthy and nonfluent aphasic subjects”. In: *Journal of Cognitive Neuroscience* 23.9, pp. 2309–2323.

Fridriksson, Julius, Jessica D Richardson, Julie M Baker, and Chris Rorden (2011). “Transcranial direct current stimulation improves naming reaction time in fluent aphasia: a double-blind, sham-controlled study”. In: *Stroke* 42.3, pp. 819–821.

Kang, Eun Kyoung, Yu Kyeong Kim, Hae Min Sohn, Leonardo G Cohen, and Nam-Jong Paik (2011). "Improved picture naming in aphasia patients treated with cathodal tDCS to inhibit the right Broca's homologue area". In: *Restorative Neurology and Neuroscience* 29.3, pp. 141–152.

Marangolo, P, CV Marinelli, S Bonifazi, V Fiori, MG Ceravolo, L Provinciali, and F Tomaiuolo (2011). "Electrical stimulation over the left inferior frontal gyrus (IFG) determines long-term effects in the recovery of speech apraxia in three chronic aphasics". In: *Behavioural Brain Research* 225.2, pp. 498–504.

Monti, Alessia, Filippo Cogiamanian, Sara Marceglia, Roberta Ferrucci, Francesca Mameli, Simona Mrakic-Sposta, Maurizio Vergari, Stefano Zago, and Alberto Priori (2008). "Improved naming after transcranial direct current stimulation in aphasia". In: *Journal of Neurology, Neurosurgery & Psychiatry* 79.4, pp. 451–453.

Knopman, David S, Joel H Kramer, Bradley F Boeve, Richard J Caselli, Neill R Graff-Radford, Mario F Mendez, Bruce L Miller, and Nathaniel Mercaldo (2008). "Development of methodology for conducting clinical trials in frontotemporal lobar degeneration". In: *Brain* 131.11, pp. 2957–2968.

Rapp, Brenda and Brian Glucroft (2009). "The benefits and protective effects of behavioural treatment for dysgraphia in a case of primary progressive aphasia". In: *Aphasiology* 23.2, pp. 236–265.

Beeson, Pélagie M and Heather Egnor (2006). "Combining treatment for written and spoken naming". In: *Journal of the International Neuropsychological Society* 12.6, pp. 816–827.

Szekely, Anna, Thomas Jacobsen, Simona D'Amico, Antonella Devescovi, Elena Andonova, Daniel Herron, Ching Ching Lu, Thomas Pechmann, Csaba Pléh, Nicole Wicha, et al. (2004). "A new on-line resource for psycholinguistic studies". In: *Journal of Memory and Language* 51.2, pp. 247–250.

Benton, Arthur Lester, B Abigail, Abigail B Sivan, Kerry deS Hamsher, Nils R Varney, and Otfried Spreen (1994). *Contributions to neuropsychological assessment: A clinical manual*. Oxford University Press, USA.

Mori, Susumu, Dan Wu, Can Ceritoglu, Yue Li, Anthony Kolasny, Marc A Vaillant, Andreia V Faria, Kenichi Oishi, and Michael I Miller (2016). "MRICloud: delivering high-throughput MRI neuroinformatics as cloud-based software as a service". In: *Computing in Science & Engineering* 18.5, pp. 21–35.

Tang, Xiaoying, Kenichi Oishi, Andreia V Faria, Argye E Hillis, Marilyn S Albert, Susumu Mori, and Michael I Miller (2013). "Bayesian parameter estimation and segmentation in the multi-atlas random orbit model". In: *PLoS ONE* 8.6.

Behzadi, Yashar, Khaled Restom, Joy Liau, and Thomas T Liu (2007). "A component based noise correction method (CompCor) for BOLD and perfusion based fMRI". In: *NeuroImage* 37.1, pp. 90–101.

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Rubin, Daniel and Mark J van der Laan (2006). "Doubly robust censoring unbiased transformations". In: *U.C. Berkeley Division of Biostatistics Working Paper Series* Working Paper 208.

Chapter 6

Conclusions

This thesis focused on the reliability and application of resting state fMRI functional connectivity measures. We particularly focused on the general problem of measurement reliability. In an analysis of relationships between a series of recently proposed reliability measures, we focused on discriminability and investigated how it is related with other measures. The results show that inter-study interpretability changes when different reliability measures are used as measures of reliability.

We first confirmed that population discriminability is a reliability measure by showing that it is deterministically linked with the classical univariate reliability measure, ICC, under an assumed true univariate ANOVA model. The non-decreasing transformation of ICC to discriminability (Equation 2.14; Figure 2.1) allows for better interpretation and comparison between studies conducted with either one. Under MANOVA models, the relationship is not deterministic, but the non-decreasing link in general maintains via an approximation of the discriminability bounded by two non-decreasing functions of I2C2 (Figure 2.2).

We also showed that, under mild conditions, the population fingerprint index converges to a fixed proportion of discriminability, which depends on a positive correlation of $corr(\mathbb{I}_{\{\delta_{i,1,2}<\delta_{i,i',1,2}\}}, \mathbb{I}_{\{\delta_{i,1,2}<\delta_{i,i',1,2}\}})$. This correlation can be estimated, as long as the distance matrix generated from the repeated measured data is available. Therefore the instability the fingerprint index with small samples can be detected by estimating its limit.

We argued that a form of the rank sum statistic, (2.7), can be directly transformed into a discriminability statistic, (2.17). Moreover, there exists another consistent discriminability estimator, (2.15), as a function of the sum of ranks. Interestingly, this alternate estimator reduces the computational complexity for calculating discriminability. Potentially, the approximation for the rank sum statistic may be generalized for discriminability, an interesting avenue for future work.

For the evaluation of all aforementioned measures in the terms of simulated power in the reliability test, the results showed that I2C2 or ICC outperformed all others when parametric assumptions were met. However, discriminability or any other non-parametric measures are preferable when Gaussian assumptions were violated. The rank sum statistic can be superior to discriminability with the existence of strong batch effects and small numbers of repeated measurements. Such benefits disappear when more measurements are included in the calculation of discriminability.

In addition, we showed that permutation-based reliability tests result in a Poisson(1) limiting permutation distribution when fingerprinting is applied. We approximated the rank sum permutation distribution up to the second moments with normal distributions.

Admittedly, such tests can be impacted by clustering or demographic factors to various extents. An association analysis can be follow the matching results to investigate these properties. Covariates with strong effects, such as MZ twin status generated relatively high reliability across completely distinct groups of people. Here, an interesting consequence of this investigation was the study of heritability of multivariate brain connectomes using reliability measures. In the future, the individual discriminability scores can also be applied for such association studies. In fact, a follow-up study focusing on these issues is warranted, since the violation of homogeneity can result in violation of the model assumptions for measures, such as fingerprinting and discriminability.

Lastly, we demonstrated in a real data example that the FC data can be applied in a clinical setting, where precision health decisions are of concern. The FC data showed potential as a more informative personal characteristic. However, precaution should be taken that the level of reliability can still vary across different brain networks and locations, as network analysis illustrated in Section 5.3. A high whole brain reliability level does not guarantee reliable or consistent measurements on all subnetworks.

Major challenges of FC applications come from the data dimensionality, the limit of domain knowledge, and the expense of rsfRMI studies. In fact, dimensionality is of greater concern in FC analysis, since an upper-right vector of l by l correlation matrix has a length of $\binom{l}{2}$. Thus, the rate of comparisons increases at the order of l^2, as opposes to l for task-based analysis. The consequence is that a small number of additional ROIs can dramatically increase the FC dimension. High reliability alone cannot guarantee the validity and reproducibility of a small sample study, and it is especially true when we take into account the loss of power due to multiple testing. Except for larger sample sizes, such challenge calls for effective and data-adaptive methods that better utilize the strong correlations within the FC data.

Zeyi Wang
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
667-225-2356 (mobile)
zwang107@jhmi.edu

Education

2015-now **Johns Hopkins Bloomberg School of Public Health**, Baltimore, MD
Ph.D. in Biostatistics, Anticipated May 2020
Thesis title: *Statistical Analysis of Functional Connectivity in Brain Imaging: Measurement Reliability and Clinical Applications*
Advisor: Brian Caffo

2015 **School of Mathematical Sciences, Nankai University**, Tianjin, China
B.Sc. in Statistics

Professional Experience

2016-now **Research Assistant**
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
Advisor: Brian Caffo

2016-now **Research Assistant**
The Language Neuromodulation Lab, Johns Hopkins School of Medicine
Principal Investigator: Kyrana Ksapkini

2013-2015 **Research Scholar**
Bioinformatics Laboratory, School of Mathematical Sciences, Nankai University

Peer-Reviewed Journal Articals

de Aguiar V, Zhao Y, Faria A, Ficek B, Webster KT, Wendt H, **Wang Z**, Hillis A, Onyike C, Frangakis C, Caffo B, and Tsapkini K. (2020). Brain Volumes as Predictors of tDCS Effects in Primary Progressive Aphasia. *Brain and Language*, 200: 104707.

Caffo B, **Wang Z**, Faria A, Zipunnikov V, and Miller M. (in press). A Survey of Visualizing Compositional Data: an Application to Brain Volumetrics. *Encyclopedia of Biostatistics*.

Harris AD, **Wang Z**, Ficek B, Webster K, Edden RA, and Tsapkini K. (2019). Reductions in GABA Following a tDCS-Language Intervention for Primary Progressive Aphasia. *Neurobiology of Aging*, 79: 75-82.

Caffo B, Zhao Y, Eloyan A, **Wang Z**, Mejia A, and Lindquist M. (2018). A Survey of Statistics in the Neurological Sciences with a Focus on Human Neuroimaging. *Wiley StatsRef: Statistics Reference Online*, 1-47.

Wang Z*, Ficek BN*, Zhao Y, Webster KT, Desmond JE, Hillis AE, Frangakis C, Faria AV, Caffo B, and Tsapkini K. (2018). The Effect of tDCS on Functional Connectivity in Primary Progressive Aphasia. *NeuroImage: Clinical*, 19: 703-715.

* Authors contributed equally to the work.

Manuscripts in Preparation

Wang Z, Geuter S, Qi G, Welsh RC, Wager TD, Caffo B, Lindquist M. *Quantification and Correction of Nonindependence Error in fMRI Studies*. To be submitted for publication in May 2020.

Wang Z, Bridgeford EW, Vogelstein JT, and Caffo B. *Statistical Analysis of Data Repeatability Measures*. To be submitted for publication in April 2020.

Wang Z*, Ficek BN*, Hillis AE, Caffo B, Frangakis C, and Ksapkini K. *Transcranial Direct Current Stimulation over the Left Inferior Frontal Gyrus Improves Semantic Fluency in Primary Progressive Aphasia*. (2020) Submitted for publication.

Bridgeford EW, Wang S, Yang Z, **Wang Z**, Xu T, Craddock C, Kiar G, Gray-Roncal W, Priebe CE, Caffo B, Milham M, Zuo X, and Vogelstein JT. (2019). *Optimal Experimental Design for Big Data: Applications in Brain Imaging*. Submitted for publication.

Wang Z, Sair H, Crainiceanu C, Lindquist M, Landman BA, Resnick S, Vogelstein JT, and Caffo B. (2019). *On Statistical Tests of Functional Connectome Fingerprinting*. Submitted for publication.

* Authors contributed equally to the work.

Editorial Activities

Reviewer, International Conference on Machine Learning (ICML) 2019

Teaching

Teaching Assistant, Statistical Machine Learning, 2019
Teaching Assistant, Statistical Consulting, 2018
Teaching Assistant and Lab Lecturer, Survival Analysis, 2018
Teaching Assistant and Lab Lecturer, Advanced Methods in Biostatistics, 2017-2018
Teaching Assistant and Lab Lecturer, Essentials of Probability and Statistical Inference, 2016-2017

Presentations and Posters

Statistical Analysis of Data Reproducibility Measures. JSM, Denver, CO, July 2019
On Statistical Tests of Functional Connectome Fingerprinting. ENAR, Philadelphia, PA, March 2019