Optimal Decisions for Discovery Science via Maximizing Discriminability: Applications in Neuroimaging

Shangsi Wang, Zhi Yang, Xi-Nian Zuo, Michael Milham, Cameron Craddock,
Greg Kiar, William Gray Roncal, Eric Bridgeford, Consortium for Reliability and Reproducibility,
Carey E. Priebe, Joshua T. Vogelstein

August 23, 2016

Contents

I Introduction

In this era of big data, many scientific, government, and corporate groups are collecting and processing massive data sets (1, 2). To obtain optimal quantitative answers to any inquiry about data requires making two decisions: (i) how should the data be collected?, and (ii) how should the data be processed? When the downstream inference task is specified, a priori, we can collect and process data to optimize the performance of task (3, 4). However, recently, across industry, governmental, and academic settings, certain datasets become benchmark or reference datasets. Such data sets are then used for a wide variety of different inferential problems. Collecting and processing these data sets requires massive institutional investments, and choices related to questions(i) and (ii) above have dramatic effects on all subsequent analyses. Optimally addressing experimental design decisions can yield significant savings in both the financial and human costs, and also improve accuracy of analytical results (5–7). Therefore, a theoretical framework to enable investigators to select from a set of possible design decisions in the absence of an explicit task or for multiple tasks could reap great rewards.

This frame work should provide a measure of discriminability (or reliabiliity) which is intuitive to understand and easy to implement. It should be non-parametric and robust; therefore, it is ready to be applied under a variety of settings. It should not be computationally expensive and can be applied to large data sets. Furthermore, it should be simple and unified; as a consequence, we can easily compare it across data sets. Lastly, theories and real data experiments should provide solid support to use discriminability to guide data collection and processing.

To this end, we have proposed and developed a formal definition of discriminability to guide data collection and processing. Discriminability is a non-parametric statistical property of a joint distribution in a hierarchical model, to differentiate between classes of objects. We prove that discriminability (which may be more aptly called reliability), provides a lower bound on predictive accuracy for any downstream inference task, even if we have never seen the covariates to predict in the processing. We then design an estimator of discriminability computed from test-retest data set, demonstrate that it is unbiased, and derive our estimators asymptotic distribution.

Numerical simulations are conducted to demonstrate the basic property of our discriminability estimator in a variety of settings. Then, we apply our approach to choose amongst a set of choices one must make when designing a neuroimaging study to investigate functional connectomics (8, 9). We start by finding the maximally discriminable threshold for converting correlation connectome matrices into binary graphs. Indeed, consistent with our theoretical and simulated results, maximizing the discriminability also maximizes performances on a suite of different downstream inference tasks. We then ask about a series of pre-processing steps: should one motion correct or not, and should one implement global signal regression or not, etc. We determine the optimal choice for each pre-processing steps, and find the maximally discriminable pipelines amongst 64 pre-processing pipelines.

Thus, in total, our discriminability analysis is a powerful tool for making decisions about how to collect and process data sets designed for discovery science. We expect this method to be useful in a wide variety of applications, and therefore have made all the code open source and available from http://openconnecto.me.

II Related Work

There are some successful attempts to quantify reliability or reproducibility in neuroimaging studies (10–17). We are going to review a subset of them which is related to our approach.

- Intraclass correlation coefficient (ICC) is introduced to measure consistency or reproducibility of scalar quantitative measurements (10). In neuroimaging, people attempt to extract one or a few summary scalar statistics from each image and then evaluate the ICC of the statistics (13, 14). They report moderate-to-high test-retest reliability for different statistics. The problem with this approach is that the summary statistics may not be representative. Also, there is no principled approach to average over multiple ICCs.

- Image intraclass correlation coefficient (I2C2) is proposed by Shou et al. to measure reliability (15).

It generalizes classic intraclass coefficient to high dimensional observations. It computes reliability estimates based on the traces of within subject and across subject covariance matrix. It relies on the critical assumption that noise is additive and observations lies in the space equipped with Euclidean distance. As a consequence, it is not suitable to apply to more general settings.

- Graphical intraclass correlation coefficient (GICC) is a reproducibility measure proposed by Yue et al. (16). It is designed specifically for the case when data of interest are binary graphs. It takes a parametric approach by first assuming a probit link function and estimating latent edge feature vectors. Then, it computes GICC based on variation of latent edge feature vectors. In practice, its assumptions is hard to justify and it is computationally expensive to estimate latent features for graphs of moderate size.

- Distance components (DISCO) is proposed by Rizzo and Székely as a measure of dispersion(12). It computes one distance statistic for multiple empirical distributions based on pairwise distances between samples. It can also be used to test the hypothesis that whether multiple sets of samples are drawn from the same distribution or not. Our approach is similar to DISCO in the sense that we all rely on pairwise distance matrix. However, DISCO is designed for testing which requires a fixed number of subjects and a large amount of measurements from each subject. In our studies, we only have a few measurements from each subject which makes DISCO hard to apply.

- NPAIRS data analysis framework is proposed in (11). It takes a resampling approach by splitting data in half. After performing a series of dimension reduction on the data, a label is predicted using Gaussian mixture model. Then, correlation between all pairs of spatially aligned voxels is calculated. A signal-to-noise ratio measure is computed based on the correlation.

- A statistics called estimation stability (ES) is proposed in (17). It is similar to a variance estimator computed through delete-d jacknife resampling. It is applied to smoothing parameter selection in Lasso and is shown to obtain a great reduction of model without sacrificing prediction performance in a task fMRI study.

III Results

III.A Theory

III.A.1 Discriminability as a framework to guide processing

Discriminability measures the overall consistency and differentiability of observations. For example, if a subject is measured twice under the same conditions, two observations should be close to each other given the measure is consistent. In addition, one should be able to tell these two observations come from the same subject when compared to observations from other subjects given the measure is differentiable. We quantify this idea of consistency and differentiability through discriminability.

To formalize the definition of discriminability, consider the following generative process. For each sample i, there exists some true physical property $\boldsymbol{v}_i$. Unfortunately, we do not get directly to observe $\boldsymbol{v}_i$, rather, we measure it with some device, that transforms the truth from $\boldsymbol{v}_i$ to $\boldsymbol{w}_i$ via $f_{\boldsymbol{\phi}}$. The parameter $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ characterizes all options in the measurement, including, for example, which scanner to use, which resolution, the number of images, sampling rate, etc. The output of $f_{\boldsymbol{\phi}}$ is the "raw" observation data $\boldsymbol{w}_i$, but it is corrupt in various ways, including movement or intensity artifacts introduced by the measurement process. Therefore, rather than operating directly on $\boldsymbol{w}_i$, we intentionally "pre-process" the data, in an effort to remove a number of nuisance variables. This pre-processing procedure further transforms the data from $\boldsymbol{w}_i$ to $\boldsymbol{x}_i$ via $g_{\boldsymbol{\psi}}$. The parameter $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ indexes all pre-processing options. In neuroimaing, these options may include whether to perform motion correction, which motion correction, deconvolution, etc. More specifically, the entire code base, including dependencies, and even the hardware the pre-processing is running on, could count as $\boldsymbol{\psi}$. For brevity, we define $\boldsymbol{x}_i := g_{\boldsymbol{\psi}}\big(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i)\big)$. We should notice that $g_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ by their natures are random functions which means even if we measure the same physical property $\boldsymbol{v}_i$ twice the results could be different.

Let i denote the sample's unique *identity* (hereafter, referred to as the *subject*) and t denote the trial number. Thus, there is a single $\boldsymbol{v}_i$ for subject i, but we have $\boldsymbol{x}_{i,t}$, which is the t^{th} trial, implicitly also a function of ϕ and ψ, which encodes all the details of the measurement and pre-processing. If both g_ψ and f_ϕ together do not introduce too much noise, then we would expect that $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{i,t'}$ are *closer* to one another than either are to any other subject's measurement, $\boldsymbol{x}_{i',t''}$. Define δ to be a metric computing the distance between two measurements, $\delta\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Formally, we expect that $\delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'}) \leq \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''})$, for most combinations of $i, i' \neq i, t, t' \neq t, t''$. For brevity, let $\delta_{i,t,t'} := \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'})$ and $\delta_{i,i',t,t''} := \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''})$. This intuition leads to our definition of discriminability:

$$D(\boldsymbol{\psi}, \boldsymbol{\phi}) = \mathbb{P}(\delta_{i,t,t'} \leq \delta_{i,i',t,t''}) \tag{1}$$

In words, discriminability is the probability that within subject distance is smaller than across subject distance. $D(\psi, \phi)$ depends on three matters, namely measurement options f_ϕ, processing options g_ψ and the distribution of true physical property $\boldsymbol{v}$. To understand the equation 1 better, we can expand it

$$D(\boldsymbol{\psi}, \boldsymbol{\phi}) = \mathbb{E}(\mathbb{P}(\delta(g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i))_t, g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i))_{t'}) \leq \delta(g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i))_t, g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{v}_{i'}))_{t''}) | \boldsymbol{v}_i, \boldsymbol{v}_{i'})) \tag{2}$$

The distribution of $\boldsymbol{v}$ is usually out of the control of researchers. However, we want to find the best data collection and processing options. To achieve this, we consider maximizing the discriminability of processed data, that is

$$\underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}, \boldsymbol{\phi} \in \boldsymbol{\Phi}}{\text{maximize}} \quad D(\boldsymbol{\psi}, \boldsymbol{\phi}) \tag{3}$$

It is often the case that data collection is out of control of researchers, that is ϕ is a fixed element in $\boldsymbol{\Phi}$. Therefore, we are only interested in finding the best processing routine encoded by ψ. This is also the focus of this paper, since we do not have opportunity to make decision on data collection choices. In this case, we drop ϕ in our notation and only maximize the discriminability over set $\boldsymbol{\Psi}$

$$\underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}}{\text{maximize}} \quad D(\boldsymbol{\psi}) \tag{4}$$

This approach is intuitive and easy to understand. We will show that maximizing discriminability leads to good prediction performance. In addition, an unbiased estimator is designed to compute discriminability from test-retest data set. In the simulation and application section, we will demonstrate the utility of discriminability through data experiments.

III.A.2 Optimizing discriminability optimizes bound on performance for any task

Consider the situation that the downstream inference task is classification, that is in addition to $\boldsymbol{v}_i$, there are other properties of sample i of interest; we call all of them $\boldsymbol{y}_i \in \mathcal{Y}$. These may include, for example, the phenotype of the subject, including personality tests, demographic information, and genetic data. In this paper, we focus on binary classification problem that is $\mathcal{Y} = \{0, 1\}$. The goal of experimental design, in this context, is to choose $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ to make good prediction of $\boldsymbol{y}_i$ based on observation $\boldsymbol{x}_i$. In this section, we will see that given two pipelines $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, the one with larger discriminability is more likely to have better prediction performance.

To quantify the performance of our choice, we introduce some assumptions. First, assume that each $(\boldsymbol{v}_i, \boldsymbol{y}_i)$ pair is sampled independently and identically from some distribution, $(\boldsymbol{v}_i, \boldsymbol{y}_i) \overset{iid}{\sim} F_{V,Y}$. The goal is to predict the binary-valued *target* variable $\boldsymbol{y}_i$, using $\boldsymbol{x}_i$ as the *predictor* variables. Given a classifier $C : \mathcal{X} \to \mathcal{Y}$, to quantify the performance of classifier, we define the loss function $L(C)$ to be the probability of making error in prediction that is

$$L(C) = \mathbb{P}(C(\boldsymbol{x}_i) \neq \boldsymbol{y}_i)$$

It is known that the minimal prediction error $L^*(\boldsymbol{x}_i, \boldsymbol{y}_i)$ among all possible prediction function is achieved by Bayes classifier (18).

$$L^*(\boldsymbol{x}_i, \boldsymbol{y}_i) := L(C^B)$$

where C^B is the Bayes classifier which is defined by

$$C^B(\boldsymbol{x}_i) := \underset{y \in \{0,1\}}{\text{argmax}}\, \mathbb{P}(\boldsymbol{y}_i = y | \boldsymbol{x}_i)$$

Since $\boldsymbol{x}_i$ depends on pipeline $\boldsymbol{\psi}$, we denote the loss of pipeline $\boldsymbol{\psi}$ by $\ell(\boldsymbol{\psi})$ which is the Bayes prediction error of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

$$\ell(\boldsymbol{\psi}) := L^*(\boldsymbol{x}_i, \boldsymbol{y}_i) = L^*(g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i)), \boldsymbol{y})$$

The next theorem shows the relationship between Bayes classification error and discriminability. Under assumptions that the noise is additive, we can prove theorem 1 which asserts that Bayes classification error is bounded by a decreasing function of discriminability.

Theorem 1. *There is a decreasing function h which only depends on $\boldsymbol{v}$ and $\boldsymbol{y}$, such that*

$$\ell(\boldsymbol{\psi}) \leq h(D(\boldsymbol{\psi}))$$

As a consequence, we expect the classification error to be small when the discriminability is large. An immediate corollary justifies using discriminability to select the optimal processing pipeline.

Corollary 2. *Given two processing pipelines $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, suppose $\boldsymbol{\psi}_1$ is more discriminable than $\boldsymbol{\psi}_2$, that is $D(\boldsymbol{\psi}_1) > D(\boldsymbol{\psi}_2)$. If $\ell(\boldsymbol{\psi}_2) \geq h(D(\boldsymbol{\psi}_1))$, then*

$$\ell(\boldsymbol{\psi}_1) \leq \ell(\boldsymbol{\psi}_2)$$

Also, we must have

$$\ell(\boldsymbol{\psi}_1) \leq h(D(\boldsymbol{\psi}_2))$$

It tells us for any distribution of $\boldsymbol{y}$, we have a tighter bound on Bayes error using the more discriminable pipeline. When choosing from two processing pipelines $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, we should first compute $D(\boldsymbol{\psi}_1)$ and $D(\boldsymbol{\psi}_2)$. We then select the pipeline which yields larger discriminability to have lower bound on the Bayes classification error. This theorem justifies maximizing discriminability for subsequent classification tasks. Figure 1 summarizes the framework to find the optimal processing pipeline.

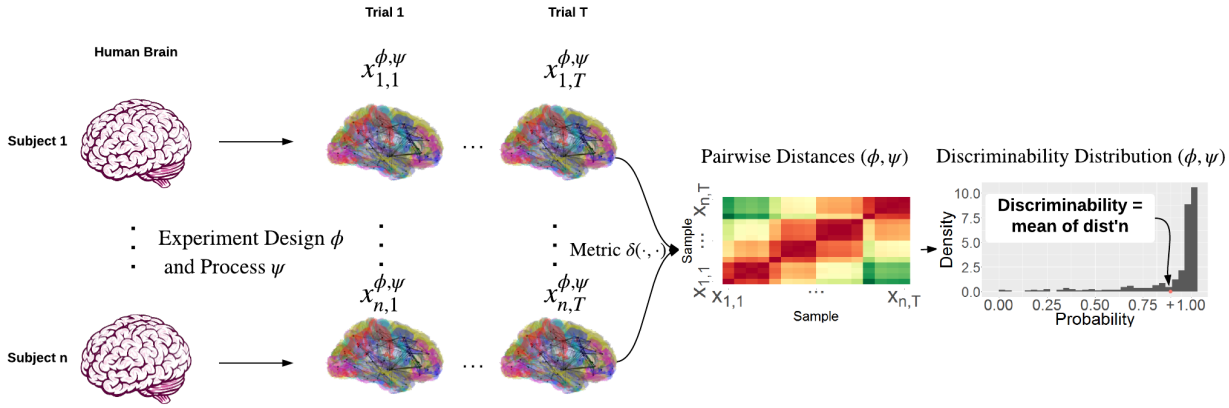

Figure 1: **Decision Making Through Discriminability Framework.** Test-retest data set is collected under experiment design options phi and processed by pipeline ψ. The pairwise distances of all measurements are computed using a metric $\delta(\cdot, \cdot)$. For each pair of measurements of the same subject, we estimate the probability of across subject distances being larger than the within subject distance. Discriminability is the mean of estimated probabilities. Select the option and pipeline with maximum discriminability.

III.A.3 Estimating discriminability

In real applications, distribution of $\boldsymbol{x}_{i,t}$ may never known to us; hence, it is not possible to compute discriminability $D(\boldsymbol{\psi})$ or D in short when there is no ambiguity in processing pipelines under consideration. However, samples $x_{i,t}$ are observed, and we can approximate true discriminability D using an estimator $\hat{D}$ which is a function of observed samples. For each pair of observations $x_{i,t}$ and $x_{i,t'}$ from subject i, we first define

$$\hat{D}_{i,t,t'} = \frac{\sum_{i' \neq i}^{n} \sum_{t''=1}^{s} \mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}}{(n-1)s}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, n is the number of subjects, and s denotes the number of observations per subject. $\hat{D}_{i,t,t'}$ is the faction of observations from other subjects farther away from $x_{i,t}$ than $x_{i,t'}$. It approximates the probability that distances from observations of other subjects to the t^{th} observation of subject i is larger than the distance between t^{th} and t'^{th} trial of subject i. Then, we define the discriminability estimator $\hat{D}$ to be the mean of $\hat{D}_{i,t,t'}$ averaged over all pairs of observations from same subjects.

$$\hat{D} := \frac{\sum_{i=1}^{n} \sum_{t=1}^{s} \sum_{t' \neq t}^{s} \hat{D}_{i,t,t'}}{ns(s-1)}$$

$\hat{D}$ is the sample discriminability which approximates discriminability or population discriminability. The next two lemmas asserts that the discriminability estimator $\hat{D}$ is unbiased and converges to D as the number of subjects n goes to infinity (19).

Lemma 1. *$\hat{D}$ is an unbiased estimator of D, that is*

$$\mathbb{E}(\hat{D}) = D$$

Lemma 2. *As $n \to \infty$, $\hat{D}$ converges to D in probability, that is*

$$\hat{D} \xrightarrow{p} D$$

III.A.4 Testing discriminability

In applications, we are sometimes interested in whether there is any subject specific information in the data. In other words, we want to know whether within subject distance is statistical significantly smaller than across subject distance. This is equivalent to test the null hypothesis that discriminability is differ from 0.5, that is

$$H_0 : D = 0.5$$
$$H_A : D > 0.5$$

We have two valid approaches to address this problem. The first test relies on the bound on variance of $\hat{D}$ which we derived in proving Lemma 2. Specifically, we show that the variance of $\hat{D}$ is less than $\frac{1}{n}$. Based on Chebyshev's inequality, we can derive a 95 percent confidence interval $(\hat{D} - \frac{2\sqrt{5}}{\sqrt{n}}, \hat{D} + \frac{2\sqrt{5}}{\sqrt{n}})$. If 0.5 lies in the confidence interval, we do not reject the null hypothesis; otherwise, we reject the null hypothesis. The second test based on estimating a null distribution for $\hat{D}$ through permutation. In particular, we random permute subject labels for each trial and then estimate discriminability based on permuted labels. Repeat this procedure a large number of times and find the 95^{th} quantile of permuted discriminability. If $\hat{D}$ is less than the 95^{th} quantile, we do not reject the null hypothesis; otherwise, we reject the null hypothesis. The first test is computationally simple and can be generalized to comparing discriminability of two data sets; however, it generally has smaller power than the second test.

III.B Simulations

III.B.1 Convergence of discriminability estimator

In Lemma 1 and 2, we claim discriminability $\hat{D}$ is unbiased and converges to the true population discriminability in probability. We demonstrate these two lemmas through simulation. We consider a simple case that $g_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ together introduce independent additive Gaussian noise ϵ, that is

$$\boldsymbol{x}_{i,t} = g_{\boldsymbol{\psi}}\big(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i)\big) = \boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t} \tag{5}$$

$\boldsymbol{v}_i$ and $\boldsymbol{\epsilon}_{i,t}$ are both independent and identically distributed standard Gaussian random variable that is

$$\boldsymbol{v}_i \overset{i.i.d.}{\sim} \mathbb{G}(0,1) \text{ and } \boldsymbol{\epsilon}_{i,t} \overset{i.i.d.}{\sim} \mathbb{G}(0,1)$$

In addition, $\boldsymbol{v}_i$ and $\boldsymbol{\epsilon}_{i,t}$ are assumed to be independent.
For each subject, we sample one true physical property v_i and two noises $\epsilon_{i,t}$ with $t \in \{1,2\}$. Then, two measurements are generated by $x_{i,t} = v_i + \epsilon_{i,t}$. We let the number of subjects n vary from 10 to 200. For each value of n, we repeatedly generate data and compute discriminability 100 times using Euclidean distance. It leaves us 100 estimates of discriminability $\hat{D}$. With this data generation scheme, we can actually compute the population discriminability D through numerical integration. It turns out to be 0.6150. Subtracting D from 100 $\hat{D}$s, we can estimate the distribution of estimation error. The figure 2 shows the difference between $\hat{D}$ and D. We can see that the mean of difference is centered around 0 and discriminabiity estimates $\hat{D}$ tends to converge to D as the number of subject increases.

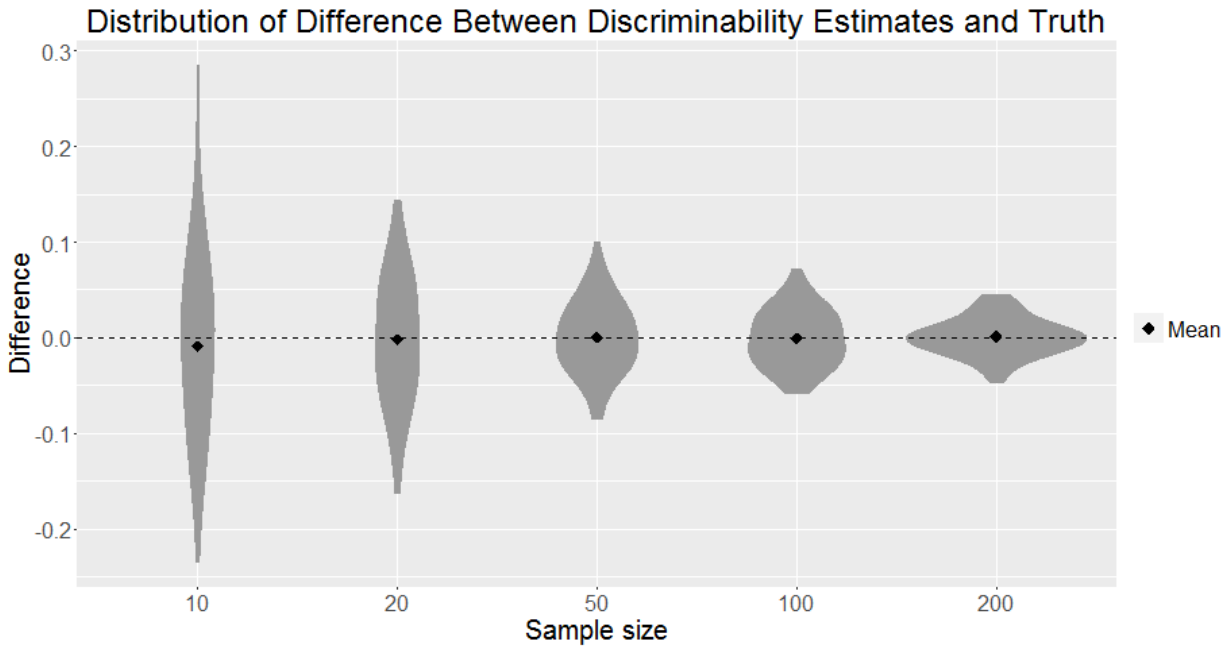

Figure 2: **Convergence of $\hat{D}$.** Distribution of difference between discriminability estimates and truth is ploted. The black dots indicate the mean over 100 repeats. As the number of subjects increases, the sample discriminability converges to the true population discriminability.

III.B.2 Parameter selection through discriminability

In this simulation, we consider the task of projecting 2-dimensional observations linearly into 1-dimensional space. Like in the previous experiment, we assume independent additive noise. In addition to $\boldsymbol{x}_{i,t}$, there is a binary class label $\boldsymbol{y}_i$ associated with subject i. The true physical property is Gaussian distributed conditioned on $\boldsymbol{y}_i$,

$$\boldsymbol{v}_i|\boldsymbol{y}_i = 1 \overset{i.i.d.}{\sim} \mathbb{G}(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}) \text{ and, } \boldsymbol{v}_i|\boldsymbol{y}_i = 0 \overset{i.i.d.}{\sim} \mathbb{G}(\begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix})$$

We consider two cases for the distribution $\boldsymbol{\epsilon}_{i,t}$. The first case is that $\boldsymbol{\epsilon}_{i,t}$ has larger variance in the first coordinate; the other case is that $\boldsymbol{\epsilon}_{i,t}$ has larger variance in the second coordinate, that is

$$\text{Case 1: } \boldsymbol{\epsilon}_{i,t} \sim \mathbb{G}(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}2 & 0\\0 & 1\end{bmatrix})$$

$$\text{Case 2: } \boldsymbol{\epsilon}_{i,t} \sim \mathbb{G}(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 4\end{bmatrix})$$

The noise is assumed to be independent of $\boldsymbol{v}_i$ and $\boldsymbol{y}_i$. The figure 3 shows the scatter plot of measurements. Under this generation scheme, the class signal only exists in the first coordinate. Therefore, the optimal linear projection should only keep the first coordinate.

We sample 200 subjects with v_i from each class conditional distribution. Furthermore, 2 measurements are sampled for each subject. We use both discriminability and principal component analysis (PCA) (20) to find the optimal linear projection. After finding the projection, we estimate two class conditional distribution through a kernel density estimator (21). The results of two cases are provided in two columns of figure 3. In the first case, both methods find the optimal linear projection which separates two classes. However, in the second case only discriminability recovers the optimal projection. PCA finds linear projection with little class signal.

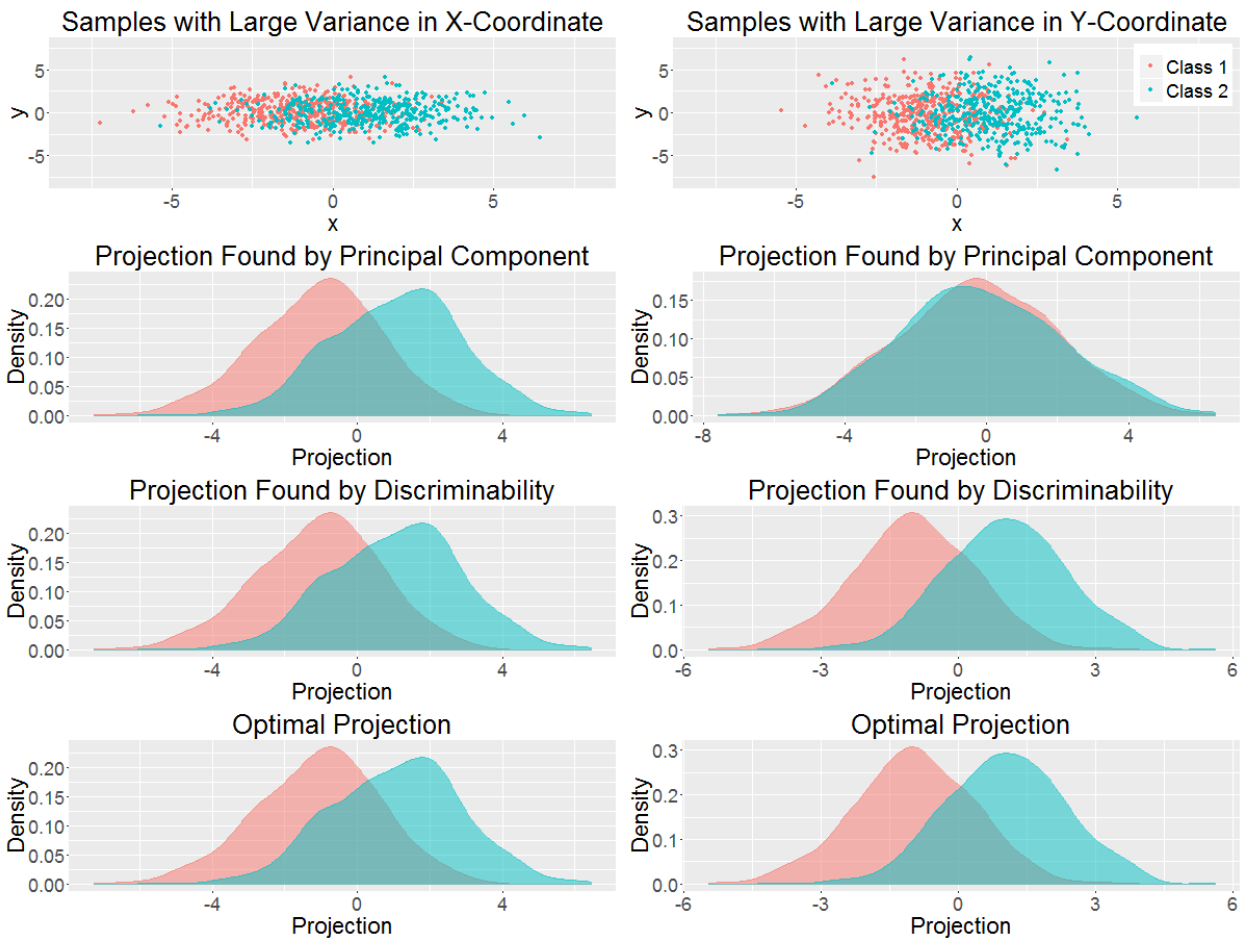

Figure 3: **Finding the Oprimal Projection.** Linear projections are computed using PCA and optimizing discriminability. Physical properties v_i of 200 subjects are sampled from 2-D two class conditional Gaussian distribution. 2 measurements are sampled for each subject with additive Gaussian noise. Noise could have large variance in x-coordinate or y-coordinate. The results for two cases are shown in two columns. Maximizing discriminability yields separated samples which have Bayes optimal classification error.

III.C Connectome Processing Applications

III.C.1 Optimal discriminability yields optimal predictive accuracy

In this experiment, we are going to investigate the thresholding step in processing resting state functional magnetic resonance imaging (fMRI). In fMRI processing, time series is first extracted for each region of interest (ROI) of brain (22). Then, a pairwise connectivity matrix is estimated through computing absolute Pearson correlation (23). To remove noise and obtain a binary graph, the pairwise connectivity matrix needs to be thresholded by a value which lies in $[0, 1]$ (24, 25). We would like to find the optimal value for the threshold. In addition to neuroimages, demographic information and five neuro factors (26) are also collected from each subject. We also want to find the threshold which leads to graphs with the best prediction performance.

HCP100 data set is used in this experiment (27). It contains data from 461 subjects with 4 measurements per subject. We let the threshold vary from 0 to 1. For each value of the threshold, binary graphs is constructed by thresholding correlations. Then, the discriminability is computed with Euclidean distance. In addition, sex, age and the neuro factors are predicted using k-nearest neighbor (28). For comparison,

another reliability statistics, namely image intraclass correlation coefficient is also computed which generalizes intraclass correlation coefficient for high dimensional observations (15). The discriminability, I2C2, and prediction errors versus the values of threshold are shown in figure 4. The threshold which maximizes discriminability is close to the thresholds yielding smallest predicting errors for three covariates.

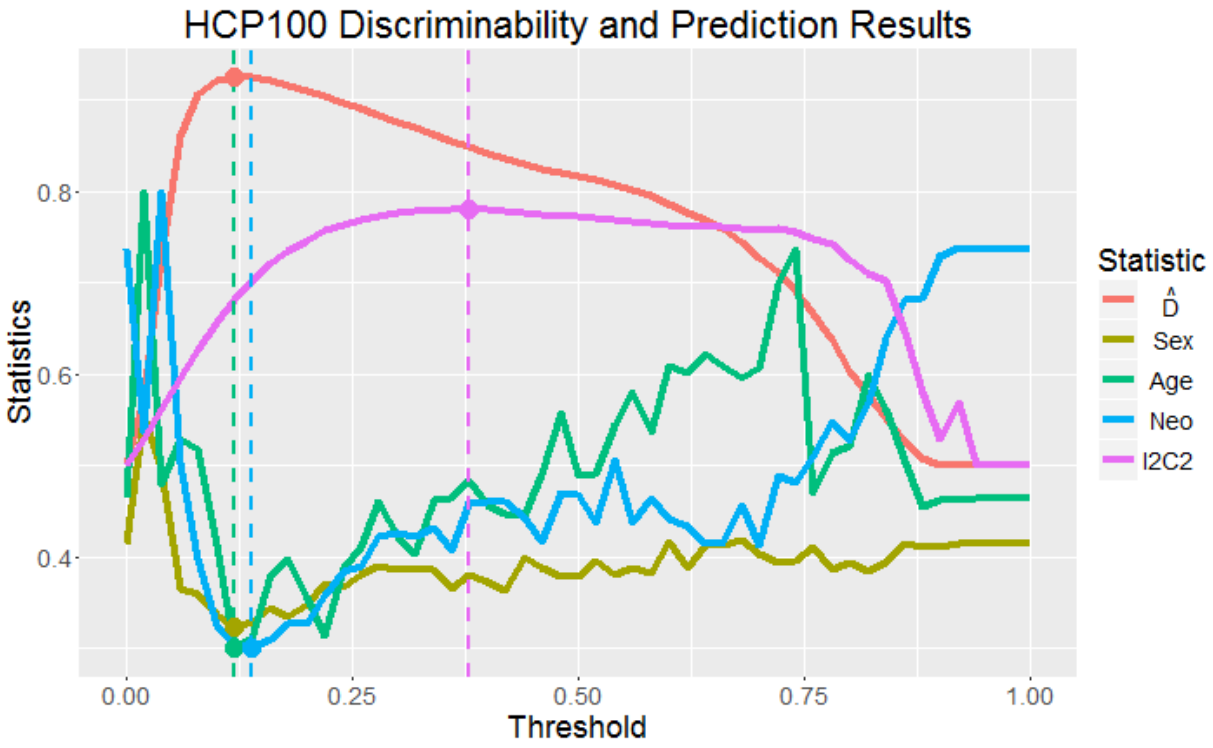

Figure 4: **Optimizing discriminability yields optimal prediction accuracy for multiple covariates.** HCP100 is used to investigate optimal threshold to convert correlation graphs into binary graphs. Curves are scaled to have similar value range. For each statistic, the optimal threshold and value pair is indicated by a circle on the curve. The threshold maximizing discriminability is close to the optimal thresholds for predicting three covariates.

III.C.2 fMRI processing pipelines

In this experiment, we are going to investigate the pre-processing options in acquiring resting state fMRI graphs (29). There have been a lot of steps proposed for pre-processing connectomes in the last decade. Here, we study a subset of them. In particular, we are interested in options include atlas (30), anatomical registration (31), temporal filtering (32), motion correction (33) and nuisance signal regression (34). We want to find the optimal pre-processing pipeline and the best decision for each option. We are going to index each pipeline by five letters which is explained in the table below.

Option	Letter
Atlas	C for CC200, H for HOX, A for AAL, D for DES (35, 36)
Anatomical Registration	F for FSL, A for ANTS (37, 38)
Temporal Filtering	F for frequency filtering, X for not (32)
Motion Correction	S for scrubbing, X for not (33)
Nuisance Signal Regression	G for global signal regression , X for not (34)

As an example, the best pipeline found is CFXSG which means the data is pre-processed using CC200 atlas, registered with FSL, no frequency filtering, with scrubbing and with global signal regression. There are

4 possible choices for atlas and 2 possible choices for other options. This leaves us 64 different combinations of options. We select 13 test-retest fMRI data sets with the number of measurements ranging from 50 to 300. These data sets are pre-processed by the 64 pipelines through the configurable pipeline for the analysis of connectomes (c-pac) (39). We also consider an extra rank conversion step which proves to be helpful in boosting discriminability. Rank conversion transforms a weighted undirected graph into a graph with rank weights. Specifically, in the previous experiment all edge weights are absolute correlations which lie in $[0, 1]$. In rank conversion step, for each edge in a graph, its weight w is replaced by the rank of w among all edge weights. If we denote a graph by a node set and an edge weight set pair (V, E) with $E = \{w_{i,j}\}$, rank conversion is a function

$$(V, E) \to (V, E') \text{ , where } E' = \{\text{rank}(w_{i,j})\}$$

The rank conversion is designed to improve signal to noise ratio by removing background noise. We carry out this step on the 13 data sets pre-processed by 64 pipelines. We compare the difference in discriminability with rank conversion and without rank conversion. The figure 5 shows the discriminability rank fMRI graphs and the discriminability graphs are provided in appendix. It turns out that rank conversion does help improving mean discriminability in all pipelines. When global signal regression is not performed, rank conversion significantly boosts discriminability.

There is notable variation in discriminability. The mean discriminability across 13 data sets can vary from 0.77 to 1.00. CFXSG turns out to be the best pipeline with maximum mean discriminability. Furthermore, we carried out a multi-factor analysis of variance test to study each option (40). Specifically, we fix decision for all options except one, and attempts to see whether there is significant difference in discriminability. It turns out that FSL, no frequency filtering, no scrubbing, global signal regression and rank conversion is better than their alternatives in terms of mean discriminability. However, no scrubbing is not statistical significantly better at level 0.05. FSL, no frequency filtering, global signal regression and rank conversion is better than their alternatives at level 0.001. Figure 6 shows the distribution of paired difference in discriminability.

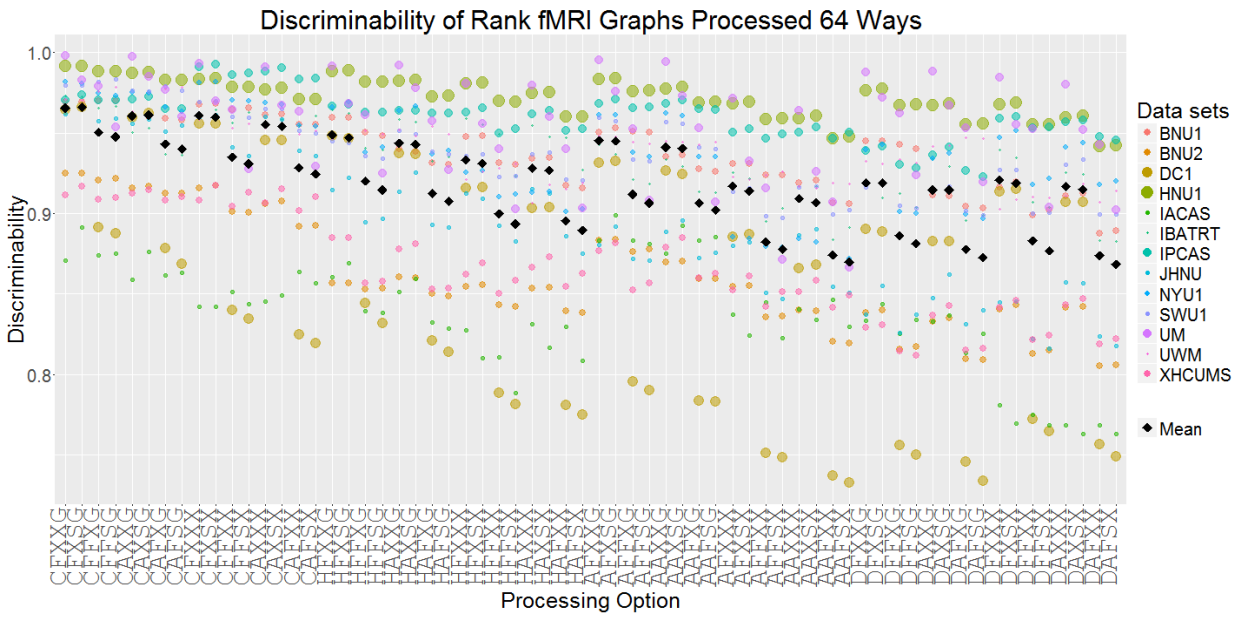

Figure 5: **Discriminability of rank fmri graphs from 13 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1,HNU1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS pre-processed by 64 pipelines are computed and shown in the top panel. Color of each dot indicates data set and size indicates the number of measurements in data set. The black square indicates the weighted mean discriminability across 13 data sets. CFXSG pipeline has the best mean discriminability across data sets.

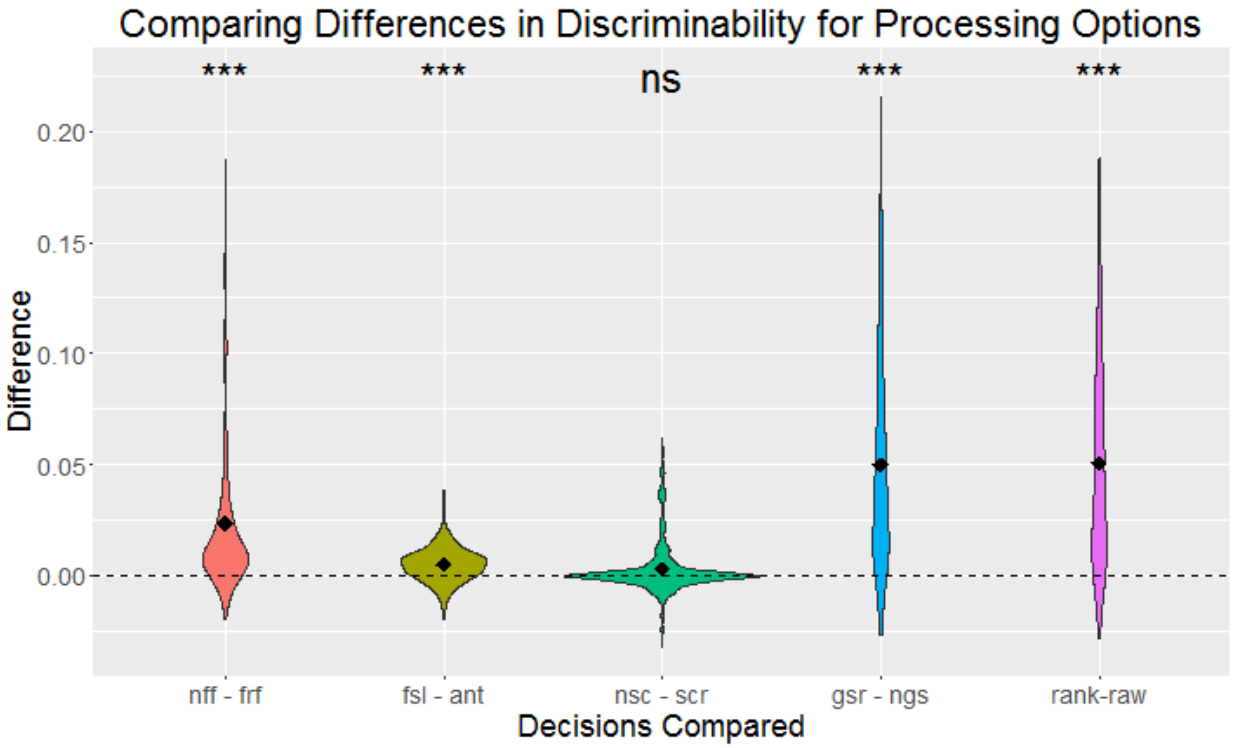

Figure 6: **Paired difference in discriminability of pre-processing options.** Difference in discriminability for each option is compared by fixing the other options and data set. The symbols at top indicates the significance. FSL, no frequency filtering, global signal regression and rank conversion are statistical significantly better than their alternatives at level 0.001. No scrubbing are not significantly better.

III.C.3 DTI experiment design

In this experiment, we consider the experiment design of collecting DTI data. In particular, we are interested the effect of b-value and number of directions on discriminability (41). We pick four data sets with different b-value and number of directions and compute discriminability. The result is show in the right panel of figure 7. We can see they have comparable discriminability. Given four data sets, we cannot conclude the optimal value for the parameters. It would be ideal if we could carry out a more controlled study with more data.

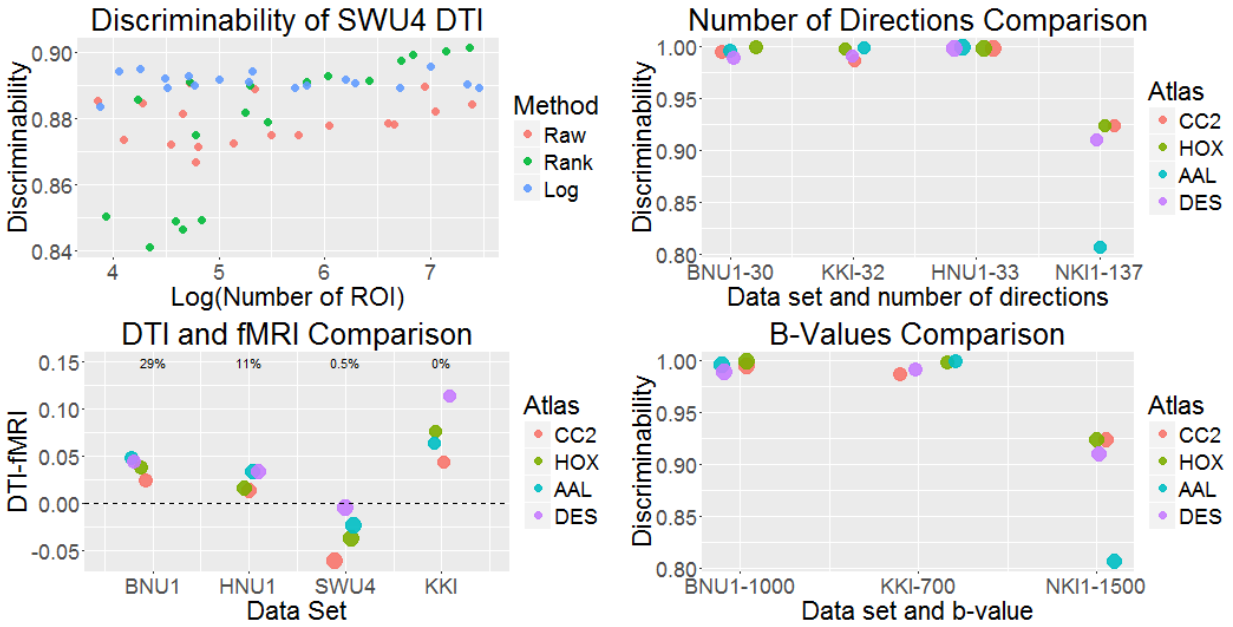

Figure 7: **Discriminability of DTI data sets.** The top left plot shows the discriminability of BNU1, HNU1, SWU4 and KKI registered with 15 atlases are computed and shown in the top panel. Raw, rank and log edges weights are considered. Discriminability of DTI and fMRI graphs are computed for BNU1, HNU1, SWU4 and KKI data set. The results are shown in the bottom left panel. The number at the top indicates the percentage of outliers in DTI data sets. After removing outliers, DTI data sets tend to be more discriminable than fMRI data sets. The right column shows the result of discriminability of different data sets with different b-value and number of directions.

III.C.4 DTI processing pipelines

In this experiment, we consider the processing of diffusion tensor imaging (DTI) (41). In particular, we are interested in finding the optimal number of ROI, and the optimal approach to process edge weights. BNU1, HNU1, SWU4 and KKI data sets are used in this experiment. We process four DTI data sets using 15 atlases with the number of ROI ranging from 48 to 1875 (42). For edge weights, we consider three options. First, raw edge weights are used which are fiber counts. Furthermore, we consider two alternatives: log weights and rank weights as discussed in the previous experiment. Top left panel of figure 7 shows the results. We see discriminability is basically stable across different atlases when raw and log edge weights are used. When using the rank weights, discriminability is low when the number of ROI is small. For three out of four data sets, the discriminability is very close to 1. As a consequence, we cannot find any statistical relationship between the number of ROI and discriminability.

III.C.5 fMRI vs. DTI

In this experiment, we want to compare discriminanility of fMRI and DTI data sets. Four data sets with both fMRI and DTI images are selected for the comparison. In processing fMRI data sets, the most dicriminable pipeline (*FXXG) is used. In processing DTI data sets, we use the raw edge weights. Some DTI measurements fail to pass the processing pipeline or have a dubious small number of edges. In this case, these measurements are labeled as outliers and removed from discriminability calculation. The result is shown in the bottom left panel of figure 7. Our conclusion is that DTI data sets after outlier removal have comparable discriminability as fMRI data sets. Actually, DTI measurements are better than fMRI in three out of four data sets.

IV Discussion

Summary We propose a non-parametric statistics of discriminability which is define to be the probability that within subject distance is smaller than across subject distance. We prove discriminability bounds Bayes prediction error. An estimator is designed to approximate the discriminability based on test-retest data set. We show the estimator is unbiased and converges to the discriminability asymptotically. We apply the discriminability framework under various setups in neuroimaging processing. We find the best processing pipeline for fMRI pre-processing and look into options in DTI processing. Furthermore, fMRI and DTI are shown to have comparable discriminability.

Next Steps First, more experiments should be carried out to analyze processing options. In particular, we could investigate processing of DTI more thoroughly given more data sets. Also, the effect of the number of ROI on discriminability is still not determined. Second, metrics other than Euclidean distance could be studied. Third, a testing procedure could be developed for comparing discriminability of multiple data sets.

V Appendix

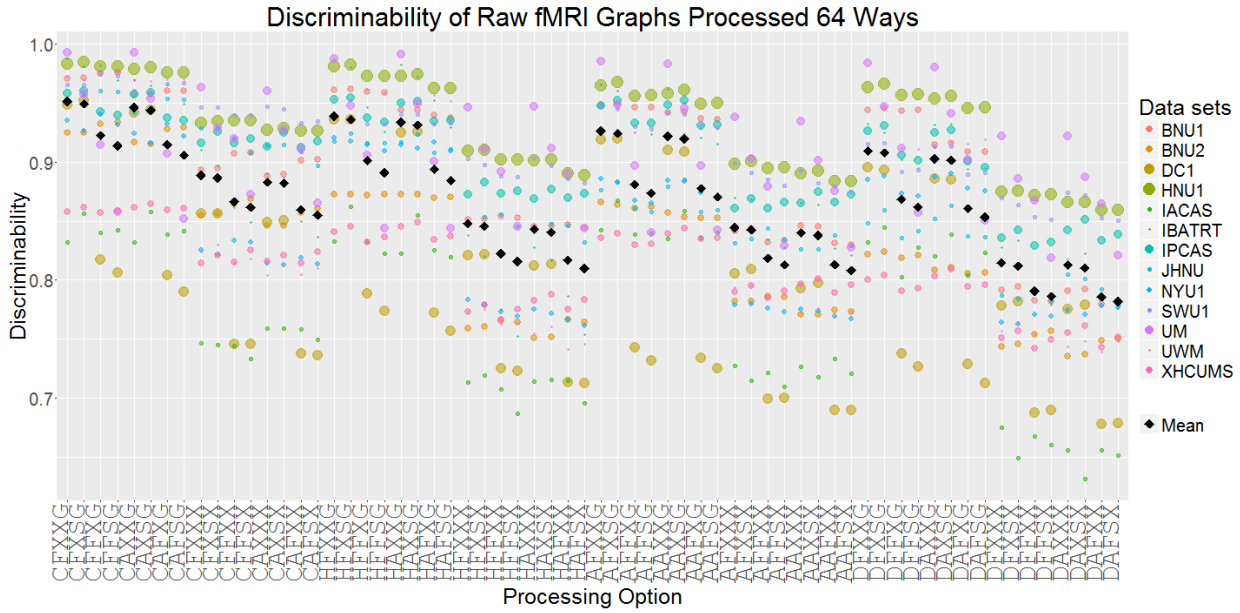

Figure 8: **Discriminability of raw fmri graphs from 13 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1, HNU1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS pre-processed by 64 pipelines are computed and shown in the top panel. Color of each dot indicates data set and size indicates the number of measurements in data set. The black square indicates the weighted mean discriminability across 13 data sets. CFXXG pipeline has the best mean discriminability across data sets.

Proof of Theorem 1. Consider the additive noise setting, that is $\boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t}$,

$$
\begin{aligned}
& \mathbb{P}(\delta_{i,t,t'} \leq \delta_{i,i',t,t''}) \\
= \; & \mathbb{P}(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i,t'}\| < \|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t''}\|) \\
= \; & P(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| < \|\boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t} - \boldsymbol{v}_{i'} + \boldsymbol{\epsilon}_{i',t''}\|) \\
\leq \; & \mathbb{P}(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| < \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\| + \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|) \\
= \; & \mathbb{P}(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\| < \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|) \\
= \; & \frac{1}{2}\mathbb{P}(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\| < \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\| | \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\| < 0) + \\
& \frac{1}{2}\mathbb{P}(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\| < \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\| | \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\| > 0) \\
= \; & \frac{1}{2} + \frac{1}{2}\mathbb{P}(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\| < \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\| | \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\| > 0) \\
= \; & \frac{1}{2} + \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big| < \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|) \\
= \; & 1 - \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big| > \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|)
\end{aligned}
$$

To bound the probability above, we bound the $\|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|$ and $\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big|$ separately. We start with the first term.

$$
\begin{aligned}
& \mathbb{E}(\|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|^2) \\
= \; & \mathbb{E}(\boldsymbol{v}_i^T \boldsymbol{v}_i + \boldsymbol{v}_{i'}^T \boldsymbol{v}_{i'} - 2\boldsymbol{v}_i^T \boldsymbol{v}_{i'}) \\
= \; & 2\sigma_2^2
\end{aligned}
$$

Here, σ_2^2 is the trace of covariance matrix of $\boldsymbol{v}_i$. We can apply Markov's Inequality,

$$
\mathbb{P}(\|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\| < t) \geq 1 - \frac{2\sigma_2^2}{t^2}
$$

Let σ_1^2 denote the trace of covariance matrix of $\epsilon_{i,t}$, and let a and b be two constants satisfy

$$
\mathbb{E}(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big|^2) \geq a^2\sigma_1^2
$$

$$
\frac{\mathbb{E}^2(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big|^2)}{\mathbb{E}(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big|)^4} \geq b
$$

Then, we can apply Paley-Zygmund Inequality (43),

$$
\mathbb{P}(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big|^2 > t^2) \geq b(1 - \frac{t^2}{a^2\sigma_1^2})^2
$$

Understand the fact that $\boldsymbol{v}$s and $\boldsymbol{\epsilon}$s are independent, we can combine the two inequalities and get a bound on $\mathbb{P}(\delta_{i,t,t'} \leq \delta_{i,i',t,t''})$.

$$
\begin{aligned}
& \mathbb{P}(\delta_{i,t,t'} \leq \delta_{i,i',t,t''}) \\
= \; & \mathbb{P}(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i,t'}\| < \|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t''}\|) \\
\leq \; & 1 - \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big| > \|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|) \\
\leq \; & 1 - \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\big|^2 > t^2)P(\|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|^2 < t^2) \\
\leq \; & 1 - \frac{1}{2}b(1 - \frac{t^2}{a^2\sigma_1^2})^2(1 - \frac{2\sigma_2^2}{t^2})
\end{aligned}
$$

Assume $a^2\sigma_1^2 \geq 2\sigma_2^2$ and set $t^2 = \sqrt{2}a\sigma_1\sigma_2$,

$$\mathbb{P}(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i,t'}\| < \|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t''}\|) \leq 1 - \frac{1}{2}b(1 - \frac{\sqrt{2}\sigma_2}{a\sigma_1})^3$$

By definition, $D = \mathbb{P}(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i,t'}\| < \|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t''}\|)$, we can have a bound on $\frac{\sigma_2}{\sigma_1}$.

$$\frac{\sigma_2}{\sigma_1} \geq \frac{a}{\sqrt{2}}(1 - (\frac{2-2D}{b})^{1/3}) \tag{6}$$

To obtain a bound on Bayes error, we apply Devijver and Kittler's result (44),

$$L \leq \frac{2\pi_0\pi_1}{1 + \pi_0\pi_1\Delta\mu^T\Sigma^{-1}\Delta\mu}$$

Here, π_0 and π_1 are prior probabilities for two classes. $\Delta\mu$ is the difference between means of two classes. Since $\boldsymbol{\epsilon}$ is assumed to be independent of $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\Delta\mu = \mathbb{E}(\boldsymbol{x}|\boldsymbol{y}=0) - \mathbb{E}(\boldsymbol{x}|\boldsymbol{y}=1) = \mathbb{E}(\boldsymbol{v}|\boldsymbol{y}=0) - \mathbb{E}(\boldsymbol{v}|\boldsymbol{y}=1)$$

Σ is the weighted covariance matrix of $\boldsymbol{x}$,

$$\begin{aligned}\Sigma &= \pi_0\mathsf{Var}(\boldsymbol{x}|\boldsymbol{y}=0) + \pi_1\mathsf{Var}(\boldsymbol{x}|\boldsymbol{y}=1) \\ &= \pi_0\mathsf{Var}(\boldsymbol{v}|\boldsymbol{y}=0) + \pi_1\mathsf{Var}(\boldsymbol{v}|\boldsymbol{y}=1) + \mathsf{Var}(\boldsymbol{\epsilon})\end{aligned}$$

If we further assume $\mathsf{Var}(\boldsymbol{\epsilon}) = \lambda\Sigma'$ where the trace of Σ is 1, then equation 6 implies $\lambda \leq \lambda_*$, where

$$\lambda_* = \frac{\sqrt{2}\sigma_2}{a(1 - (\frac{2-2D}{b})^{1/3})}$$

Hence, $\Sigma \leq \Sigma_*$ where

$$\Sigma_* = \pi_0\mathsf{Var}(\boldsymbol{v}|\boldsymbol{y}=0) + \pi_1\mathsf{Var}(\boldsymbol{v}|\boldsymbol{y}=1) + \lambda^*\Sigma'$$

Therefore, $\Sigma^{-1} \geq \Sigma_*^{-1}$, and we have

$$\begin{aligned}L &\leq \frac{2\pi_0\pi_1}{1 + \pi_0\pi_1\Delta\mu^T\Sigma^{-1}\Delta\mu} \\ &\leq \frac{2\pi_0\pi_1}{1 + \pi_0\pi_1\Delta\mu^T\Sigma_*^{-1}\Delta\mu}\end{aligned}$$

□

Proof of Lemma 1. By definition of $\hat{D}$,

$$\hat{D} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{s}\sum_{t'\neq t}^{s}\hat{D}_{i,t,t'}}{ns(s-1)}$$

The expectation of $\hat{D}_{i,t,t'}$ is actually D,

$$\begin{aligned}
& \mathbb{E}(\hat{D}_{i,t,t'}) \\
= & \frac{\sum_{i' \neq i}^{n} \sum_{t''=1}^{s} \mathbb{E}(\mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\})}{(n-1)s} \\
= & \frac{\sum_{i' \neq i}^{n} \sum_{t''=1}^{s} \mathbb{P}[\delta_{i,t,t'} \leq \delta_{i,i',t,t''}]}{(n-1)s} \\
= & \frac{\sum_{i' \neq i}^{n} \sum_{t''=1}^{s} D}{(n-1)s} \\
= & D
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \mathbb{E}(\hat{D}) \\
= & \frac{\sum_{i=1}^{n} \sum_{t=1}^{s} \sum_{t' \neq t}^{s} \mathbb{E}(\hat{D}_{i,t,t'})}{ns(s-1)} \\
= & \frac{\sum_{i=1}^{n} \sum_{t=1}^{s} \sum_{t' \neq t}^{s} D}{ns(s-1)} \\
= & D
\end{aligned}$$

This concludes that $\hat{D}$ is an unbiased estimator of discriminability D. □

Proof of Lemma 2. By definition of $\hat{D}$,

$$\begin{aligned}
\hat{D} &= \frac{\sum_{i=1}^{n} \sum_{t=1}^{s} \sum_{t' \neq t}^{s} \hat{D}_{i,t,t'}}{ns(s-1)} \\
&= \frac{\sum_{i=1}^{n} \sum_{t=1}^{s} \sum_{t' \neq t}^{s} \sum_{i' \neq i}^{n} \sum_{t''=1}^{s} \mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}}{ns(s-1)(n-1)s} \\
&= \frac{\sum_{i,i',t,t',t''} \mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}}{ns(s-1)(n-1)s}
\end{aligned}$$

In the last step, we simplify the sum, but keep in mind that $i \neq i'$ and $t \neq t'$. We show in the previous lemma that $\mathbb{E}(\hat{D}) = D$. To demonstrate that $\hat{D}$ converges to D in probability, it is suffice to show that $\mathsf{Var}(\hat{D}) \to 0$. Since then, by Chebyshev's inequality,

$$\mathbb{P}[|\hat{D} - D| \geq \epsilon] \leq \frac{\mathsf{Var}(\hat{D})}{\epsilon^2} \to 0$$

If we expand the variance of R,

$$\mathsf{Var}(\hat{D}) = \frac{\sum_{i,i',t,t',t''} \sum_{j,j',r,r',r''} \mathsf{Cov}(\mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}, \mathbb{I}\{\delta_{j,r,r'} \leq \delta_{j,j',r,r',r''}\})}{(ns(s-1)(n-1)s)^2}$$

There are $(ns(s-1)(n-1)s)^2$ covariance terms in the sum of nominator; however, most of them are actually 0. $\mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}$ is a function of $\boldsymbol{x}_{i,t}$, $\boldsymbol{x}_{i,t'}$ and $\boldsymbol{x}_{i',t''}$; therefore, is independent of any observations of subjects other than i and i'. This implies $\mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}$ is independent of $\mathbb{I}\{\delta_{j,r,r'} \leq \delta_{j,j',r,r',r''}\}$ as long as $\{i,i'\} \cap \{j,j'\} = \emptyset$. As a consqeunce, there are $(4n-6)(s(s-1)s) = ns(s-1)(n-1)s - (n-2)s(s-1)(n-3)s$ combinations of j, j', r, r', r'' such that covariance between $\mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}$ and $\mathbb{I}\{\delta_{j,r,r'} \leq \delta_{j,j',r,r',r''}\}$ maybe non-zero. Furthermore, the covariance must be less $\frac{1}{4}$ due to the fact that they are indicator random variables. Therefore, we have

$$
\begin{aligned}
\mathsf{Var}(\hat{D}) &= \frac{\sum_{i,i',t,t',t''} \sum_{j,j',r,r',r''} \mathsf{Cov}(\mathbb{I}\{\delta_{i,t,t'} \leq \delta_{i,i',t,t''}\}, \mathbb{I}\{\delta_{j,r,r'} \leq \delta_{j,j',r,r',r''}\})}{(ns(s-1)(n-1)s)^2} \\
&\leq \frac{\sum_{i,i',t,t',t''} (4n-6)(s(s-1)s)}{4(ns(s-1)(n-1)s)^2} \\
&= \frac{(4n-6)(s(s-1)s)}{4ns(s-1)(n-1)s} \\
&= \frac{4n-6}{4n(n-1)} \\
&\leq \frac{1}{n} \\
&\to 0 \text{ , as } n \to \infty
\end{aligned}
$$

As discussed before, this concludes that $\hat{D}$ converges to D in probability. □

A Bibliography

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011. 2

[2] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014. 2

[3] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145. 2

[4] L. Reiter, O. Rinner, P. Picotti, R. Hüttenhain, M. Beck, M.-Y. Brusniak, M. O. Hengartner, and R. Aebersold, "mprophet: automated data processing and statistical validation for large-scale srm experiments," *Nature methods*, vol. 8, no. 5, pp. 430–435, 2011. 2

[5] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," *Management science*, vol. 31, no. 2, pp. 150–162, 1985. 2

[6] A. M. Dale, "Optimal experimental design for event-related fmri," *Human brain mapping*, vol. 8, no. 2-3, pp. 109–114, 1999.

[7] J. R. Banga and E. Balsa-Canto, "Parameter estimation and optimal experimental design," *Essays in biochemistry*, vol. 45, pp. 195–210, 2008. 2

[8] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle, "The human brain is intrinsically organized into dynamic, anticorrelated functional networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9673–9678, 2005. 2

[9] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe *et al.*, "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4734–4739, 2010. 2

[10] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979. 2

[11] S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg, "The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework," *NeuroImage*, vol. 15, no. 4, pp. 747–771, 2002. 3

[12] M. L. Rizzo, G. J. Székely *et al.*, "Disco analysis: A nonparametric extension of analysis of variance," *The Annals of Applied Statistics*, vol. 4, no. 2, pp. 1034–1055, 2010. 3

[13] X.-N. Zuo, C. Kelly, J. S. Adelstein, D. F. Klein, F. X. Castellanos, and M. P. Milham, "Reliable intrinsic connectivity networks: test–retest evaluation using ica and dual regression approach," *Neuroimage*, vol. 49, no. 3, pp. 2163–2177, 2010. 2

[14] U. Braun, M. M. Plichta, C. Esslinger, C. Sauer, L. Haddad, O. Grimm, D. Mier, S. Mohnke, A. Heinz, S. Erk *et al.*, "Test–retest reliability of resting-state connectivity network characteristics using fmri and graph theoretical measures," *Neuroimage*, vol. 59, no. 2, pp. 1404–1412, 2012. 2

[15] H. Shou, A. Eloyan, S. Lee, V. Zipunnikov, A. Crainiceanu, M. Nebel, B. Caffo, M. Lindquist, and C. Crainiceanu, "Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (i2c2)," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 13, no. 4, pp. 714–724, 2013. 2, 10

[16] C. Yue, S. Chen, H. I. Sair, R. Airan, and B. S. Caffo, "Estimating a graphical intra-class correlation coefficient (gicc) using multivariate probit-linear mixed models," *Computational statistics & data analysis*, vol. 89, pp. 126–133, 2015. 3

[17] B. Yu *et al.*, "Stability," *Bernoulli*, vol. 19, no. 4, pp. 1484–1500, 2013. 2, 3

[18] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31. 4

[19] P. J. Bickel and K. A. Doksum, *Mathematical statistics: basic ideas and selected topics*. CRC Press, 2015, vol. 2. 6

[20] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002. 8

[21] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26. 8

[22] S. C. Strother, "Evaluating fmri preprocessing pipelines," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 2, pp. 27–41, 2006. 9

[23] X. Liang, J. Wang, C. Yan, N. Shu, K. Xu, G. Gong, and Y. He, "Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: a resting-state functional mri study," *PloS one*, vol. 7, no. 3, p. e32766, 2012. 9

[24] M. Hampson, B. S. Peterson, P. Skudlarski, J. C. Gatenby, and J. C. Gore, "Detection of functional connectivity using temporal correlations in mr images," *Human brain mapping*, vol. 15, no. 4, pp. 247–262, 2002. 9

[25] M. P. Van Den Heuvel and H. E. H. Pol, "Exploring the brain network: a review on resting-state fmri functional connectivity," *European Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010. 9

[26] P. T. Costa and R. R. MacCrae, *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources, 1992. 9

[27] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss *et al.*, "The human connectome project: a data acquisition perspective," *Neuroimage*, vol. 62, no. 4, pp. 2222–2231, 2012. 9

[28] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1. 9

[29] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional magnetic resonance imaging*. Sinauer Associates Sunderland, 2004, vol. 1. 10

[30] J. K. Mai, M. Majtanik, and G. Paxinos, *Atlas of the human brain*. Academic Press, 2015. 10

[31] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009. 10

[32] A. M. Smith, B. K. Lewis, U. E. Ruttimann, Q. Y. Frank, T. M. Sinnwell, Y. Yang, J. H. Duyn, and J. A. Frank, "Investigation of low frequency drift in fmri signal," *Neuroimage*, vol. 9, no. 5, pp. 526–533, 1999. 10

[33] J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen, "Spurious but systematic correlations in functional connectivity mri networks arise from subject motion," *Neuroimage*, vol. 59, no. 3, pp. 2142–2154, 2012. 10

[34] M. D. Fox, D. Zhang, A. Z. Snyder, and M. E. Raichle, "The global signal and observed anticorrelated resting state brain networks," *Journal of neurophysiology*, vol. 101, no. 6, pp. 3270–3283, 2009. 10

[35] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fmri atlas generated via spatially constrained spectral clustering," *Human brain mapping*, vol. 33, no. 8, pp. 1914–1928, 2012. 10

[36] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman *et al.*, "An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest," *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006. 10

[37] J. L. Andersson, M. Jenkinson, S. Smith *et al.*, "Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2," *FMRIB Analysis Group of the University of Oxford*, vol. 2, 2007. 10

[38] B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ants)," *Insight J*, vol. 2, pp. 1–35, 2009. 10

[39] S. Sikka, B. Cheung, R. Khanuja, S. Ghosh, C. Yan, Q. Li, J. Vogelstein, R. Burns, S. Colcombe, C. Craddock *et al.*, "Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac)," in *5th INCF Congress of Neuroinformatics, Munich, Germany*, vol. 10, 2014. 11

[40] J. F. Hair, “Multivariate data analysis,” 2009. 11

[41] C.-F. Westin, S. E. Maier, H. Mamata, A. Nabavi, F. A. Jolesz, and R. Kikinis, “Processing and visualization for diffusion tensor mri,” *Medical image analysis*, vol. 6, no. 2, pp. 93–108, 2002. 13, 14

[42] S. Mori, S. Wakana, P. C. Van Zijl, and L. Nagae-Poetscher, *MRI atlas of human white matter*. Elsevier, 2005. 14

[43] R. Paley and A. Zygmund, “On some series of functions,(3),” in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 28, no. 02. Cambridge Univ Press, 1932, pp. 190–205. 16

[44] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice hall, 1982. 17