Optimal Design 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

June 23, 2016

Contents

I Introduction

In this era of big data, many scientific, government, and corporate groups are collecting and processing massive data sets. 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. 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. 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.

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. 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 analyze datasets 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 Results

II.A Theory

II.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_ϕ. The parameter $\phi \in \Phi$ 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_ϕ 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 $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, which encodes all the details of the measurement and pre-processing. If both $g_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\phi}}$ 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. $D(\boldsymbol{\psi}, \boldsymbol{\phi})$ depends on three matters, namely measurement options $f_{\boldsymbol{\phi}}$, processing options $g_{\boldsymbol{\psi}}$ 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 $\boldsymbol{\phi}$ is a fixed element in $\boldsymbol{\Phi}$. Therefore, we are only interested in finding the best processing routine encoded by $\boldsymbol{\psi}$. 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 $\boldsymbol{\phi}$ 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.

II.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.

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

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.

II.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

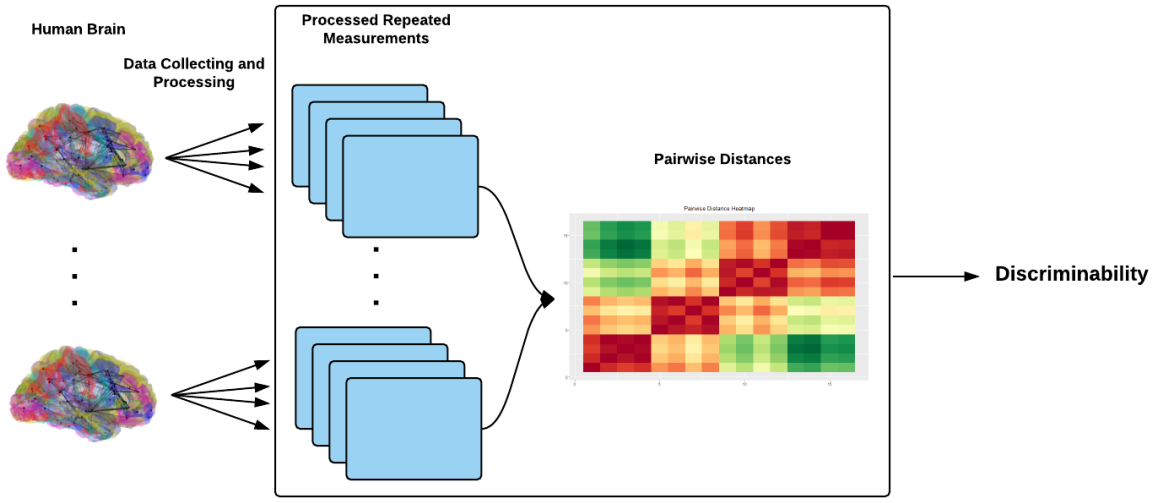

Figure 1: **Decision Making Through Discriminability Framework.** Test-retest data is collected from multiple subjects. The data is processed by a set of pipelines. For each processing pipeline, pairwise distances between observations are computed and discriminability is estimated.

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.

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*

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

II.B Simulations

II.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 this idea with simulation. We consider a simple case that g_ψ and f_ϕ together introduce independent Gaussian noise ϵ, that is

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

We sample $\boldsymbol{v}_i$ and $\boldsymbol{\epsilon}_{i,t}$ independently from standard Gaussian distribution. That is $\boldsymbol{v}_i \overset{i.i.d.}{\sim} \mathbb{G}(0,1)$ and $\boldsymbol{\epsilon}_{i,t} \overset{i.i.d.}{\sim} \mathbb{G}(0,1)$. For each subject, we sample two observations and 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 to estimate the distribution of $\hat{D}$. With this data generation scheme, we can compute the population discriminability D through numerical integration. It turns out to be 0.6150. The figure 2 shows the difference $\hat{D}$ and D. We can see from the figure that sample discriminabiity $\hat{D}$ converges to D as the number of subject increases.

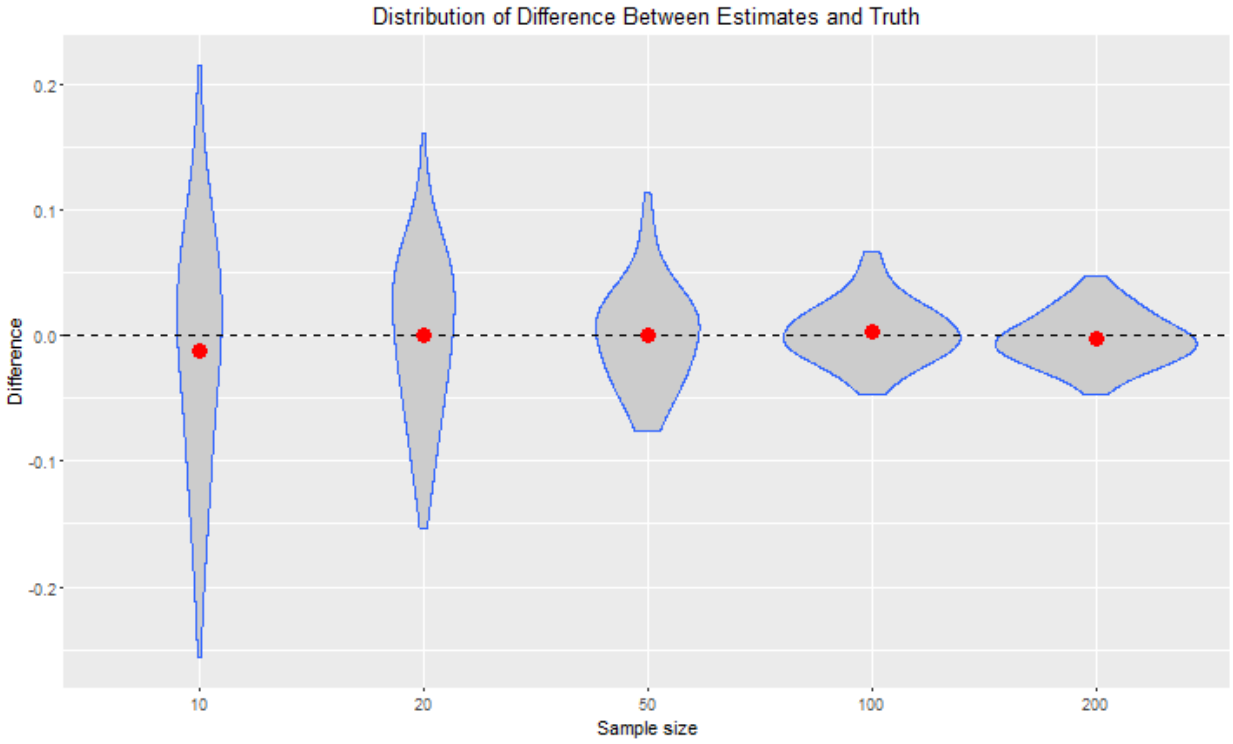

Figure 2: **Convergence of sample discriminability.** Distribution of difference between discriminability estimates and truth is ploted. The red dots indicate the mean over 100 repeats. As the number of subjects increases, the sample discriminability converges to the true population discriminability.

II.B.2 Parameter selection through discriminability

In this simulation, we consider the task of projecting 2-dimensional observations linearly into 1-dimensional space. Again, we assume 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 $\boldsymbol{y}_i$,

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

The optimal linear projection should keep two classes separated which is just keep the first dimension of the observations. 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 dimension; the other case is that $\boldsymbol{\epsilon}_{i,t}$ has larger variance in the second dimension. We use both discriminability and principal component analysis to find the optimal linear projection. The results of two cases are provided in two columns of figure 3. In the first case, both methods find the linear projection which separates two classes. However, in the second case only discriminability recovers the projection which separates two classes.

II.C Connectome Processing Applications

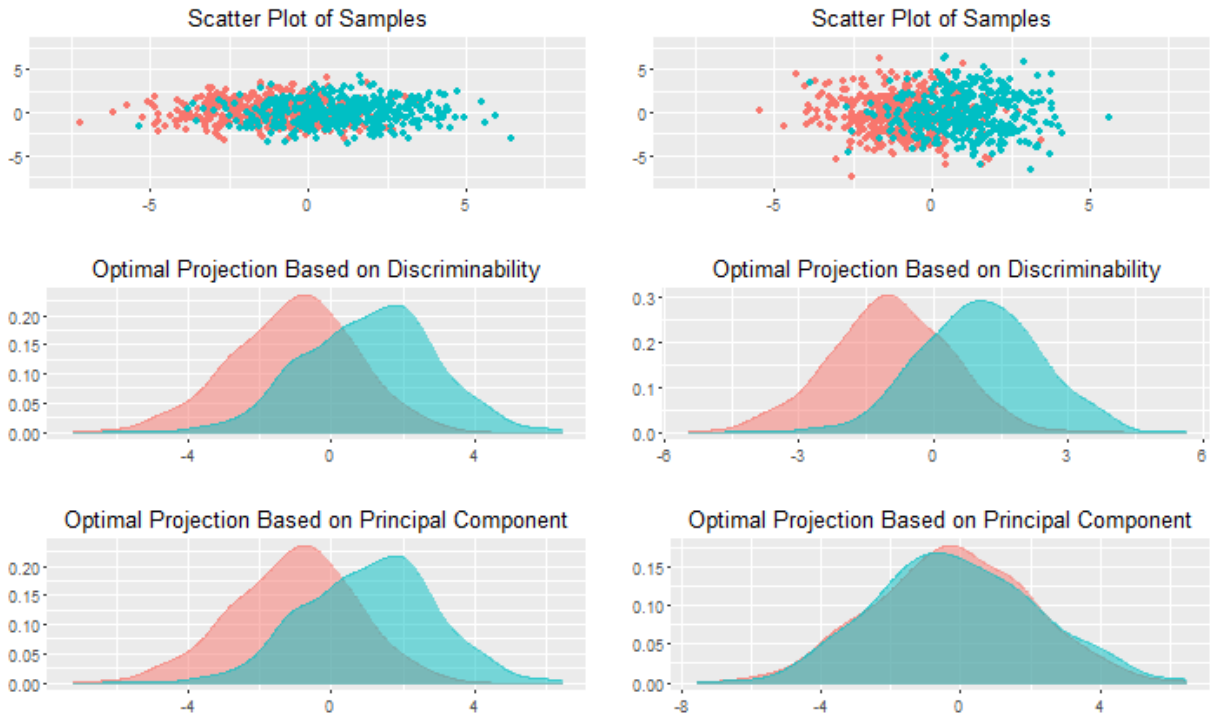

Figure 3: **Linear projection based on by PCA and Discriminability.** Linear projections are computed using PCA and optimizing Discriminability. Maximizing discriminability yield separated samples which have Bayes optimal classification error.

II.C.1 Optimal discriminability yields optimal predictive accuracy

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

HCP100 data set is used in this experiment. It has 461 subjects with 4 measurements per subject. We let the threshold vary from 0 to 1. For each value of threshold, the discriminability is computed; sex, age and the neuro factor are predicted using k-nearest neighbor classifier. As a comparison, another reliability statistics, namely image intraclass correlation coefficient (I2C2) is also computed. It is proposed by (Han Zhou et al.) which generalizes intraclass correlation coefficient for high dimensional observations. The discriminability, I2C2, and prediction errors are shown in figure 4. The threshold which maximizes discriminability is close to the thresholds yielding smallest predicting errors for three covariates.

II.C.2 fMRI processing pipelines

In this experiment, we are going to investigate the pre-processing options in acquiring fMRI graphs. There have been a lot of steps proposed for pre-processing connectomes in the last decade. We could only study a subset of them. In particular, we are interested in options include atlas, anatomical registration, temporal filtering, motion correction and nuisance signal regression. 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

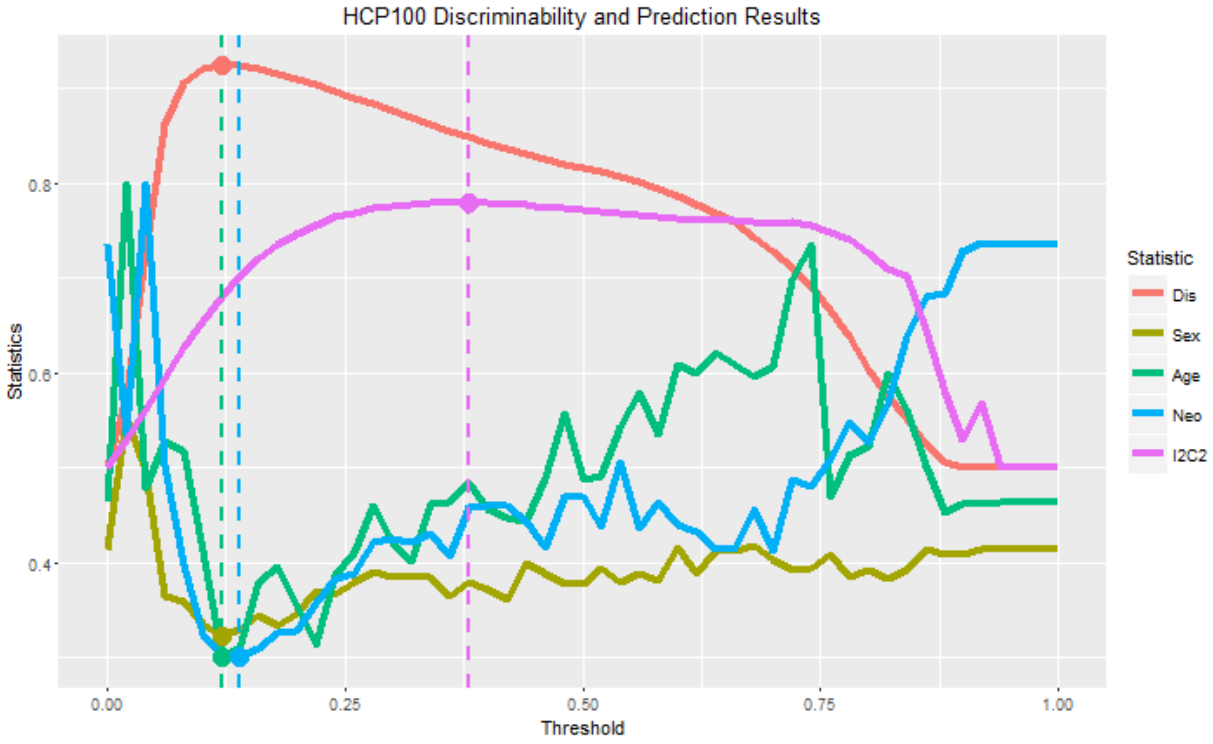

Figure 4: **Optimizing discriminability yields optimal prediction accuracy for multiple covariates.** HCP100 is used to investigate optimal threshold to convert correlation graphs into binary graphs. Curves are scaled to have similar value range. For each statistic, the optimal threshold and value pair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

letters which is explained in the table below.

Option	Letter
Atlas	C for CC200, H for HOX, A for AAL, D for DES
Anatomical Registration	F for FSL, A for ANTS
Temporal Filtering	F for frequency filtering, X for not
Motion Correction	S for scrubbing, X for not
Nuisance Signal Regression	G for global signal regression, X for not

For example, the best pipeline found is CFXXG which means the data is pre-processed using CC200 atlas, registered with FSL, no frequency filtering, no scrubbing and with global signal regression. There are 64 possible combinations of options. We pre-processed 12 data sets with the 64 pipleines and compute discriminability thereafter. The result is shown in figure 5. CFXXG turns out to be the best pipeline with maximal mean discriminability across all data sets. Furthermore, we carried out a multi-factor analysis of variance test to study each option. It turns out that FSL, no frequency filtering, no scrubbing and global signal regression is better than their alternatives in terms of mean discriminability. However, only no frequency filtering and global signal regression are statistical significantly better based on the test. Figure 6 shows the distribution of paired difference in discriminability.

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 replace 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 supposed to improve signal to noise ratio by removing background noise. We carry out this step on 12 data sets pre-processed by the 64 pipleines. We compare the difference in discriminability with rank conversion and without. Figure 7 shows the results. To summarize, rank conversion does help improving mean discriminability in all pipelines. When global signal regression is not performed, rank conversion significantly boosts discriminability.

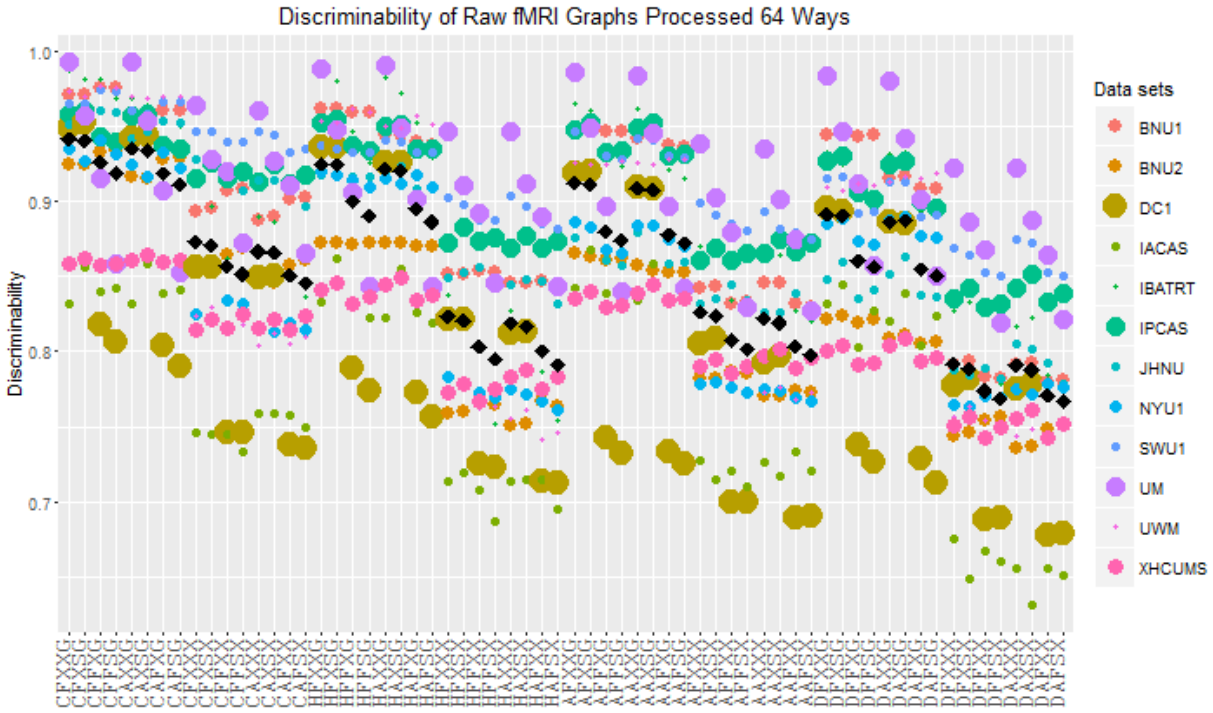

Figure 5: **Discriminability of raw fmri graphs from 12 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS pre-processed by 64 pipelines are computed. Color of each dot indicates data set and size indicates the number of measurements in data set. The black square indicates the mean discriminability across 12 data sets. CFXXG pipeline has the best mean discriminability across data sets.

II.C.3 DTI processing pipelines

In this experiment, we consider the processing of diffusion tensor images (DTI). In particular, we are interested in finding the optimal number of region of interest (ROI), and the optimal approach to process edge weights. We process four DTI data sets using 15 atlases with the number of ROI ranging from 48 to 1875. 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 conversion. Figure 8 shows the results. We see discriminability is stable across different atlases when using raw and log edge weights. When using the rank weights, discrinability 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.

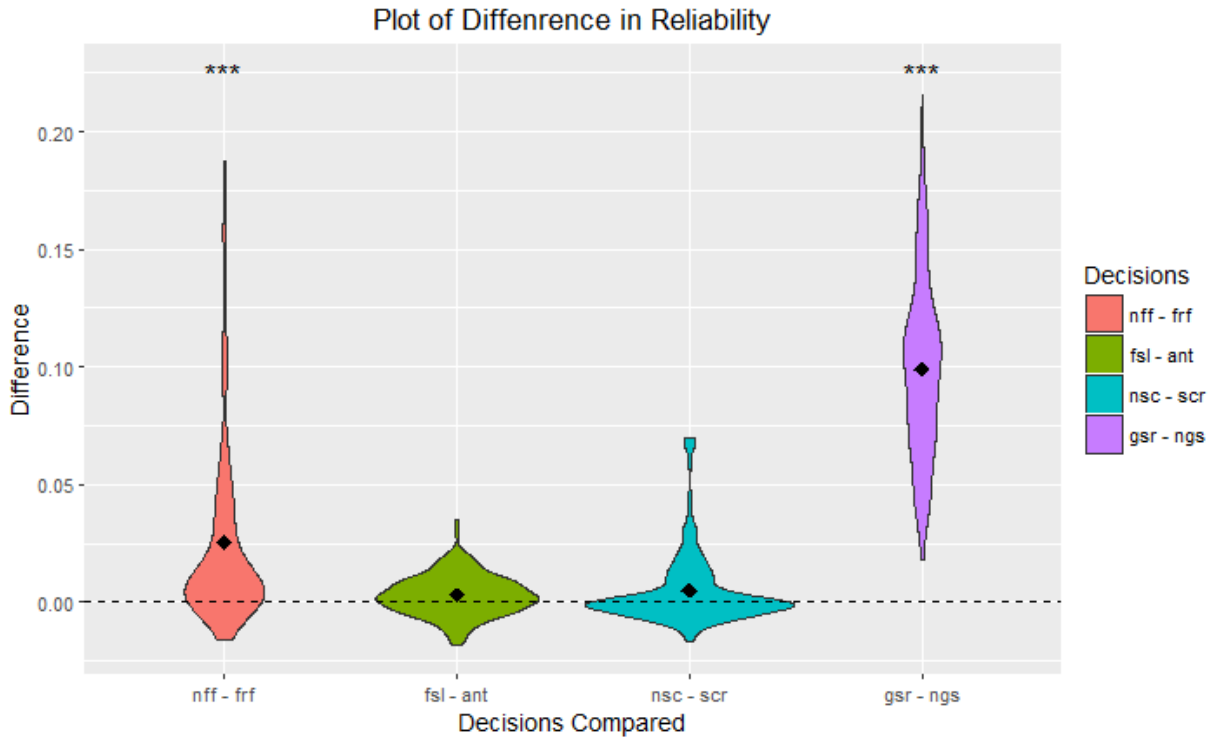

Figure 6: **Paired difference in discriminability of pre-processing options.** Difference in discriminability for each option is compared by fixing the other options and the data set. no frequency filtering and global signal regression are statistical significantly better than the alternatives. FSL and no scrubbing are not significantly better than the alternatives.

II.C.4 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 measurements are selected for the comparison. In processing fMRI data sets, the most dicriminable pipeline (CFXXG) is used. In processing DTI data sets, some measurements fail to pass the processing pipeline or have a very small number of edges. In this case, these measurements are labeled as outliers and removed from discriminability calculation. The result is shown in figure 9. 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.

III 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 discriminabilility 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.

Related Work Image intraclass correlation coefficient (I2C2) is proposed by Shou H et al to measure reliability. It generalizes classic image intraclass coefficient to high dimensional observations. It relies on the assumption that noise is additive and computes reliability estimates based on the traces of within subject

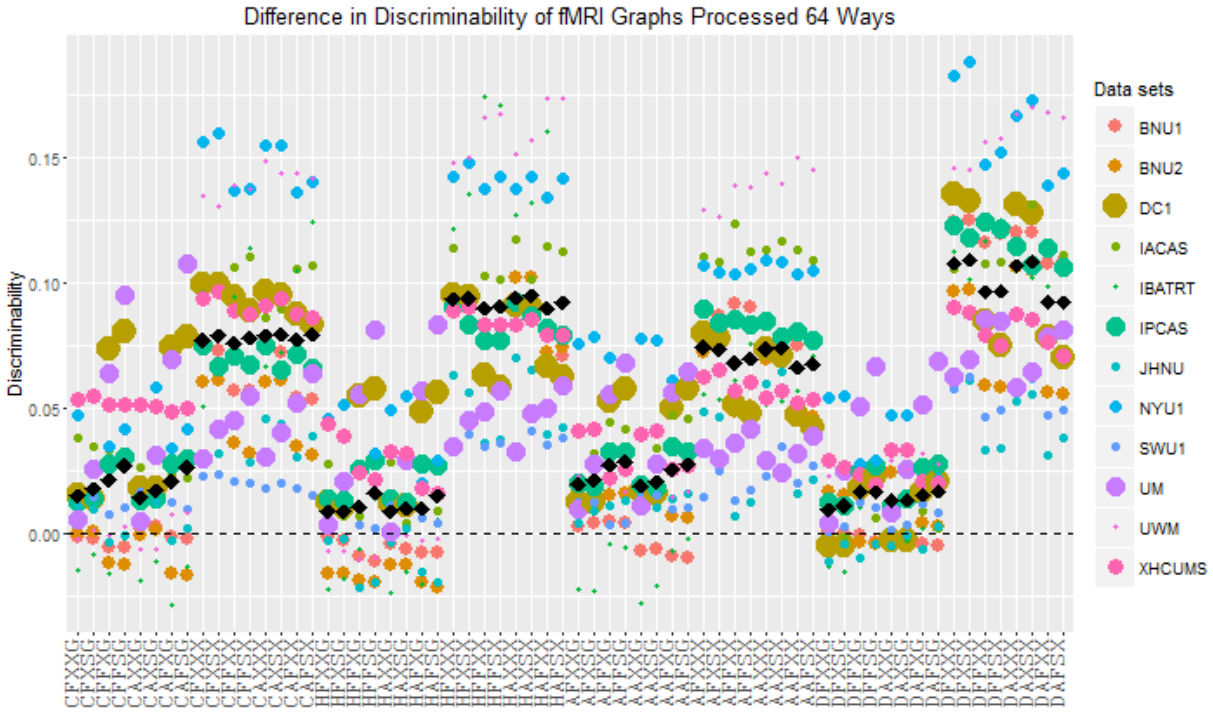

Figure 7: **Paired difference in discriminability between rank and raw graphs.** Difference in discriminability for rank and raw fmri graphs are computed for 12 data sets processed using 64 pipelines. Overall, rank fMRI graphs are more discriminable than raw fMRI graphs.

and across subject covariance matrix.

Graphical intraclass correlation coefficient (GICC) is another reproducibility measure proposed by Chen Y et al. It is designed specifically for the case when data of interest are binary graphs. It takes a parametric approach by first estimating latent edge feature vectors and computes GICC based on variation of latent edge feature vectors.

Distance components (DISCO) is proposed by MARIA L. RIZZO AND GBOR J. SZKELY as a measure of dispersion. 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 multiple set samples are drawn from the same distribution or not.

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 two discriminability.

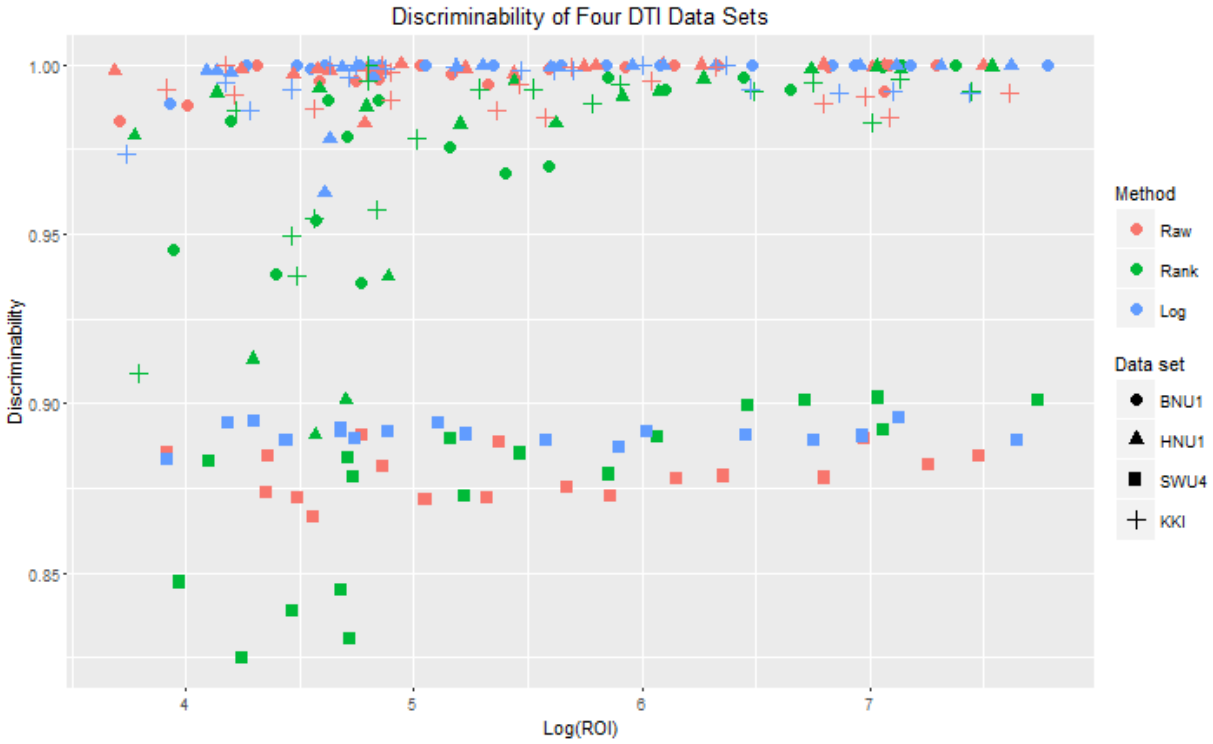

Figure 8: **Discriminability of four DTI data sets.** Discriminability of BNU1, HNU1, SWU4 and KKI registered with 15 atlases are computed. Raw, rank and log edges weights are considered.

IV Appendix

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}
$$

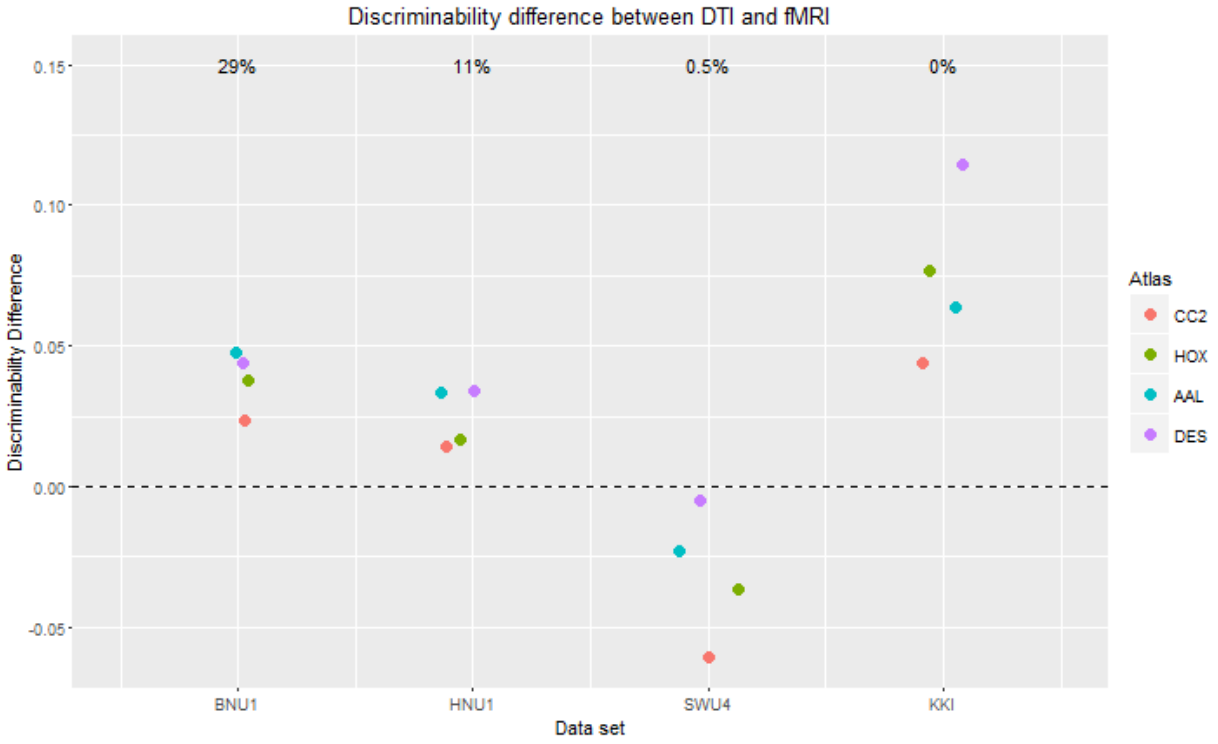

Figure 9: **Paired difference in discriminability between dti and fmri data sets.** Discriminability of DTI and fMRI graphs are computed for BNU1, HNU1, SWU4 and KKI data set. The number at the top indicates the percentage of outliers in DTI data sets. After removing outliers, DTI data sets are more discriminable than fMRI data sets.

To bound the probability above, we bound the $\|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\|$ and $\left|\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| - \|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|\right|$ 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

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

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

Then, we can apply Paley-Zygmund Inequality,

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

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,

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

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,

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

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

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

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

A Bibliography