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

November 26, 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 data sets become benchmark or reference data sets. 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 framework should provide a measure of consistency of data collection and processing, 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 this measure 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, which can be used 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. Furthermore, one sample and two sample tests for discriminability are developed. These tests determines the statistical significance of hypothesis of interest based on the discriminability estimator.

Numerical simulations are conducted to demonstrate the basic property of our discriminability estimator and tests 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 most 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, should one perform frequency filtering 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 most 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 (10–18). 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 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.

- Correspondence curve is introduced to study reproducibility of signals (18). It first ranks all the signals by a scalar score within each replicates, and then the proportion of signals which ranked among top percentile of both replicates is computed. It generalizes Spearman's rank correlation coefficient and can be used to detect irreproducible signals. In our studies, we are interested in reproducibility of measurements instead of signals and the measurements are vectors or matrices, which makes this approach not immediately applicable.

- 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_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 $\delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'}) < \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''})$, 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:

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

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,

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

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 also 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 decisions on the 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. Furthermore, we have developed a one sample test procedure to determine whether there are subject specific information in the data, and a two sample test procedure to compare two processing pipelines. 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 subject 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 (19)

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

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.

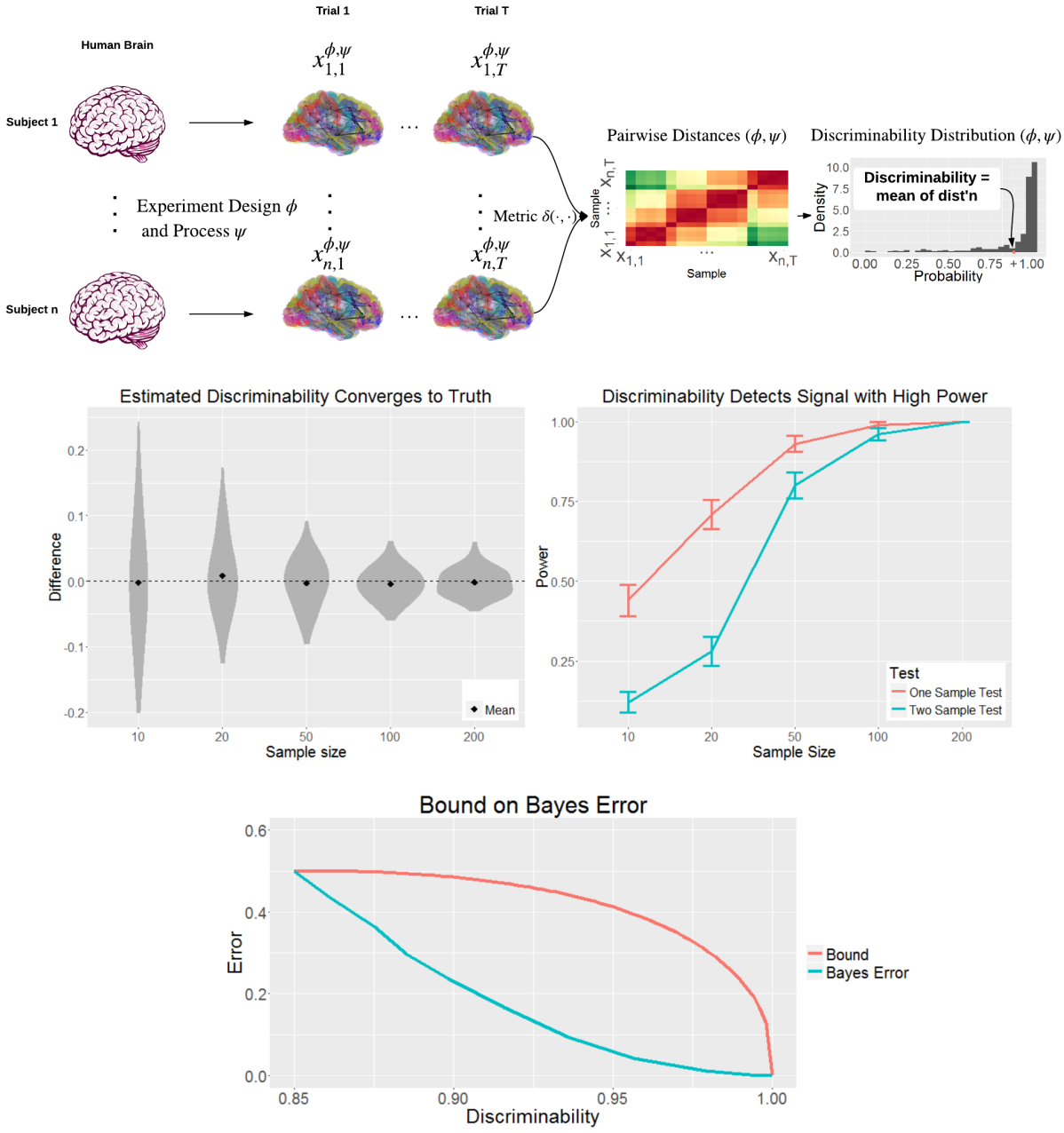

Figure 1: **Decision Making Through Discriminability Framework.** The top panel shows the decision framework of discriminability. Test-retest data set is collected under experiment design options $/phi$ and processed by pipeline ψ. The pairwise distances of all measurements are computed using a metric $\delta(\cdot,\cdot)$. 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.
Convergence of $\hat{D}$. Distribution of difference between discriminability estimates and truth is shown in the middle left panel. The physical property and noise are generated from standard Gaussian distribution as described in the simulation section. The black dots indicate the mean over 100 repeats. As the number of subjects increases, the sample discriminability converges to the true population discriminability.
Discriminability Test Power. One sample and two sample test power of discriminability with varying sample size is shown in the middle right panel. The physical property and additive noise are generated from standard Gaussian distribution as described in the simulation section. At level of 0.05, the power is estimated based on 100 repeats. The power of two tests become close to 1 with more than 100 samples.
Discriminability Bound Bayes Error. Bayes error and theoretical bound based on Theorem 1 are shown. Samples are generated from conditional Gaussian distribution with additive Gaussian noise.

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

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

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 (20).

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 One sample test for discriminability

In applications, we sometimes are interested in whether there is any subject specific information in the data. In other words, we want to know whether $\boldsymbol{x}_{i,t}$ is independent of $\boldsymbol{v}_i$. Formally, it is equivalent to test the hypothesis that $\boldsymbol{x}_{i,t}$ is independent of $\boldsymbol{v}_i$. If we fail to reject the hypothesis, it implies the measurement $\boldsymbol{x}_{i,t}$ reveals no information of true physical property $\boldsymbol{v}_i$. As a consequence, $\boldsymbol{x}_{i,t}$ is independent of any phenotype $\boldsymbol{y}_i$, and there is no hope in predicting $\boldsymbol{y}_i$ based on $\boldsymbol{x}_{i,t}$. If this is the case, the researchers should consider collecting more data or processing data differently. Since $\boldsymbol{v}_i$ is unobserved and $\boldsymbol{y}_i$ is unknown, a direct independence test is not applicable. We consider a test through discriminability. If measurements are independent of physical properties, $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{i',t'}$ should follow the same distribution. In this case, within subject distances should not differ across subject distances in distribution; therefore, discriminability should be 0.5. Conversely, we have the following lemma.

Lemma 3. *Under some regularity conditions, disrcriminability is 0.5 implies measurements are independent of physical property, that is*

$$D = 0.5 \Rightarrow \boldsymbol{x} \perp \boldsymbol{v}.$$

An immediate consequence of the lemma is that we can test the null hypothesis that measurements $\boldsymbol{x}_{i,t}$ are independent of any phenotype $\boldsymbol{y}_i$ through testing the hypothesis whether discriminability is 0.5.

$$H_0 : \boldsymbol{x} \perp \boldsymbol{y} \text{ , and } H_A : \boldsymbol{x} \not\perp \boldsymbol{y}.$$

We reject the null hypothesis above when there are strong evidences suggesting that $D > 0.5$.

We have two valid approaches to determine $D > 0.5$ through discriminability estimate $\hat{D}$. The first approach takes the advantage of 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 or equal to $1/n$. 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. This approach is computationally simple; however, generally has small power due to the bound on variance is not tight. The second approach based on estimating a null distribution for $\hat{D}$ through permutation. In particular, we randomly permute subject labels for each trial and then estimate discriminability based on permuted labels. We repeat this procedure a large number of times and find the 95^{th} quantile of permuted discriminability estimates. If $\hat{D}$ is less than the 95^{th} quantile, we do not reject the null hypothesis; otherwise, we reject the null hypothesis. This approach has larger power than the first approach, the only downside is that estimating discriminability for permuted samples takes sometime. In most applications, with less than a few hundred measurements, we recommend using the second approach.

III.A.5 Two sample test for discriminability

In many applications, we want to know whether one data processing pipeline $\boldsymbol{\psi}_1$ yields more discriminable data set than another pipeline $\boldsymbol{\psi}_2$. Based on the theory, by choosing the processing pipeline with larger discriminability, we can have a lower bound on Bayes prediction error. To achieve this, we consider testing the null hypothesis that two discriminabilities are equal:

$$H_0 : D(\boldsymbol{\psi}_1) = D(\boldsymbol{\psi}_2) \text{ , and } H_A : D(\boldsymbol{\psi}_1) > D(\boldsymbol{\psi}_2).$$

However, $D(\boldsymbol{\psi}_1)$ and $D(\boldsymbol{\psi}_2)$ are not known to us, we have to decide based on estimators $\hat{D}(\boldsymbol{\psi}_1)$ and $\hat{D}(\boldsymbol{\psi}_2)$. We have two valid approaches to test this. The first approach takes the advantage of the bound on variance of $\hat{D}$ which we derived in proving Lemma 2. Specifically, we show that the variance of discriminability estimate is bounded by $1/n$. Therefore, we can derive two confidence intervals centered at $\hat{D}(\boldsymbol{\psi}_1)$ and $\hat{D}(\boldsymbol{\psi}_2)$. Then, the null hypothesis is rejected if two confidence intervals does not overlap. Unfortunately, due to the fact that inequalities are not tight, this approach has very low power. For this method to work, the number of subjects n usually needs to be larger than a thousand. It is impractical for most of the data set. The second approach estimates null distribution of $\hat{D}(\boldsymbol{\psi}_1) - \hat{D}(\boldsymbol{\psi}_2)$ through bootstrapping. We can bootstrap copies of the original data set and compute discriminability on bootstrapped data set to approximate the null distribution. Specifically, let $\hat{D}^{(i)}(\boldsymbol{\psi}_j)$ denote the discriminability estimate for ith bootstrapped copy with data processed by pipeline j. If the null hypothesis is true, $\hat{D}(\boldsymbol{\psi}_1) - \hat{D}(\boldsymbol{\psi}_2)$ should have similar distribution as $\hat{D}^{(i)}(\boldsymbol{\psi}_j) - \hat{D}^{(i')}(\boldsymbol{\psi}_j)$. To bootstrap a copy of original data set, we need to make sure that the copy have the same number of subjects and number of measurements per subject as the original data set. To bootstrap measurements for a subject, we first randomly choose two subjects from original data sets, and then take a random convex linear combination of measurements of these two subjects. We keep repeating this step until the bootstrapped data set has the same number of subjects as the original data set, and discriminability $\hat{D}^{(i)}(\boldsymbol{\psi}_j)$ is estimated. To approximate the null distribution, a large number of bootstrapped discriminabilities are computed, and their pairwise differences $\hat{D}^{(i)}(\boldsymbol{\psi}_j) - \hat{D}^{(i')}(\boldsymbol{\psi}_j)$ are used to compute a p-value for $\hat{D}(\boldsymbol{\psi}_1) - \hat{D}(\boldsymbol{\psi}_2)$. We should notice that bootstrapped data tends to be less discriminable than the original data due to the fact bootstrapped subjects are closer to each other. However, we only use differences in bootstrapped discriminability. The algorithm 1 summarizes the steps to estimate p-value for testing $D(\boldsymbol{\psi}_1) = D(\boldsymbol{\psi}_2)$.

Pseudocode 1 Test the null hypothesis $D(\psi_1) = D(\psi_2)$

Input: Data set, Pipeline ψ_1, Pipeline ψ_2
Output: Reject or do not reject the null hypothesis
  Process the data set with pipelines ψ_1 and ψ_2
  Compute $\hat{D}(\psi_1)$ and $\hat{D}(\psi_2)$
  for i in 1 through number of repeats **do**
    for j in 1 through number of subjects **do**
      Randomly select two subjects from data set
      Linearly combine measurements of these subjects
    end for
    Form two bootstrapped data sets processed by ψ_1 and ψ_2
    Compute $\hat{D}^{(i)}(\psi_1)$ and $\hat{D}^{(i)}(\psi_2)$
  end for
  Compute pairwise differences $\hat{D}^{(i)}(\psi_1) - \hat{D}^{(i')}(\psi_1)$ and $\hat{D}^{(i)}(\psi_2) - \hat{D}^{(i')}(\psi_2)$
  Compute p-value which is the fraction of times that $\hat{D}(\psi_1) - \hat{D}(\psi_2) > \hat{D}^{(i)}(\psi_j) - \hat{D}^{(i')}(\psi_j)$
  Reject the null hypothesis if p-value is less than 0.05.

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_ψ and f_ϕ together introduce independent additive Gaussian noise ϵ, that is

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

$\boldsymbol{v}_i$ 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, which turns out to be 0.615. Subtracting D from 100 $\hat{D}$s, we can estimate the distribution of estimation error. The bottom left panel of figure 1 shows the difference between $\hat{D}$ and D. We can see that the mean of difference is centered around 0, and discriminabiity estimates $\hat{D}$ converge to D as the number of subject increases.

III.B.2 Test power of discriminability

In this section, we investigate the power of one sample and two sample tests for discriminability through simulation. For one sample test for discriminability, we consider the simple additive noise case as in the previous section, that is,

$$\boldsymbol{x}_{i,t} = \boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t}, \boldsymbol{v}_i \overset{i.i.d.}{\sim} \mathbb{G}(0,1); \boldsymbol{\epsilon}_{i,t} \overset{i.i.d.}{\sim} \mathbb{G}(0,1).$$

Again, we let the number of subjects n to increase from 10 to 200, and for each subject we sample two observations. For each generated data set, we first estimate discriminability, and a p-value is computed based on 100 permutations. The null hypothesis that $D = 0.5$ is rejected when p-value is less than 0.5. Under this data generation scheme, the true discriminability is 0.615; therefore, rejecting the null hypothesis is preferred. For each value of n, we independently generate 100 data sets and perform the one sample test. The fraction of times in which the null hypothesis is rejected with its standard error is shown in the

bottom right panel of Figure 1. The power of the test quickly increases as the number of subjects increases, and is close to 1 with more than 50 subjects.

For two sample test for discriminability, we generate two sets of measurements $\boldsymbol{x}_{i,t}^1$ and $\boldsymbol{x}_{i,t}^2$. The superscript is to denote the pipeline which the measurements come from. The noise is still Gaussian and additive; however, the pipeline 1 has smaller noise level compared to pipeline 2. Specifically,

$$\boldsymbol{x}_{i,t}^1 = \boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t}^1; \boldsymbol{x}_{i,t}^2 = \boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t}^2; \boldsymbol{v}_i \overset{i.i.d.}{\sim} \mathbb{G}(0,1); \boldsymbol{\epsilon}_{i,t}^1 \overset{i.i.d.}{\sim} \mathbb{G}(0,0.25); \boldsymbol{\epsilon}_{i,t}^1 \overset{i.i.d.}{\sim} \mathbb{G}(0,1).$$

We let the number of subjects n to increase from 10 to 200, and for each subject we sample two observations. We generate measurements for both pipelines, and apply the two sample test procedure as described in Algorithm 1. Under this data generation scheme, the true discriminability of pipeline 1 is larger than that of pipeline 2; therefore, rejecting the null hypothesis is preferred. For each value of n, we independently generate 100 pairs of data sets and perform the two sample test. The fraction of times in which the null hypothesis is rejected with its standard error is shown in the bottom right panel of Figure 1. The power of the test quickly increases as the number of subjects increases, and is close to 1 with more than 100 subjects.

III.B.3 Parameter selection through discriminability

In this simulation, we consider the task of projecting 2-dimensional measurements 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 2 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) (21) to find the optimal linear projection. After finding the projection, we estimate two class conditional distribution through a kernel density estimator (22). The results of two cases are provided in two columns of figure 2. 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.

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 (23). Then, a pairwise connectivity matrix is estimated through computing absolute Pearson correlation (24). To remove noise and obtain a binary graph, the pairwise connectivity matrix needs to be thresholded by a value which lies in $[0,1]$ (25, 26). We would like to find the optimal value

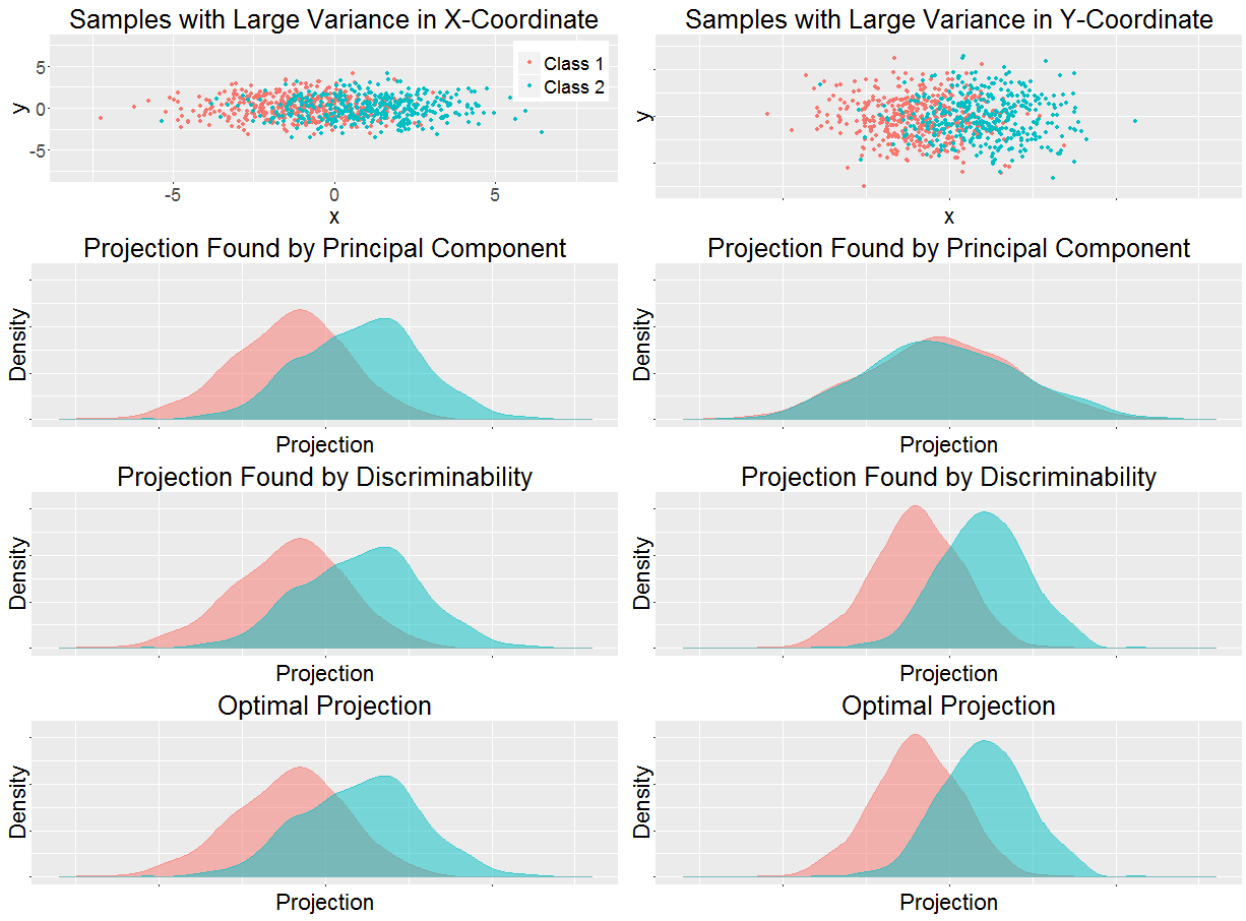

Figure 2: **Finding the Oprimal Projection.** Linear projections are computed using PCA and optimizing discriminability. Physical properties $\boldsymbol{v}_i$ of 200 subjects are sampled from 2-D two class conditional Gaussian distribution. 2 measurements are sampled for each subject with additive Gaussian noise. Noise could have either large variance in x-coordinate or y-coordinate. The details of generating data can be found in simulations section. The results for two cases are shown in two columns. Maximizing discriminability yields separated samples which have Bayes optimal classification error.

for the threshold. In addition to neuroimages, demographic information and five neuro factors (27) 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 (28). 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 (29). For comparison, another reliability statistics, namely image intraclass correlation coefficient (I2C2) 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 3. The threshold which maximizes discriminability is close to the thresholds yielding smallest predicting errors for three covariates.

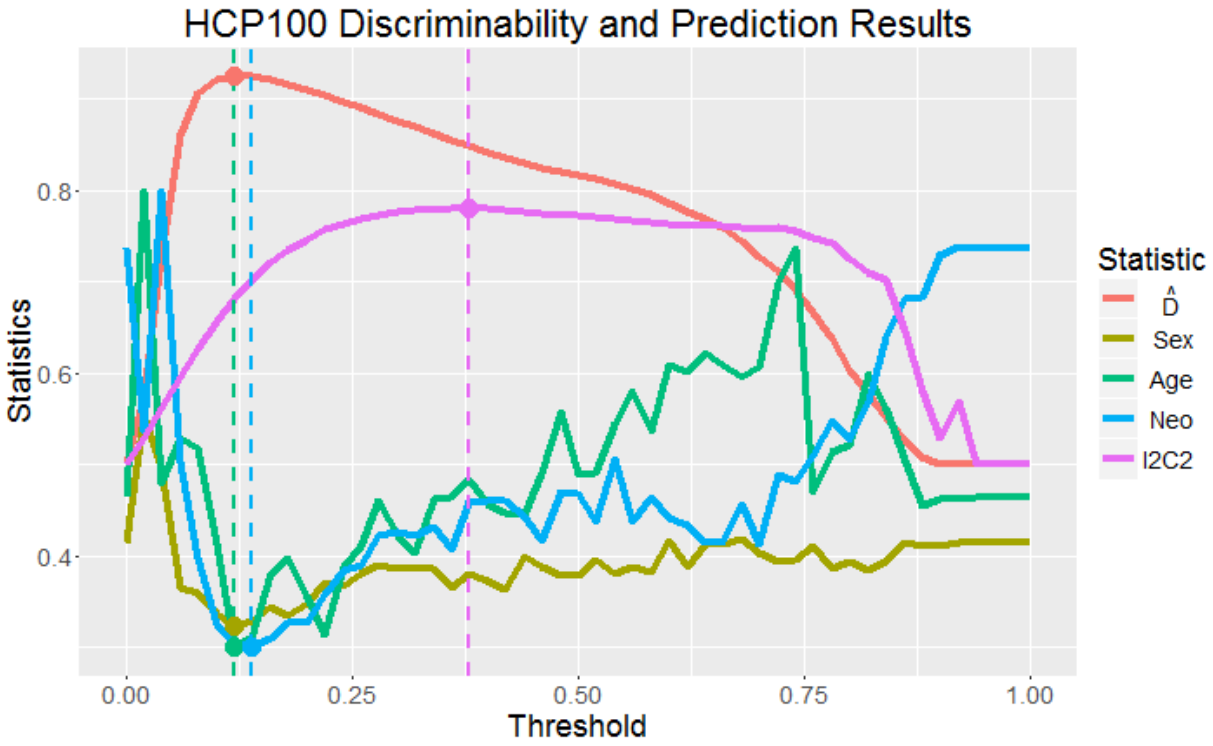

Figure 3: **Optimizing discriminability yields optimal prediction accuracy for multiple covariates.** HCP100 is used to investigate optimal threshold to convert correlation graphs into binary graphs. Curves are scaled to have similar value range. For each statistic, the optimal threshold and value pair 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 (30). 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 (31), anatomical registration (32), temporal filtering (33), motion correction (34) and nuisance signal regression (35). 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 (36, 37)
Anatomical Registration	F for FSL, A for ANTS (38, 39)
Temporal Filtering	F for frequency filtering, X for not (33)
Motion Correction	S for scrubbing, X for not (34)
Nuisance Signal Regression	G for global signal regression , X for not (35)

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) (40). 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 maps (V, E) to (V, E'), that is

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

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 and compare the difference in discriminability with and without rank conversion. It turns out that the rank conversion does help improving mean discriminability in all pipelines. When global signal regression is not performed, rank conversion significantly boosts discriminability. The Figure 4 shows the discriminability of rank fMRI graphs and the discriminability of raw fMRI graphs are provided in 7 which is included in appendix.

There is notable variation in discriminability. The discriminability of 13 data sets processed by 64 pipelines vary from 0.732 to 0.997. The sample-size weighted mean discriminability of 64 pipelines vary from 0.868 to 0.966. CFXSG turns out to be the best pipeline with maximum mean discriminability. In Figure 4, for each data set, we compare CFXSG to all the other pipelines using the two sample test. We combine the p-values by Fisher's method (41), then we group pipelines by the magnitude of their p-values and order them by mean discriminability. Furthermore, we carried out a multi-factor analysis of variance test to study each option (42). Specifically, we fix decisions for all options except one, and investigates 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, fsl and no scrubbing is not statistical significantly better at level 0.05. No frequency filtering, global signal regression and rank conversion is better than their alternatives at level 0.001. Figure 5 shows the distribution of paired difference in discriminability.

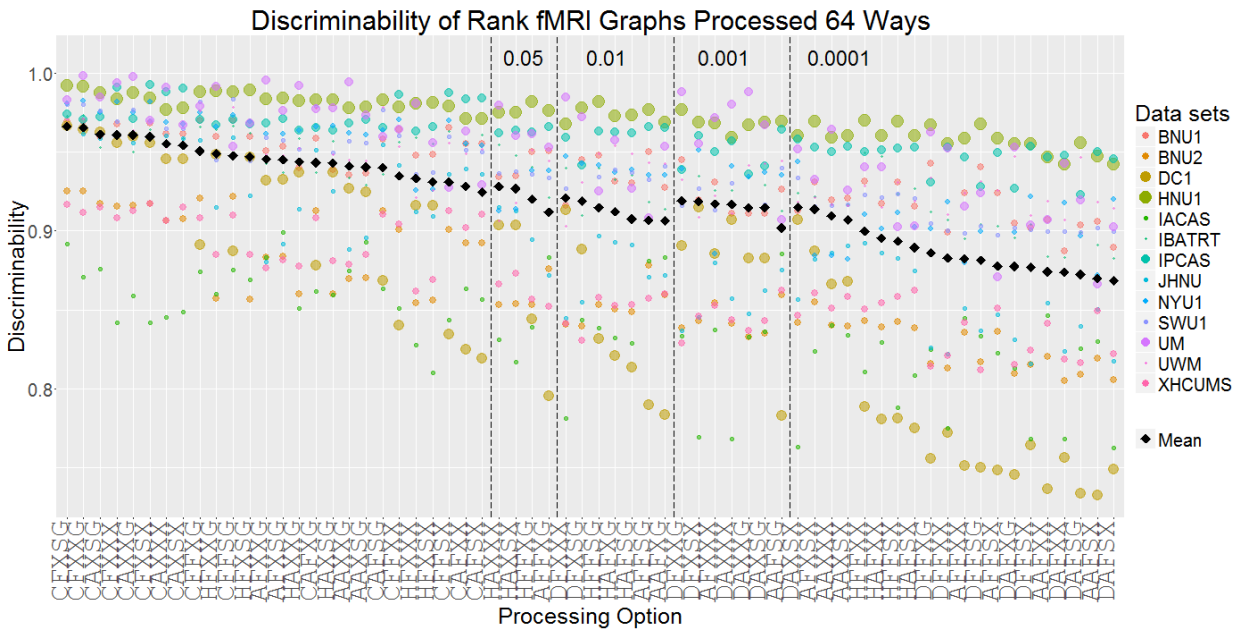

Figure 4: **Discriminability of rank fmri graphs from 13 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1,HNU1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS pre-processed by 64 pipelines are shown in the plot. Color of each dot indicates data set and size indicates the number of measurements in data set. The black square indicates the weighted mean discriminability across 13 data sets. For each data set, all pipelines are compared to the pipeline CFXSG using two sample test, and a single p-value is calculated by Fisher's method. The pipelines are grouped by p-values. The number at the top indicates the range of the p-values. Within each group, the pipelines are ordered by the mean discriminability. CFXSG pipeline has the best mean discriminability across data sets.

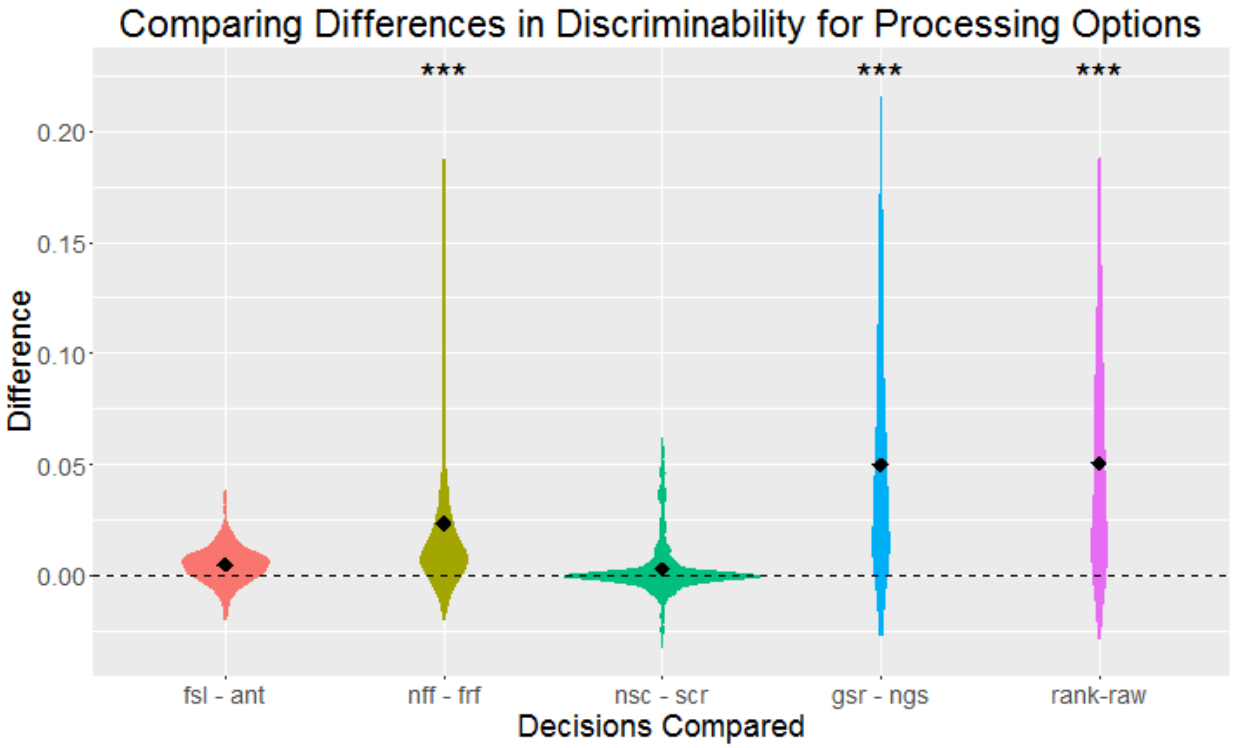

Figure 5: **Paired difference in discriminability of pre-processing options.** Difference in discriminability for each option is compared by fixing the other options and data set. The symbols at top indicates the significance. No frequency filtering, global signal regression and rank conversion are statistical significantly better than their alternatives at level 0.001. Fsl and 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 (43). 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 6. 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.

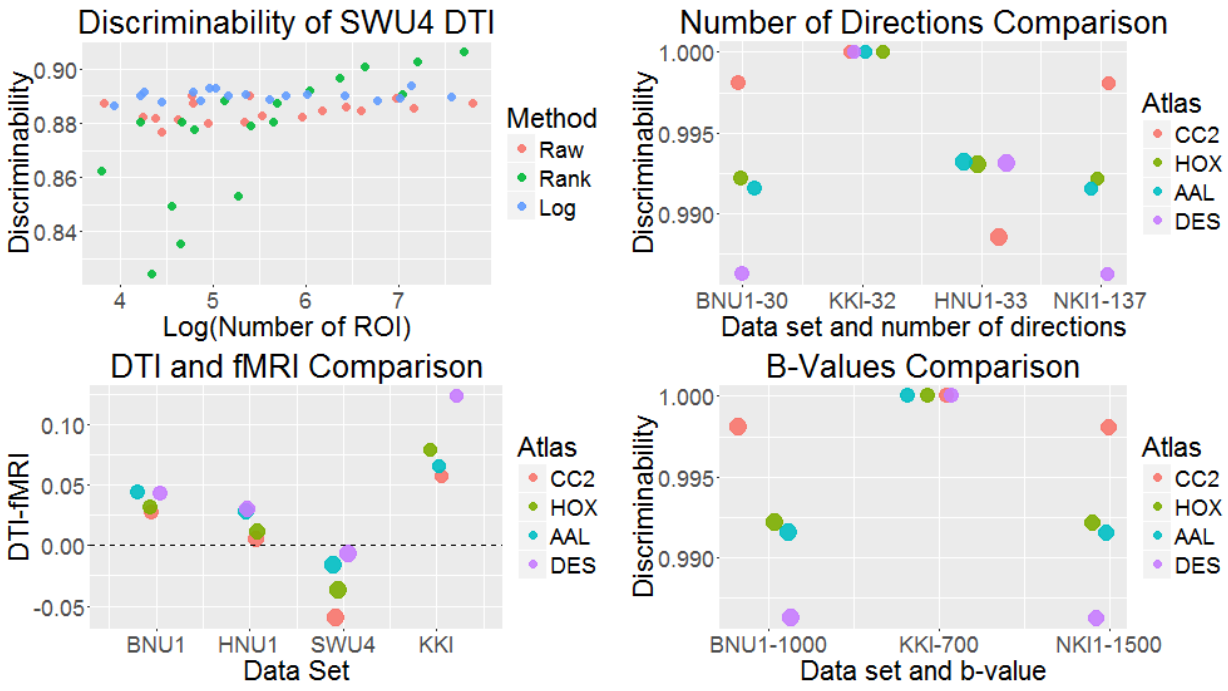

Figure 6: **Discriminability of DTI data sets.** The top left plot shows the discriminability of SWU4 registered with 15 atlases with ROI varying from 48 to 1875. Raw, rank and log edges weights are considered. Discriminability of DTI and fMRI graphs are compared for BNU1, HNU1, SWU4 and KKI data set. The results are shown in the bottom left panel. DTI data sets tend to be more discriminable than fMRI data sets. The plots in the right column show the discriminability of different data sets with different b-values and number of directions.

III.C.4 DTI processing pipelines

In this experiment, we consider the processing of diffusion tensor imaging (DTI) (43). In particular, we are interested in finding the optimal number of ROI, and the optimal approach to process edge weights. SWU4 data set is used in this experiment. We process four DTI data sets using 15 atlases with the number of ROI ranging from 48 to 1875 (44). 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 6 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 and high when ROI is large. 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) and raw edge weights are used. In processing DTI data sets, the raw edge weights are also used. The detailed DTI processing configurations and parameters are provided in the appendix. The result is shown in the bottom left panel of figure 6. Our conclusion is that DTI data sets have at least comparable discriminability as fMRI data sets. Actually, DTI measurements are better than fMRI measurements 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 estimate the discriminability based on test-retest data set. We show the estimator is unbiased and converges to the discriminability asymptotically. Furthermore, we developed one sample and two sample tests for discriminability, which can be used to detect subject signal in data set and compare discriminability of two processing pipelines. 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 From the theoretical point of view, most of our theories require the noise to be additive and independent of subjects. The effects of subject specific noise on discriminability are left uninvestigated. As for applications, 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

Proofs

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

$$
\begin{aligned}
& \mathbb{P}(\delta_{i,t,t'} < \delta_{i,i',t,t''}) \\
= \; & \mathbb{P}(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i,t'}\| < \|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t''}\|) \\
= \; & \mathbb{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 (45),

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

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

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

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 (46), which is

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

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

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

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 $\hat{D}$,

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

Notice that the expectation of $\hat{D}_{i,t,t'}$ is actually D,

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

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 $\hat{D}$,

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

In the last sum above, we should 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,

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

There are $(ns(s-1)(n-1)s)^2$ covariance terms in the sum of nominator; however, most of them are actually 0. $\mathbb{I}\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\}$ 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 $\mathbb{I}\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\}$ is independent of $\mathbb{I}\{\delta_{j,r,r'} < \delta_{j,j',r,r''}\}$ 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 $\mathbb{I}\{\delta_{i,t,t'} < \delta_{i,i',t,t''}\}$ and $\mathbb{I}\{\delta_{j,r,r'} < \delta_{j,j',r,r',r''}\}$ 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

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

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

Proof of Lemma 3. Consider the additive noise setting, that is $\boldsymbol{x}_{i,t} = \lambda \boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t}$. We further assume $\boldsymbol{v}_i$ and $\boldsymbol{\epsilon}_{i,t}$ have continuous distributions, and $\boldsymbol{v}_i$ has spherical distribution. We will show that $D = 0.5$ implies $\lambda = 0$, hence $\boldsymbol{x}_{i,t} = \boldsymbol{\epsilon}_{i,t}$. This implies $\boldsymbol{x}_{i,t}$ is independent of physical property $\boldsymbol{v}_i$ and hence, any phenotype $\boldsymbol{y}_i$.

First, we rewrite the definition of discriminability.

$$\begin{aligned} & \mathbb{P}(\delta_{i,t,t'} < \delta_{i,i',t,t''}) \\ = \; & \mathbb{P}(\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i,t'}\| < \|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{i',t''}\|) \\ = \; & \mathbb{P}(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| < \|\lambda \boldsymbol{v}_i + \boldsymbol{\epsilon}_{i,t} - \lambda \boldsymbol{v}_{i'} - \boldsymbol{\epsilon}_{i',t''}\|) \\ = \; & \mathbb{E}(\mathbb{P}(\|\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}\| < \|\lambda \boldsymbol{v}_i - \lambda \boldsymbol{v}_{i'} + \boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}\|) | \|\lambda \boldsymbol{v}_i - \lambda \boldsymbol{v}_{i'}\| = v). \end{aligned}$$

Let $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ and $\boldsymbol{V}$ denote the $\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t'}$, $\boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i',t''}$ and $\lambda \boldsymbol{v}_i - \lambda \boldsymbol{v}_{i'}$ respectively. Due to the assumption that $\boldsymbol{v}_i$ is spherically distributed,

$$\mathbb{P}(\|\boldsymbol{A}_1\| < \|\boldsymbol{A}_2 + \boldsymbol{V}\| | \|\boldsymbol{V}\| = v) = \int_{\mathbb{S}^{d-1}} \mathbb{P}(\|\boldsymbol{A}_1\| < \|\boldsymbol{A}_2 + t\|) dS / Area(\mathbb{S}^{d-1}),$$

where $\mathbb{S}^{d-1}$ is the ball in $\mathbb{R}^d$ with radius v. We are going to show the expression above is greater than 0.5 as long as $v > 0$. Therefore, $D = 0.5$ implies $\lambda \boldsymbol{v}_i - \lambda \boldsymbol{v}_{i'} = 0$. Since $\boldsymbol{v}_i$ is not constant, we have $\lambda = 0$. Due to symmetry in $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, we have

$$\begin{aligned} 2 \int_{\mathbb{S}^{d-1}} \mathbb{P}(\|\boldsymbol{A}_1\| < \|\boldsymbol{A}_2 + t\|) dS &= \int_{\mathbb{S}^{d-1}} \mathbb{P}(\|\boldsymbol{A}_1\| < \|\boldsymbol{A}_2 + t\|) dS + \int_{\mathbb{S}^{d-1}} \mathbb{P}(\|\boldsymbol{A}_2\| < \|\boldsymbol{A}_1 + t\|) dS \\ &= \int_{\mathbb{S}^{d-1}} \mathbb{P}(\|\boldsymbol{A}_1\| < \|\boldsymbol{A}_2 + t\|) + \mathbb{P}(\|\boldsymbol{A}_2\| < \|\boldsymbol{A}_1 + t\|) dS \\ &= \int_{\mathbb{S}^{d-1}} \int \mathbb{I}(\|a_1\| < \|a_2 + t\|) + \mathbb{I}(\|a_2\| < \|a_1 + t\|) d\mathbb{P}(a_1, a_2) dS \\ &= \iint_{\mathbb{S}^{d-1}} \mathbb{I}(\|a_1\| < \|a_2 + t\|) + \mathbb{I}(\|a_2\| < \|a_1 + t\|) dS d\mathbb{P}(a_1, a_2). \end{aligned}$$

Let us consider the inner integral $\int_{\mathbb{S}^{d-1}} \mathbb{I}(\|a_1\| < \|a_2 + t\|) + \mathbb{I}(\|a_2\| < \|a_1 + t\|) dS$ and denote its value by V. Next, we show V is greater than or equal to $Area(\mathbb{S}^{d-1})$ for any a_1 and a_2. First, let us consider the case that t lies on the circle which is contained in the plane spanned by a_1 and a_2, there are three cases.

Case (1): If $\|t\| \leq |\|a_1\| - \|a_2\||$, then one of the two indicators holds for all t; hence, $V = Area(\mathbb{S}^1)$.

Case (2): If $|\|a_1\| - \|a_2\|| < \|t\| \leq |\|a_1\| + \|a_2\||$, then due to symmetry $V = Area(\mathbb{S}^1)$.

Case (3): If $\|t\| > |\|a_1\| + \|a_2\||$, then both of the two indicators holds for all t; hence, $V = 2Area(\mathbb{S}^1)$.

If t does not lie in the plane spanned by a_1 and a_2, it only adds positive number to the right hand side of the inequalities: $\|a_1\| < \|a_2 + t\|$ and $\|a_2\| < \|a_1 + t\|$. Therefore, the three cases become

Case (1): If $\|t\| \leq |\|a_1\| - \|a_2\||$, then $V = Area(\mathbb{S}^{d-1})$.

Case (2): If $|\|a_1\| - \|a_2\|| < \|t\| \leq |\|a_1\| + \|a_2\||$, then $V \geq Area(\mathbb{S}^{d-1})$.

Case (3): If $\|t\| > |\|a_1\| + \|a_2\||$, then $V = 2Area(\mathbb{S}^{d-1})$.

This shows V always greater than or equal to $Area(\mathbb{S}^{d-1})$. Since $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ have positive mass at any open ball centered at origin, case (3) must happen with positive probability. As a consequence,

$$\iint_{\mathbb{S}^{d-1}} \mathbb{I}(\|a_1\| < \|a_2 + t\|) + \mathbb{I}(\|a_2\| < \|a_1 + t\|) | dS d\mathbb{P}(a_1, a_2) > Area(\mathbb{S}^{d-1})$$

This shows $\mathbb{P}(\|\boldsymbol{A}_1\| < \|\boldsymbol{A}_2 + \boldsymbol{V}\| | \|\boldsymbol{V}\| = v) > 0.5$ as long as $v \neq 0$. As discussed above, this concludes the proof of the lemma. □

Discriminability of Raw fMRI Graphs

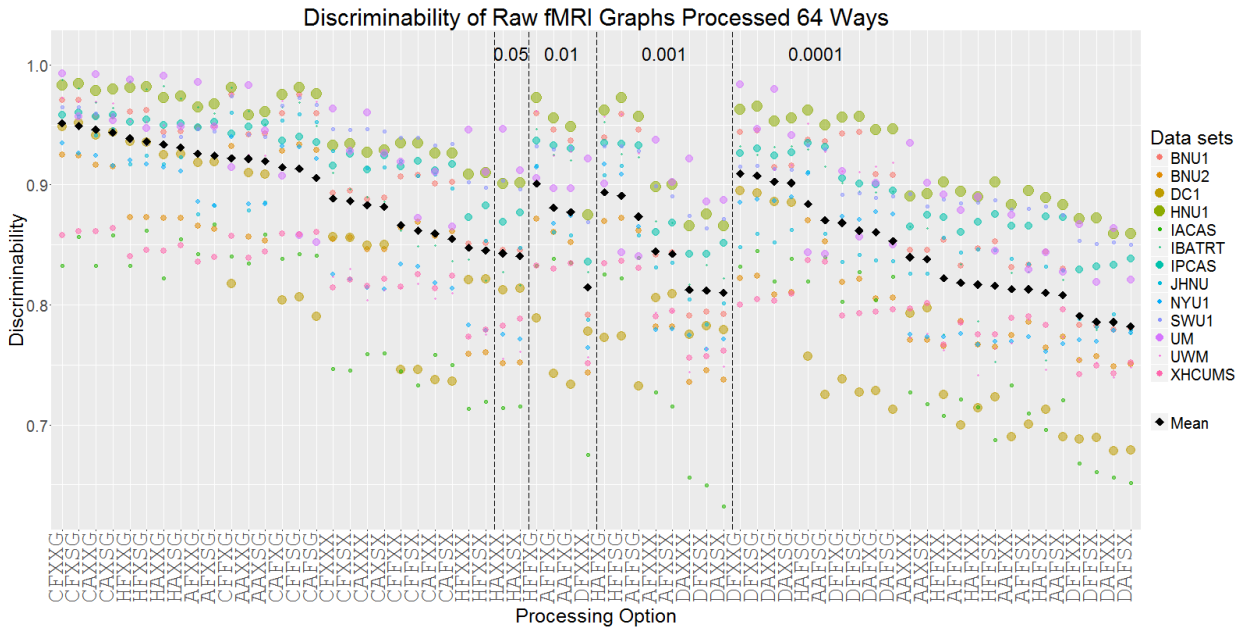

Figure 7: **Discriminability of raw fmri graphs from 13 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1, HNU1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS pre-processed by 64 pipelines are computed and shown in the figure. Color of each dot indicates data set and size indicates the number of measurements in data set. The black square indicates the weighted mean discriminability across 13 data sets. For each data set, all pipelines are compared to the pipeline CFXSG using two sample test, and a single p-value is calculated by Fisher's method. The pipelines are grouped by p-values. The number at the top indicates the range of the p-values. Within each group, the pipelines are ordered by the mean discriminability. CFXXG pipeline has the best mean discriminability across data sets.

Discriminability Data of Rank fMRI Graphs

	BNU1	BNU2	DC1	HNU1	IACAS	IBATRT	IPCAS	JHNU	NYU1	SWU1	UM	UWM	XHCUMS	Mean
AAFXG	0.927449	0.859286	0.783674	0.968958	0.883002	0.913690	0.965035	0.872126	0.935463	0.940903	0.953237	0.942917	0.860072	0.906543
CAFXG	0.958776	0.912551	0.878391	0.982635	0.861788	0.936756	0.965118	0.951149	0.965926	0.978910	0.976981	0.977500	0.908134	0.943180
DAFXG	0.904592	0.809388	0.745649	0.955347	0.813185	0.895833	0.926615	0.831034	0.899259	0.901645	0.952915	0.947083	0.815091	0.877574
HAFXG	0.931633	0.850408	0.821087	0.972802	0.832156	0.932292	0.962496	0.892816	0.937037	0.938578	0.957631	0.954167	0.852645	0.912378
AAFXX	0.906735	0.820204	0.736838	0.946837	0.846388	0.883929	0.946720	0.854310	0.873241	0.899595	0.906958	0.918750	0.841630	0.874231
CAFXX	0.955306	0.892143	0.824902	0.971091	0.863632	0.954613	0.983267	0.939080	0.954907	0.951387	0.963114	0.949167	0.901775	0.928410
DAFXX	0.887449	0.805306	0.756704	0.941923	0.768578	0.883185	0.947552	0.823563	0.917685	0.899536	0.942756	0.907083	0.818967	0.873794
HAFXX	0.917143	0.839184	0.780752	0.960229	0.829744	0.947173	0.951257	0.885920	0.901111	0.921220	0.940256	0.915000	0.854547	0.895432
AAFSG	0.926122	0.859286	0.783174	0.969225	0.885593	0.912946	0.964327	0.875575	0.935185	0.939825	0.907079	0.944167	0.862428	0.901947
CAFSG	0.958061	0.912551	0.868578	0.982884	0.863346	0.936012	0.965035	0.954598	0.963519	0.976601	0.959889	0.978333	0.910417	0.940007
DAFSG	0.903469	0.808878	0.733706	0.955676	0.825443	0.898810	0.922952	0.839655	0.896759	0.898874	0.919536	0.946667	0.816286	0.872551
HAFSG	0.930306	0.848776	0.813679	0.973060	0.828661	0.935268	0.962163	0.891379	0.938796	0.936730	0.926832	0.949167	0.853243	0.907630
AAFSX	0.906020	0.819490	0.732669	0.947456	0.829847	0.891369	0.950175	0.850575	0.872037	0.902058	0.866202	0.918333	0.849167	0.869789
CAFSX	0.955408	0.892551	0.819409	0.971189	0.856785	0.960565	0.983808	0.935632	0.954167	0.949848	0.929090	0.951667	0.910236	0.924556
DAFSX	0.889388	0.805918	0.749095	0.942073	0.762786	0.882440	0.945346	0.817529	0.919907	0.899693	0.902241	0.914167	0.822174	0.868462
HAFSX	0.915612	0.838469	0.774974	0.960006	0.808320	0.927083	0.952756	0.875000	0.902593	0.920602	0.903249	0.919167	0.862482	0.889531
AAXXG	0.935306	0.869592	0.926577	0.977625	0.874870	0.933780	0.968115	0.888506	0.962130	0.956745	0.994195	0.945000	0.878822	0.941187
CAXXG	0.959490	0.915918	0.959734	0.987269	0.858750	0.950149	0.970779	0.955747	0.975093	0.976601	0.997299	0.963333	0.912428	0.960729
DAXXG	0.911224	0.832857	0.882894	0.967266	0.832765	0.933780	0.936022	0.847414	0.934167	0.914255	0.988148	0.941667	0.836902	0.914707
HAXXG	0.939898	0.860306	0.937136	0.982195	0.851007	0.958333	0.963786	0.913793	0.963796	0.951663	0.991897	0.952083	0.877808	0.943688
AAXXX	0.919184	0.840000	0.866043	0.958981	0.840203	0.916667	0.950175	0.886207	0.884167	0.915994	0.963920	0.916667	0.851014	0.909469
CAXXX	0.961429	0.906224	0.945712	0.976868	0.845372	0.956845	0.988137	0.960057	0.968889	0.964761	0.990970	0.952917	0.906612	0.955380
DAXXX	0.910714	0.841327	0.906894	0.959308	0.768497	0.925595	0.956960	0.857471	0.940926	0.930265	0.980045	0.910833	0.843351	0.916880
HAXXX	0.934184	0.853265	0.903516	0.974729	0.831239	0.954613	0.961705	0.914943	0.912870	0.936609	0.979400	0.906250	0.866431	0.928232
AAXSG	0.936327	0.870408	0.924621	0.978401	0.892683	0.929315	0.970529	0.895690	0.960833	0.955668	0.973111	0.944167	0.884819	0.940364
CAXSG	0.959898	0.916939	0.962326	0.987526	0.876065	0.953125	0.972361	0.958621	0.974722	0.975985	0.984762	0.962083	0.914891	0.961132
DAXSG	0.910714	0.834898	0.882727	0.968451	0.836466	0.929315	0.941059	0.862069	0.937222	0.916718	0.967185	0.942500	0.842826	0.914656
HAXSG	0.938469	0.859796	0.936799	0.982765	0.859188	0.956845	0.963745	0.925287	0.966481	0.949200	0.977747	0.948333	0.880996	0.943027
AAXSX	0.920408	0.839388	0.868104	0.960410	0.834040	0.916667	0.953588	0.890517	0.881944	0.916610	0.925905	0.915833	0.858514	0.906926
CAXSX	0.961633	0.907653	0.945716	0.977725	0.848962	0.958333	0.990301	0.958333	0.967500	0.964756	0.967105	0.956250	0.915109	0.954224
DAXSX	0.912449	0.842143	0.906845	0.960479	0.763131	0.924107	0.958167	0.856897	0.943889	0.933805	0.951826	0.918333	0.846721	0.914974
HAXSX	0.934796	0.854082	0.903895	0.975124	0.816855	0.947917	0.963828	0.912644	0.913704	0.937686	0.960050	0.917917	0.873152	0.926798
AFFXG	0.950816	0.876122	0.795589	0.975918	0.883351	0.921131	0.965743	0.871839	0.938056	0.933357	0.952611	0.944167	0.852156	0.911937
CFFXG	0.969898	0.920612	0.891382	0.988116	0.874132	0.965774	0.970613	0.957184	0.974907	0.981681	0.978916	0.975833	0.908424	0.950509
DFFXG	0.942551	0.815612	0.756078	0.966950	0.825183	0.912202	0.930528	0.825862	0.901204	0.904711	0.962460	0.933750	0.814366	0.886296
HFFXG	0.950306	0.852959	0.844117	0.981547	0.839326	0.940476	0.962746	0.894540	0.937963	0.935653	0.961155	0.951250	0.856667	0.920298
AFFXX	0.923878	0.835510	0.751169	0.958355	0.844876	0.895089	0.946470	0.850862	0.879630	0.898507	0.915506	0.923750	0.842065	0.882388
CFFXX	0.964388	0.901020	0.840061	0.978351	0.851011	0.956101	0.986347	0.941379	0.970278	0.960013	0.964438	0.953333	0.904203	0.934933
DFFXX	0.898980	0.812959	0.772361	0.954986	0.775143	0.906994	0.953130	0.821839	0.918056	0.900009	0.952413	0.910417	0.821250	0.882955
HFFXX	0.931429	0.843265	0.788678	0.969664	0.810602	0.949405	0.950008	0.892241	0.910463	0.922743	0.940309	0.928333	0.850272	0.899819
AFFSG	0.950306	0.878061	0.789844	0.976540	0.881127	0.918155	0.965826	0.870977	0.935093	0.932280	0.908228	0.943333	0.856938	0.906584
CFFSG	0.970000	0.921531	0.887307	0.988175	0.875197	0.966518	0.970280	0.958908	0.973333	0.983220	0.953521	0.978333	0.909656	0.947562
DFFSG	0.940306	0.817041	0.750175	0.967727	0.833682	0.904018	0.927905	0.837069	0.899815	0.903171	0.923813	0.931250	0.811938	0.881276
HFFSG	0.947959	0.853367	0.831813	0.981808	0.838350	0.939732	0.963079	0.896552	0.941389	0.934267	0.924960	0.949167	0.857989	0.914707
AFFSX	0.923878	0.836327	0.748567	0.958796	0.822476	0.903274	0.949467	0.846839	0.878981	0.897427	0.871044	0.925833	0.851014	0.877746
CFFSX	0.965408	0.900612	0.834690	0.978636	0.843672	0.962798	0.987304	0.935920	0.969907	0.959705	0.927571	0.955833	0.912862	0.930896
DFFSX	0.901939	0.815204	0.764450	0.955244	0.768609	0.904018	0.953546	0.816092	0.921667	0.900619	0.903679	0.909583	0.824293	0.876907
HFFSX	0.930510	0.842245	0.781240	0.969452	0.788431	0.923363	0.952423	0.882759	0.911759	0.923356	0.902650	0.930833	0.858442	0.893537
AFXXG	0.950408	0.883163	0.931645	0.983162	0.883020	0.942708	0.968115	0.880460	0.961389	0.959370	0.995491	0.948333	0.876667	0.945379
CFXXG	0.969082	0.925000	0.964223	0.991333	0.870641	0.973214	0.970529	0.961494	0.982037	0.979834	0.997983	0.970833	0.911558	0.965581
DFXXG	0.945000	0.838367	0.890495	0.976575	0.833195	0.918899	0.938728	0.837069	0.939352	0.916256	0.987935	0.932500	0.829130	0.919107
HFXXG	0.959490	0.856939	0.948355	0.988527	0.860303	0.964286	0.966866	0.914655	0.966111	0.945197	0.991337	0.946667	0.884692	0.948842
AFXXX	0.930612	0.854286	0.885344	0.968144	0.837410	0.925595	0.950175	0.879885	0.885648	0.913356	0.971717	0.922500	0.852572	0.917045
CFXXX	0.967857	0.915510	0.955905	0.983235	0.842025	0.960565	0.991009	0.959483	0.981481	0.969533	0.993394	0.958333	0.907971	0.960874
DFXXX	0.916020	0.840510	0.913406	0.967674	0.781087	0.939732	0.959041	0.844828	0.947130	0.927032	0.984612	0.902917	0.841359	0.920750
HFXXX	0.947857	0.854592	0.915837	0.980805	0.827694	0.959077	0.963079	0.912356	0.925556	0.935682	0.980578	0.921250	0.861830	0.933248
AFXSG	0.953163	0.883571	0.932295	0.983765	0.899047	0.936756	0.970779	0.891667	0.961296	0.957215	0.976068	0.947500	0.881504	0.945189
CFXSG	0.969082	0.925000	0.966203	0.991739	0.891532	0.972470	0.973693	0.966667	0.980463	0.979834	0.982991	0.970417	0.916612	0.966030
DFXSG	0.945204	0.839694	0.888521	0.977577	0.843739	0.909970	0.941517	0.855172	0.942685	0.918873	0.971994	0.930833	0.830725	0.918876
HFXSG	0.959592	0.856837	0.946543	0.988891	0.869007	0.962054	0.967741	0.922414	0.969259	0.944119	0.968394	0.945417	0.884764	0.946977
AFXSX	0.930918	0.855102	0.887109	0.969110	0.823921	0.921875	0.952756	0.881897	0.884167	0.915662	0.932318	0.921250	0.860743	0.913815
CFXSX	0.967959	0.917347	0.955543	0.983783	0.841837	0.964286	0.992383	0.960920	0.981574	0.969992	0.969818	0.960417	0.917355	0.959739
DFXSX	0.919490	0.842959	0.915034	0.968571	0.769406	0.934524	0.960206	0.845402	0.951204	0.926878	0.955142	0.908750	0.845616	0.918751
HFXSX	0.948265	0.855816	0.916207	0.981056	0.810097	0.955357	0.965618	0.909195	0.926296	0.936914	0.955894	0.925833	0.869275	0.931053

Table 1: Discriminability of 13 data sets and their weighted means processed by 64 pipelines.

Registration				
Step	Package	Function	Parameters	Description
DTI De-noising	FSL	eddy-correct	None specified (defaults)	
MPRAGE Skull-stripping	FSL	bet	‘-B ‘	reduction of image bias and neck voxels
DTI align to MPRAGE	FSL	epi-reg	None specified (defaults)	
MPRAGE align to MNI 1mm	FSL	flirt	‘-cost mutualinfo‘; ‘-bins 256‘; ‘-dof 12‘; ‘-searchrx -180 180‘; ‘-searchry -180 180‘; ‘-searchrz -180 180‘	evaluated with mutual information, using 256 bin histograms of intensity, a 12 degree of freedom model, and searching the entire physical space for possible alignments
Apply MPRAGE transform to DTI	FSL	flirt	‘-interp trilinear‘; ‘-applyxfm	applying the transforms computed with trilinear interpolation
Resample aligned DTI	nilearn	resample-img	‘interpolation =”nearest”‘	resampling with nearest-neighbour interpolation
Diffusion processing				
Step	Package	Function	Parameters	Description
Tensor Fitting	dipy	TensorModel	None specified (defaults)	
Tractography	dipy	EuDX	‘a-low=0.1‘	FA stopping threshold for fiber tracking

Configurations of DTI Processing Pipelines

Data set	Scanner	Num. of channel	Structrual Sequence	Functional Sequence	Flip Angle of fMRI	Echo Time (TE in ms)	Repetition Time (TR in ms)	Dimensions (mm x mm x mm)
BNU1	Siemans TrioTim	12 Channel	3D MPRAGE	EPI	90	30	2000	3.1 x 3.1 x 3.5
BNU2 first scan	Siemans TrioTim	12 Channel	3D MPRAGE	EPI	90	30	2000	3.1 x 3.1 x 3.0
BNU2 retest	Siemans TrioTim	12 Channel	3D MPRAGE	EPI	90	30	1500	3.1 x 3.1 x 4
DC1	Philips	32 Channel	3D T1-TFE	EPI	90	35	2500	3 x 3 x 3.5
HNU1	GE Discovery MR750	8 Channel	3D SPGR	EPI	90	30	2000	3.4 x 3.4 x 3.4
JHNU	Siemans TrioTim	8 Channel	3D MPRAGE	EPI	90	30	2000	3.75 x 3.75 x 4
IACAS	GE Sigma HDx	8 Channel	3D BRAVO	EPI	90	30	2000	3.4 x 3.4 x 4
IBATRT	Siemans TrioTim	12 Channel	3D MPRAGE	EPI	90	30	1750	3.4 x 3.4 x 3.6
NYU1	Siemans Allegro				90	15	2000	3 x 3 x 4
SWU1								
UM								
UWM								
IPCAS								
XHCUMS	Siemans Allegro				90	15	2000	3 x 3 x 3

fMRI data sets and scanner parameters

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. 3, 11

[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. 3

[18] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel, "Measuring reproducibility of high-throughput experiments," *The annals of applied statistics*, pp. 1752–1779, 2011. 2, 3

[19] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31. 5

[20] P. J. Bickel and K. A. Doksum, *Mathematical statistics: basic ideas and selected topics*. CRC Press, 2015, vol. 2. 7

[21] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002. 10

[22] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26. 10

[23] S. C. Strother, "Evaluating fmri preprocessing pipelines," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 2, pp. 27–41, 2006. 10

[24] 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. 10

[25] 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. 10

[26] 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. 10

[27] 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. 11

[28] 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. 11

[29] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1. 11

[30] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional magnetic resonance imaging*. Sinauer Associates Sunderland, 2004, vol. 1. 12

[31] J. K. Mai, M. Majtanik, and G. Paxinos, *Atlas of the human brain*. Academic Press, 2015. 12

[32] 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. 12

[33] 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. 12

[34] 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. 12

[35] 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. 12

[36] 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. 12

[37] 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. 12

[38] 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. 12

[39] B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ants)," *Insight J*, vol. 2, pp. 1–35, 2009. 12

[40] 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. 12

[41] R. A. Fisher, *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925. 13

[42] J. F. Hair, "Multivariate data analysis," 2009. 13

[43] 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. 15, 16

[44] S. Mori, S. Wakana, P. C. Van Zijl, and L. Nagae-Poetscher, *MRI atlas of human white matter*. Elsevier, 2005. 16

[45] 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. 18

[46] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice hall, 1982. 18