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

April 3, 2016

Contents

I Introduction

Opportunity and Challenge In this era of big data, many scientific, government, and corporate groups are collecting and processing massive datasets. 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 datasets can then be used for a wide variety of different inferential problems. Collecting and processing these datasets 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 using pilot data could reap great rewards

Action and Resolution 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. We then design an estimator of discriminability computed from test-restest 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 specifically study functional connectomics. We start by finding the maximally reliable threshold for converting correlation matrices into graphs. Indeed, consistent with our theoretical and simulated results, maximizing the discriminability of our datasets also maximizes performance 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 Definition of Discriminability

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 to directly observed $\boldsymbol{v}_i$, 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 re-

move 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, including 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 $\boldsymbol{x}_i := g_{\boldsymbol{\psi}}(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i))$. **TODO: Emphasize $\boldsymbol{v}_i$ is a random variable or emphasize $g_{\boldsymbol{\psi}}$ is a random function or takes a random variable as input. Otherwise $\boldsymbol{x}_i|\boldsymbol{v}_i$ is not random.**

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 $\boldsymbol{w}_{i,t}$, which is the t^{th} trial, implicitly also a function of $\boldsymbol{\phi}$, which encodes all the details of the measurement. If both f and g together do not introduce too much noise, then we would expect that $\boldsymbol{x}_{i,1}$ and $\boldsymbol{x}_{i,2}$ are *closer* to one another than either are to any other subject's data, $\boldsymbol{x}_{i',t}$. Define δ to be a metric computing the distance between two data points, $\delta \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Formally, we expect that $\delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'}) \leq \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''})$, for any i, i', t, 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:

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

When trying to find the best data collection and processing pipeline, we try to maximize the discriminability of processed data, that is

$$\underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}, \boldsymbol{\phi} \in \boldsymbol{\Phi}}{\text{maximize}} \quad D(\boldsymbol{\psi}, \boldsymbol{\phi}) \tag{2}$$

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 pre-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 $\boldsymbol{\Psi}$

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

This approach is intuitive and easy to understand. We will show in the theory section 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 $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ to make prediction of $\boldsymbol{y}_i$ based on observation $\boldsymbol{x}_i$ easier. 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 well known the minimal prediction error is achieved by Bayes classification.

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

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

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

We should emphasize L^* is determined by distribution of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Moreover, $\boldsymbol{x}_i$ 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 ad 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 performance to be good 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.

II.A.3 Estimator/Test Statistic

In real world, exact 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 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 the same subject i, we first define

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

Here, n is the total number of subjects, and s is the number of observations per subject. $\hat{D}_{i,t,t'}$ approximate the probability that distance between observations from other subjects and t^{th} observation of subject i is larger than distance between t^{th} and t'^{th} trial of subject i. Then, we define $\hat{D}$ to be the mean of $\hat{D}_{i,t,t'}$.

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

$\hat{D}$ serves as the sample approximated discriminability. The next lemma asserts the unbiasedness of $\hat{D}$ as an estimator.

Lemma 1.

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

The following lemma shows the as the number of subjects increase, $\hat{D}$ converges in probability to discriminability.

Lemma 2. *As $n \to \infty$,*

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

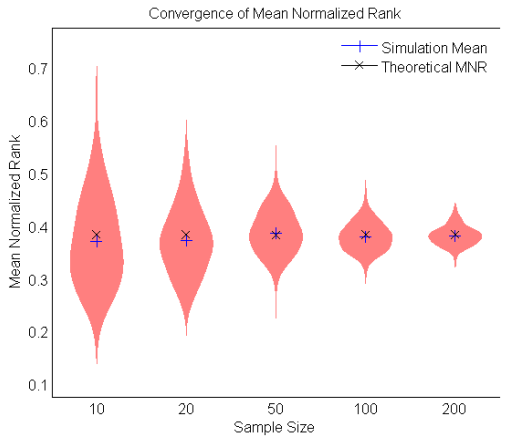

Figure 1: **Convergence of sample Discriminability.** As the number of samples increases, the sample Discriminability converges to true population Discriminability

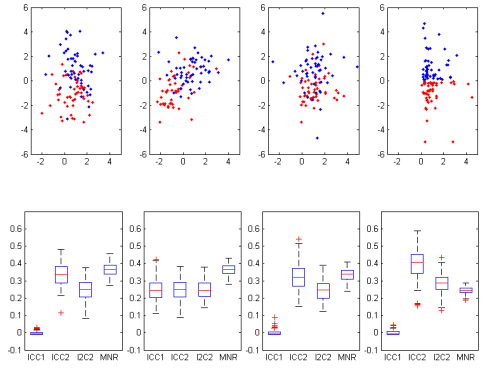

Figure 2: Two class ICC, I2C2 and Discriminability

II.B Simulations

II.B.1 Convergence of $\hat{D}$

In Lemma 1 and 2, we assert sample discriminability $\hat{D}$ is unbiased and converges to the true population discriminability in probability. We demonstrate this idea with simulation. The true physical property are generated from independent Gaussian distribution. That is $v_i \overset{i.i.d.}{\sim} \mathbb{G}(0,1)$. All subjects have two observations with additive noise: $x_{i,t}|v_i \sim \mathbb{G}(v_i,1)$. For each number of subjects, we generate data and compute discriminability 100 times.

The figure 1 shows how $\hat{D}$ is distributed when we vary the sample size from 10 to 200. With this data generation scheme, the population discriminability can be computed from numerical integration, and is marked on the plot. We can see from the figure that sample discriminabiity $\hat{D}$ converges to D as sample size increase.

II.B.2 Dhat provides a more useful bound than ICC or I2C2 for a variety of simulated settings

II.B.3 we can use Dhat to choose the most discriminable parameter (eg, threshold)

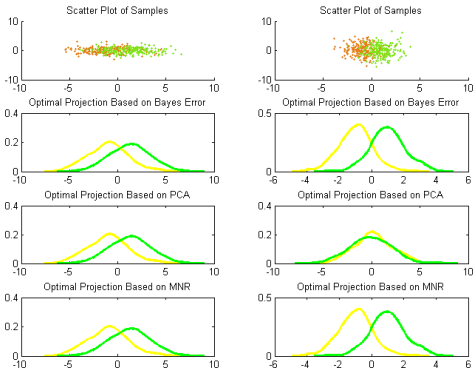

Figure 3: **Best linear projection selected by PCA and Discriminability.** Best linear projections are computed based on PCA and optimizing Discriminability. Under two settings, maximizing Discriminability always yields Bayes optimal projection.

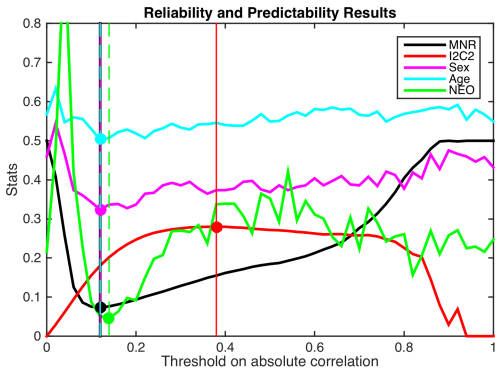

Figure 4: **Optimizing Discriminability yields optimal prediction accuracy for gender and age.** HCP100 is used to investigate optimal threshold to convert correlation graphs into binary graphs. The threshold is varied from 0 to 1. At each threshold, the discriminability is computed; sex, age and a neuro factor are predicted using k-NN. The threshold maximizing discriminability yields optimal prediction performance for all three tasks.

II.C Connectome Applications

II.C.1 optimal Discriminability yields optimal predictive accuracy

- Introduce how fmri graph is constructed
- Describe how we threshold correlation and what is being predicted, how we predict
- explain the result in fig 4 that maximal discriminability yields best prediction performance

II.C.2 Best Pipeline of 64 (raw correlation graph)

- Explain the problem of finding the best pipelines
- Explain what four decisions are and 4 atlases, nff vs frf, fsl vs ant, scrub vs no scr, gsr vs no gsr.
- Claim we find the best pipeline with maximum mean (without a hypothesis tesing)

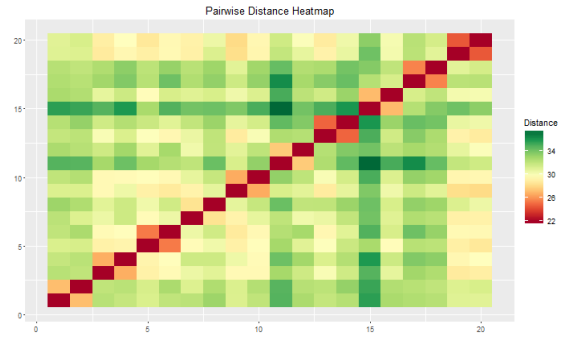

Figure 5: **Pairwise distance of an discriminable data set.** Pairwise distance of 20 observations from 10 subjects are displayed. The observations are from BNU1 data set processed by CFXXG pipeline. The small 2-by-2 block on the diagonal indicates that within subject distance is smaller than across subject distance.

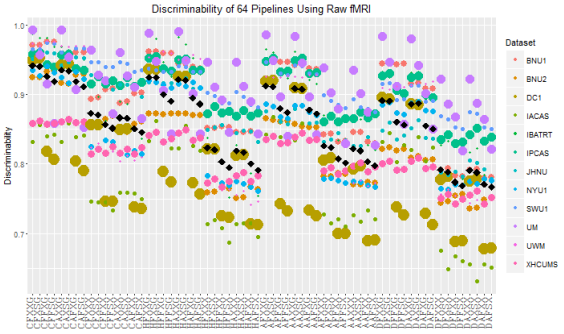

Figure 6: **Discriminability of raw fmri graphs from 12 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS processed by 64 pipelines are computed. CFXXG pipeline has the best mean Discriminability across data set.

II.C.3 best pipeline = product of marginals

- Explain we focus on individual decision for statistical significance
- Explain how we perform the multifactor analysis of variance test
- we find nff better than frf, fsl better than ant, scrub ? no scr, gsr better than no gsr, and atlas is being significant, best pipeline = product of marginals

II.C.4 which atlas/resolution

- Background info about 4 atlases
- claim cc200 better than hox better than aal better than des

II.C.5 rank graphs

- Explain how rank graphs are constructed, the motivation to robust estimate graphs
- Explain it offers large improvement over raw fmri graphs when there is no gsr, small when there is gsr.

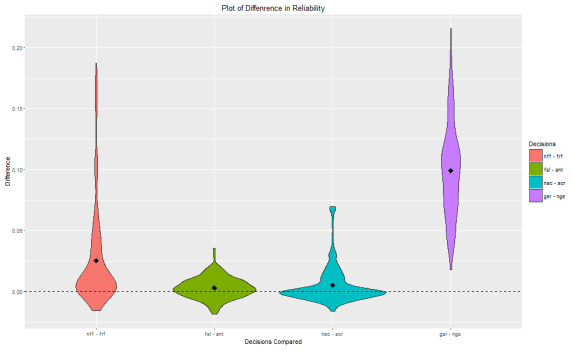

Figure 7: **Paired difference in Discriminability of each decisions.** Difference in Discriminability for each decision is compared by fixing other decisions and a data set. nff, fsl and gsr is statistical significantly better.

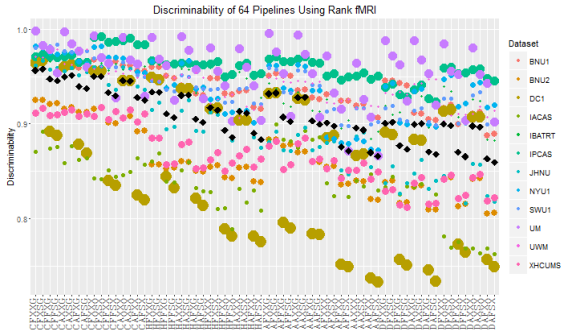

Figure 8: **Discriminability of rank fmri graphs from 12 data sets processed 64.** Discriminability of same data sets are computed with correlation graphs converting into rank graphs. CFXSG is the best pipeline in terms of mean Discriminability.

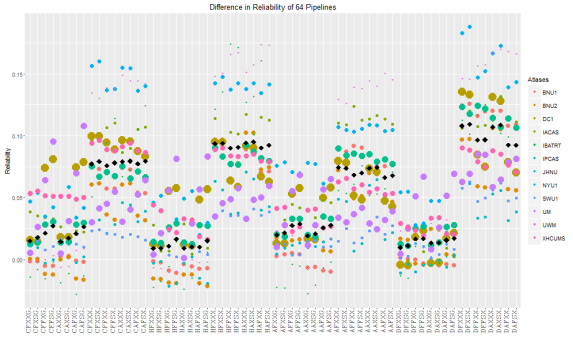

Figure 9: **Difference in Discriminability between rank vs raw fmri graphs from 12 data sets processed 64.** Difference in discriminability is computed by fixing a pipeline and a data set. Overall, rank graphs is more discriminable than raw correlation graphs, especially when no global signal regression is performed.

II.C.6 DTI vs. fMRI

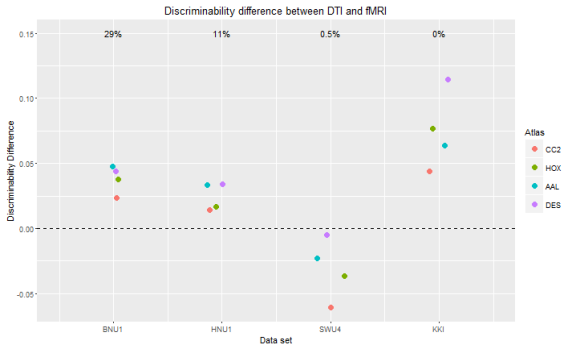

Figure 10: **Difference in Discriminability between dti and fmri of 4 data sets.** Discriminability of dti and fmri are computed for data sets: BNU1, HNU1, SWU4 and KKI. Difference in discriminability is then computed by subtracting fmri discriminability from dti discriminability. The number at top indicates the percentage of outliers removed from dti data sets. Overall, dti graphs is more discriminable than fmri graphs.

III Discussion

Summary

Related Work Our discriminability statistics is not the first attempt to quantify information of test-retest data set. Previous works include but not limit to: analysis of variance (ANOVA), intraclass correlation coefficient (ICC), image intraclass correlation coefficient (I2C2) and distance components (DISCO). The ANOVA and ICC approaches only operate on scalar data set. I2C2 is a generalization of ICC into high dimension space, however it is designed for additive Gaussian noise and assumes Euclidean distance. DISCO is a statistics designed to test whether multi-modal distribution are the same or not. It is less robust and intuitive compared to $\hat{D}$. Also, the mean of DISCO varies with the number of subject and dimension of observation which makes it not suitable to compare across data set.

Next Steps

A Bibliography