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

May 18, 2016

Contents

I Introduction

Opportunity Benchmark datasets are omnipresent

Challenge Deciding on how to collect and process the data, in the absense of an explicit task or for multiple tasks.

Action We define discriminability as a probability which bounds Bayes error.

Resolution We compute discriminability on a simulated data set and real data sets to study the optimal way to process data.

II Results

II.A Theory

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

Rigorously define discriminability Based on intuition, we define discriminability of a multi-subject distribution F as a probability of within subject distances to be smaller than across subject distances.

Define optimizing pipeline problem we are looking for most discriminable processing pipeline that is $\max_\phi \phi(F)$.

II.A.2 Optimizing discriminability Optimizes Bound on Performance for Any Task

Introduce classification Define the binary classification problem and Bayes error.

Justify discriminability Under additive noise setting, discriminability bounds bayes error.

Theorem 1. *discriminability bounds predictive accuracy*

Corollary Processing pipeline selection.

II.A.3 Estimating Discriminability

Estimate discriminability We design an estimator Dhat to estimate discriminability from test-retest data set.

- In a model free setting, $E(Dhat) = D$.
- In a model free setting, $\hat{D}_n \to D$

II.B Simulations

II.B.1 Dhat $\to$ E(D) under gaussian simulation

Simulation The means of subjects follow gaussian and the observations from a subject also follow gaussian.

II.B.2 we can use Dhat to choose the most discriminable parameter

Simulation Learn the optimal linear projection through discriminability

II.C Connectome Applications

II.C.1 optimal Discriminability yields optimal predictive accuracy

real experiment Describe the data and threshold experiment.

real experiment Emphasize that discriminability selects threshold which is close to optimal for multiple tasks.

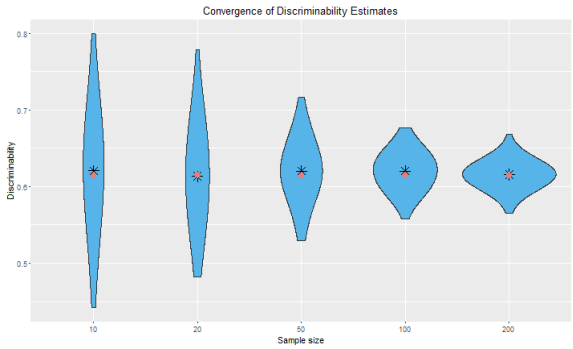

Figure 1: **Convergence of sample discriminability.** Distribution of sample discriminability is estimated. As the number of subjects increases, the sample discriminability converges to the true population discriminability.

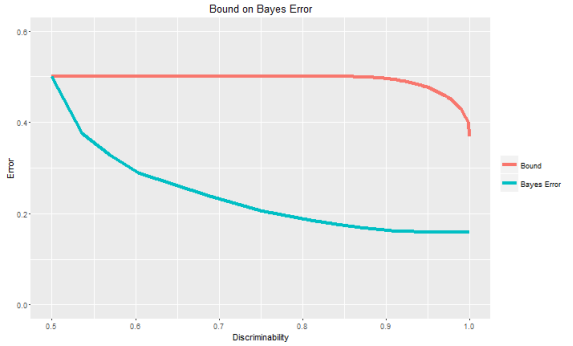

Figure 2: **Prediction error vs discriminability.** Samples are drawn from a mixture of two Gaussian distribution with additive noise. The red line is the theoretical bound derived from the theorem 1. The green line indicates the true Bayes error.

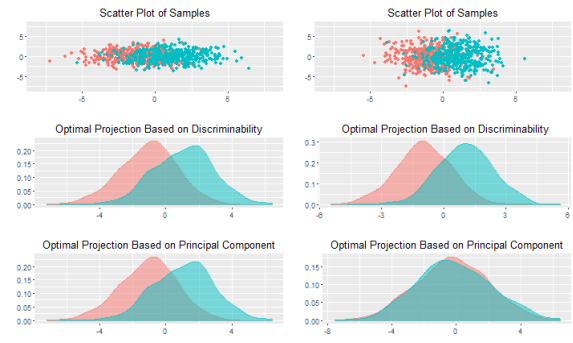

Figure 3: **Linear projection based on by PCA and Discriminability.** Linear projections are computed using PCA and optimizing Discriminability. Maximizing discriminability yield separated samples which have Bayes optimal classification error.

II.C.2 fMRI Processing pipelines

real experiment Describe the 12 data sets and 64 pipelines.

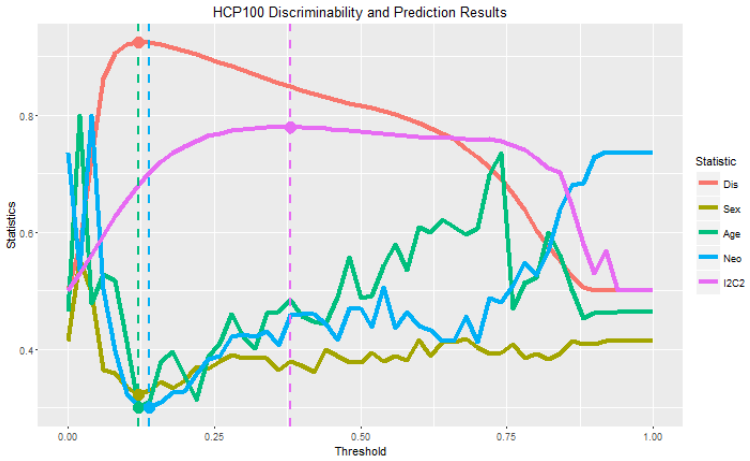

Figure 4: **Optimizing Discriminability yields optimal prediction accuracy for multiple covariates.** HCP100 is used to investigate optimal threshold to convert correlation graphs into binary graphs. The threshold is varied from 0 to 1. For each value of threshold, the discriminability is computed; sex, age and a neuro factor are predicted using k-NN. The threshold maximizing discriminability is close to optimal thresholds for predicting three covariates.

real experiment Decide the best among 64 pipelines.

real experiment Decide the optimal for each decision (atlas, nff vs frf, ant vs fsl, nsc vs scr, gsr vs ngs) using anova test.

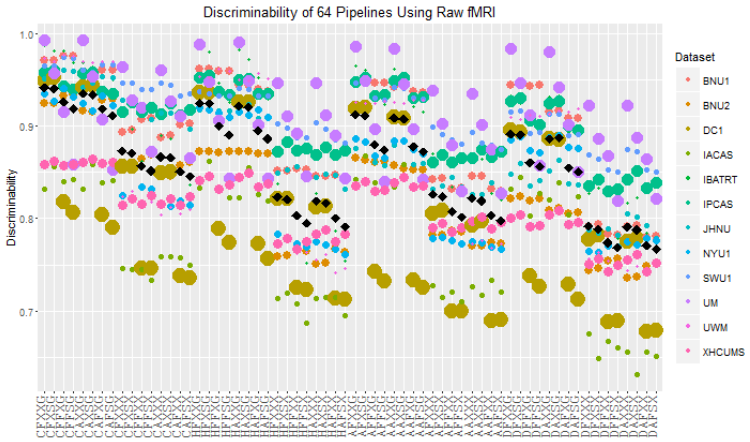

Figure 5: **Discriminability of raw fmri graphs from 12 data sets processed 64 ways.** Discriminability of BNU1, BNU2, DC1, IACAS, IBATRT, IPCAS, JHNU, NYU1, SWU1, UM, UWM and XHCUMS processed by 64 pipelines are computed. Color of dot indicates data set and size indicates the number of observations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

real experiment Describe how to convert raw graphs to rank graphs

real experiment Decide rank is better, especially when global signal regression is not performed.

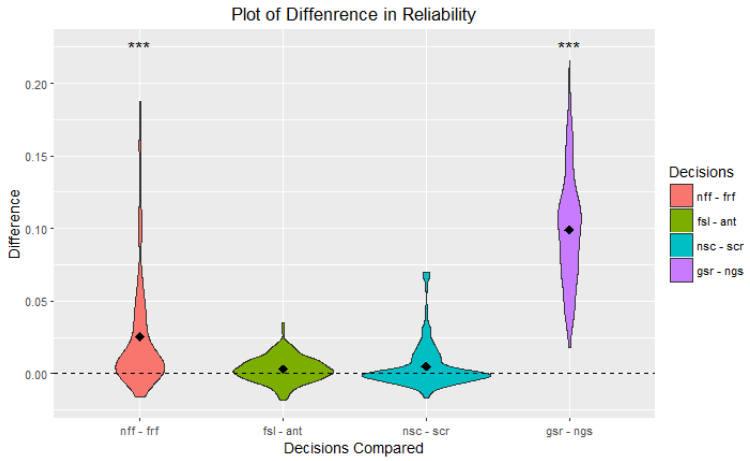

Figure 6: **Paired difference in discriminability of decisions.** Difference in discriminability for each decision in pipeline is compared by fixing other decisions and data sets. nff and gsr are statistical significantly better than the alternatives. fsl and nsc are not significantly better than the alternatives.

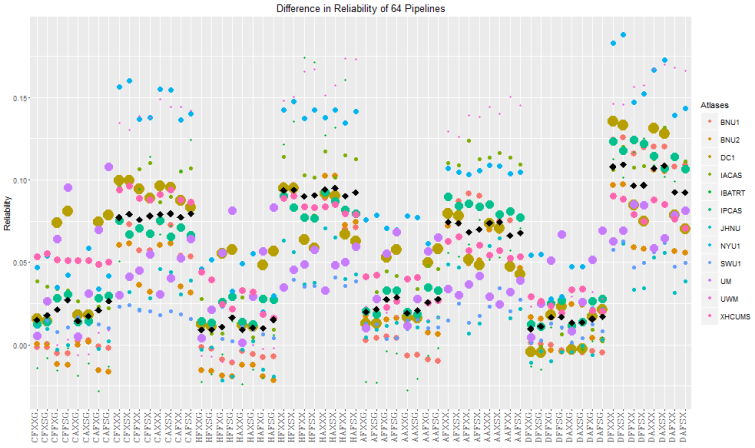

Figure 7: **Paired difference in discriminability between rank and raw graphs.** Difference in discriminability for rank and raw fmri graphs are computed for 12 data sets processed using 64 pipelines. Rank fmri graphs are more discriminable than raw fmri graphs.

II.C.3 DTI processing pipelines

real experiment Describe we process a dti data data set with rank, raw and log for 15 atlases.

real experiment We see a trend that large roi are better. Since discriminability after removing outliers is close to one, more experiments need to be done.

II.C.4 DTI vs. fMRI

real experiment Describe the 4 fmri and dti data sets.

real experiment After removing outliers, dti is more discriminable than fmri.

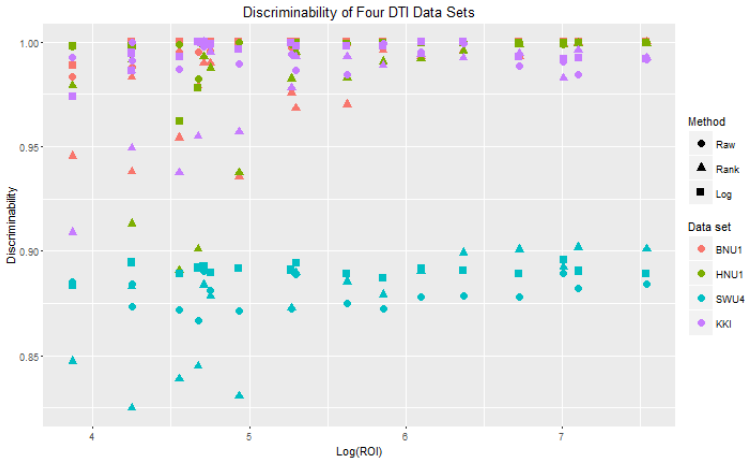

Figure 8: **Discriminability of 15 atlases.** Discriminability of SWU4 DTI registered with 15 atlases are computed. Raw, rank and log edges weights are considered. Atlases with a larger number of ROIs tend to be more discriminable.

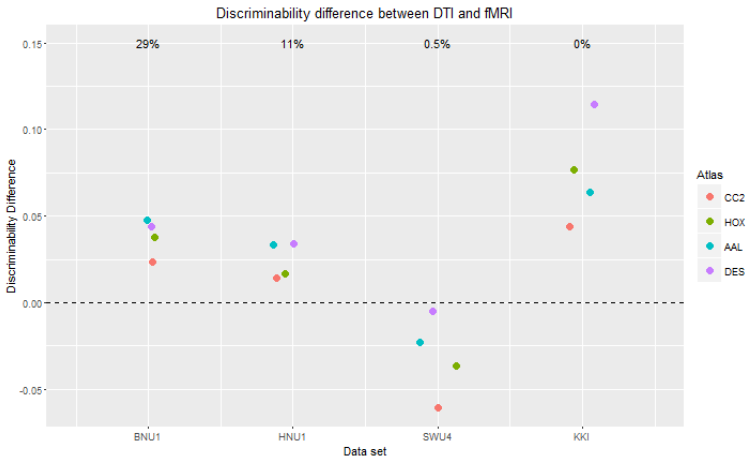

Figure 9: **Paired difference in discriminability between dti and fmri data sets.** Discriminability of DTI and fMRI graphs are computed for BNU1, HNU1, SWU4 and KKI data set. The number at the top indicates the percentage of outliers in DTI data sets. After removing outliers, DTI data sets are more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

III Discussion

Summary We propose a definition of discriminability and apply it to a variety of set ups.

Related Work I2C2, DISCO and GICC

Next Steps

IV Appendix