

neurodata.io



HOCKEY: How to Observe Connectomes for gaining Knowledge and Estimating 'Y'

Greg Kiar

March 7th, 2016

Significance

Fact: 43.8 million adults experience mental illness in a given year.



1 in 5 adults in America experience a mental illness.



Nearly 1 in 25 (10 million) adults in America live with a serious mental illness.



One-half of all chronic mental illness begins by the age of 14; three-quarters by the age of 24.

Impact



1st

Depression is the leading cause of disability worldwide, and is a major contributor to the global burden of disease.¹



-\$193b

Serious mental illness costs America \$193.2 billion in lost earning every year.³



90%

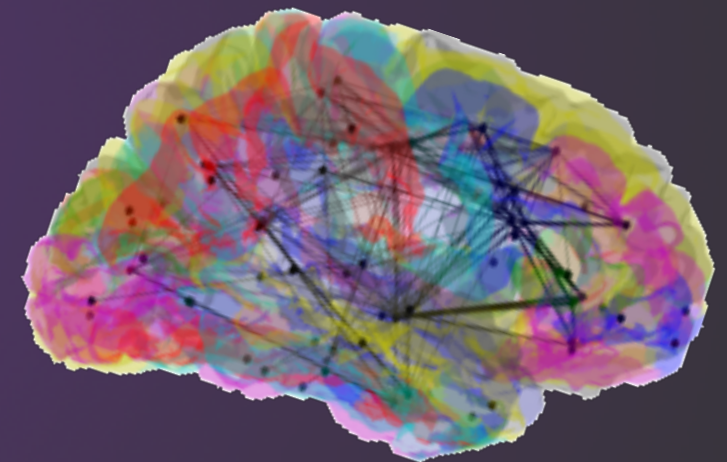
90% of those who die by suicide have an underlying mental illness. Suicide is the 10th leading cause of death in the U.S.³

Gap

- Currently, no brain imaging biomarkers exist that are clinically useful for any diagnostic category in psychiatry
- Currently, no known open source gender (or other covariate) classifiers based on connectome analysis exist

Challenge

- Graph stats are hard:
 - Large number of dimensions (D)
 - Small number of samples (N)
 - Observations are noisy
 - Batch effects across studies



Formal Statement of Problem

$G_i, Y_i \sim \mathcal{F} = \{F_{G,Y}(\cdot; \theta) : \theta \in \Theta\}$ Graphs and labels observed

$Y_i = \{0, 1\}$ Y is our binary label vector

CLASSIFY

$l = \sum \mathbb{I}(\hat{Y}_i \neq Y_i)$ We calculate loss based on correct assignment

$E[l] = \sum \mathbb{I}(\hat{Y}_i \neq Y_i) / N$ And get expected loss

Model Assumptions

$$G_i, Y_i \stackrel{i.i.d}{\sim} F \in \mathcal{F}$$

Graphs are i.i.d.

$$F_{GY} = F_{G|Y} F_Y$$

$$F_{G|Y} \in \mathcal{F}_{G|Y}$$

$$F_Y = \text{Bern}(\pi)$$

A class conditional
difference exists

Formal Statement of Algorithm

Algorithms used:

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K-Nearest Neighbours (KNN)
- Support Vector Machine (SVM)
- Random Forrest (RF)

QDA:

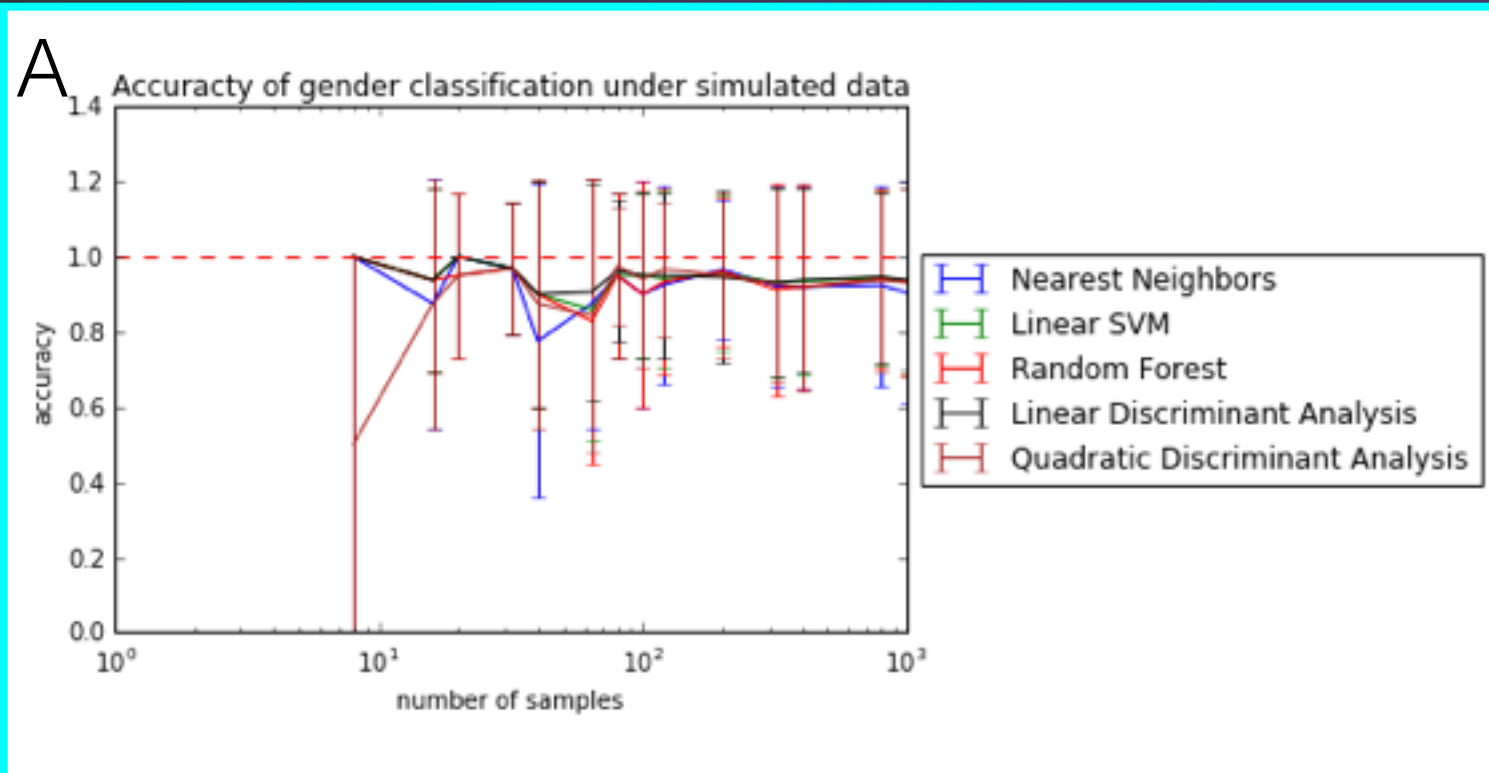
Quadratic Score Function:

$$s_i^Q(\mathbf{x}) = -\frac{1}{2} \ln |\Sigma_i| - \frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_1) + \ln p_i$$

Becomes Decision Rule:

$$s_i^Q(\mathbf{x}) = -\frac{1}{2} \ln |S_i| - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}})^T S_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}) + \ln p_i$$

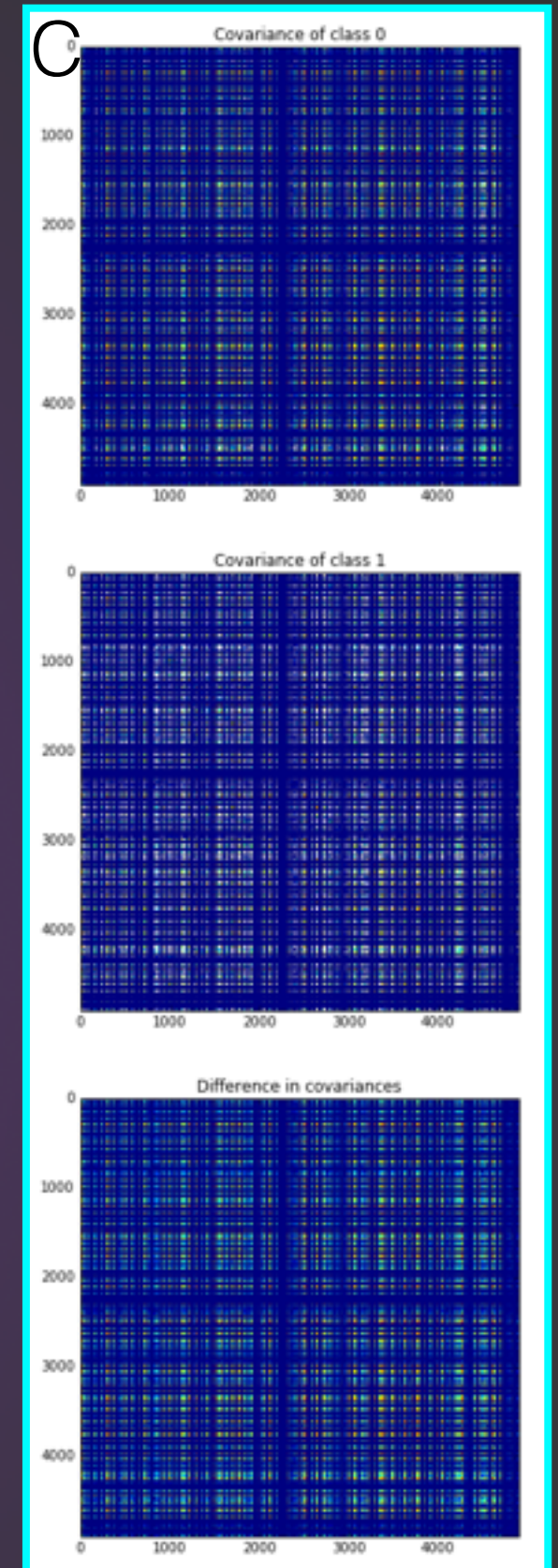
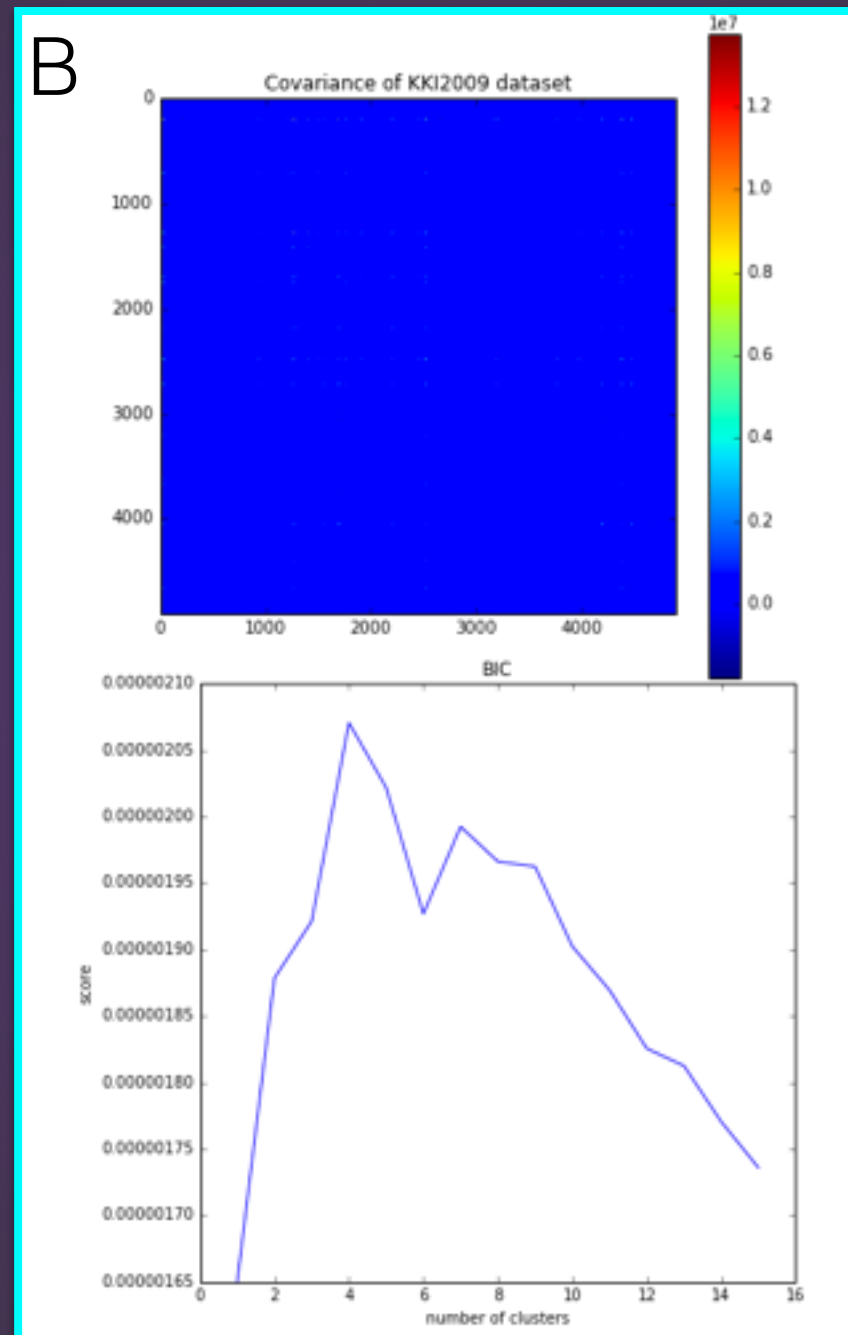
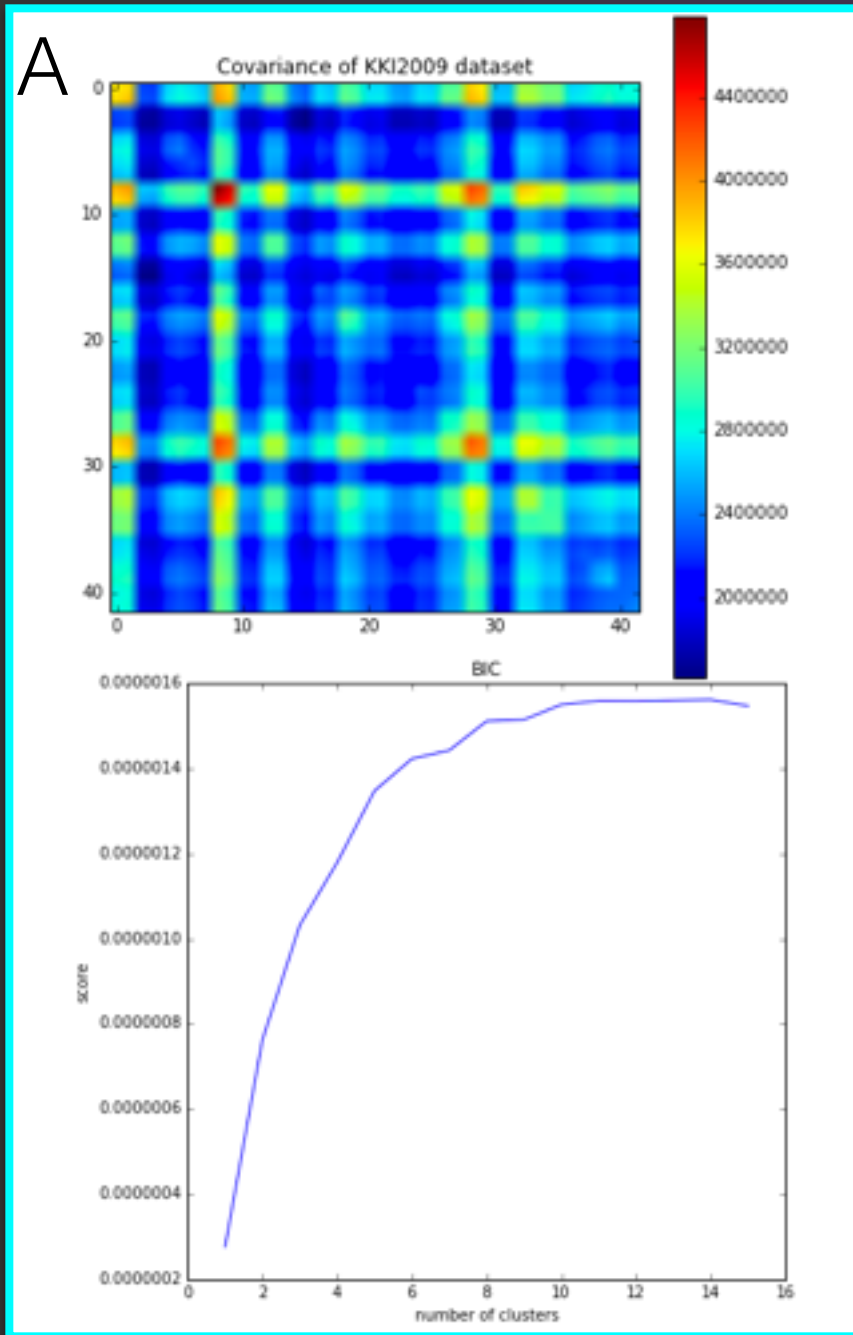
Results



B	Algorithm	Classification Accuracy
	Nearest Neighbors	0.48 (+/- 1.00)
	Linear SVM	0.55 (+/- 1.00)
	Random Forest	0.57 (+/- 0.99)
	Linear Discriminant Analysis	0.45 (+/- 1.00)
	Quadratic Discriminant Analysis	0.71 (+/- 0.90)

A. The performance of several classifiers on simulated data from our model, showing performance as the number of samples scaled. All algorithms scaled well with N and achieved near perfect accuracy in classification under these conditions. B. Performance of the same classifiers on the given dataset consisting of 42 graphs with 20 and 22 members of class 0 and 1, respectively, and each graph containing 70 nodes. The best classifier used here was the QDA method, which achieved just over 70% classification accuracy.

Model Checking



A. We show that the covariance across subjects is non zero and the ideal number of clusters is > 1 , indicating both our graph i.i.d. assumptions were false. B. Like A, the same is true for the edge i.i.d. assumption. C. We compared the covariances across classes and found that the difference was very large, suggesting why QDA performed better for classification than LDA, for instance.

Resolution

- There now exists an open source implementation of QDA for gender classification on human connectome estimates
- This is a building block for future studies which can expand the use of this classifier to other covariates