FDR-CORRECTED SPARSE CANONICAL CORRELATION ANALYSIS WITH APPLICATIONS TO IMAGING GENOMICS

ALEXEJ GOSSMANN

TULANE UNIVERSITY

2017/09/27

BACKGROUND

SPARSE CANONICAL CORRELATION ANALYSIS

$$\underset{u,\,v}{\text{maximize}}\ \text{Cor}\left(\begin{array}{c}\text{gene 1}\ \text{gene 2}\ \text{gene 3}\ \ \text{gene } P_x \\ \begin{array}{l}\text{subject 1}\\ \text{subject 2}\\ \\ \\ \text{subject n}\end{array}\begin{bmatrix} & & & \\ & & & \\ & & X & \\ & & & \\ & & & \end{bmatrix} \times \begin{bmatrix}u_1\\u_2\\u_3\\ \\ u_{P_x}\end{bmatrix}\end{array},\ \begin{array}{c}\text{voxel 1}\ \text{voxel 2}\ \text{voxel 3}\ \ \text{voxel } P_y \\ \begin{array}{l}\text{s. 1}\\ \text{s. 2}\\ \\ \\ \text{s. n}\end{array}\begin{bmatrix} & & & \\ & & & \\ & & Y & \\ & & & \\ & & & \end{bmatrix} \times \begin{bmatrix}v_1\\v_2\\v_3\\ \\ v_{P_y}\end{bmatrix}\end{array}\right)$$

subject to sparsity (and other) conditions on u and v.

👉 Find a subset of genes and a subset of brain voxels that are related to each other. 👍

CANONICAL CORRELATION ANALYSIS

Let $x_1, \dots, x_n \in \mathbb{R}^p$ be independent $\mathcal{N}(0, \Sigma_X)$,
$y_1, \dots, y_n \in \mathbb{R}^q$ be independent $\mathcal{N}(0, \Sigma_Y)$,
$\mathrm{Cov}(x_k, y_k) = \Sigma_{XY} \in \mathbb{R}^{p \times q}$ for all $k \in \{1, \dots, n\}$,
and that $\mathrm{Cov}(x_k, y_j) = 0$ whenever $k \neq j$.

$$X := \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}, \qquad Y := \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{bmatrix} \in \mathbb{R}^{n \times q}.$$

CLASSICAL CANONICAL CORRELATION ANALYSIS

$$\text{maximize}_{u \in \mathbb{R}^p, v \in \mathbb{R}^q} \widehat{\text{Cov}}(Xu, Yv) = \frac{1}{n} u^T X^T Y v,$$

$$\text{subject to} \quad \widehat{\text{Var}}(Xu) = 1, \ \widehat{\text{Var}}(Yv) = 1.$$

- Due to Hotelling, 1936.
- The solution is called first pair of canonical vectors.
- Subsequent pairs of canonical vectors are restricted to be uncorrelated with the previous ones.
- The problem is degenerate if $n \leq \max(p, q)$.

SPARSE CCA (SCCA)

- Regularization to achive sparsity (e.g., ℓ_1-norm).
- Unique solution even when $p_X, p_Y \gg n$.
- Witten et. al. (2009):

$$\text{maximize}_{u \in \mathbb{R}^p, v \in \mathbb{R}^q} \frac{1}{n} u^T X^T Y v,$$

$$\text{subject to} \quad \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1, \|u\|_1 \leq c_1, \|v\|_1 \leq$$

- *Selection of the sparsity parameters remains a challenging problem* (cross-validation, AIC, permutation-based).
- Higher-order pairs of canonical vectors can be found by applying SCCA to a residual matrix ($X^T Y - duv^T$).

FDR-CORRECTION FOR SPARSE CCA

MULTIPLE HYPOTHESES CORRECTION

- Denote $R :=$ number of rejected hypotheses, and $V :=$ number of false rejections (i.e., Type I errors).
- *Family-wise error rate*:

$$\text{FWER} = \mathbb{P}\left(\text{At least one false rejection}\right) = \mathbb{P}(V \geq 1).$$

E.g. Bonferroni correction (60ies?):

$$\mathbb{P}(V \geq 1) \leq \mathbb{P}\left(\bigcup_{i=1}^{n} \{H_i \text{ falsely rejected}\}\right) \leq \sum_{i=1}^{n} \underbrace{\mathbb{P}\left(\{H_i \text{ falsely rejected}\}\right)}_{\leq \alpha/n} \leq \alpha.$$

- *False discovery rate* ('95):

$$\text{FDR} = \mathbb{E}\left(\frac{\#\text{False rejections}}{\#\text{Rejections}}\right) = \mathbb{E}\left(\frac{V}{\min\{R, 1\}}\right).$$

E.g. Benjamini-Hochberg:

1. Sort the p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$.
2. Find the largest k such that $p_{(k)} \leq \frac{k}{n}\alpha$.
3. Reject the null hypothesis for all $H_{(i)}$ for $i = 1, \ldots, k$.

DEFINING FALSE DISCOVERY RATE (FDR) FOR SPARSE CCA

- Consider the FDR in u and in v separately.
- Consider p_X hypotheses tests $H_i : u_i = 0$.
- The null hypothesis H_i is true if the ith feature in X is uncorrelated with all features in Y, i.e., if
$$(\forall j \in \{1, 2, \ldots, p_Y\}) : \rho_{i,j}^{XY} = 0.$$
- Let $R_{\hat{u}}$ be the number of the rejected H_i, and $V_{\hat{u}}$ the number of false rejections (i.e., when $\hat{u}_i \neq 0$ but $\rho_{i,j}^{XY} = 0$ for all j).
- Define the false discovery rate in u as
$$\mathrm{FDR}(\hat{u}) := \mathbb{E}\left(\frac{V_{\hat{u}}}{\max\{R_{\hat{u}}, 1\}}\right).$$

FDR-CORRECTED SPARSE CCA

- In the classical CCA problem $u \propto X^T Y v$ (b/c SVD), and $v \propto Y^T X u$.
- Thus, the above tests are equivalent to

$$H_i : \left(X^T Y v\right)_i = 0, \quad i \in \{1, 2, \ldots, p_X\}.$$

- This motivates an FDR-correcting approach:
  1. Obtain initial estimates $\hat{u}^{(0)}$ and $\hat{v}^{(0)}$
  2. Then in order to determine which entries of u and v are truly non-zero, test null hypotheses of the form

$$\mathrm{H}_i^{(u)} : \left(X^T Y \hat{v}^{(0)}\right)_i = 0, \quad i = 1, 2, \ldots, p_X,$$

$$\mathrm{H}_j^{(v)} : \left(Y^T X \hat{u}^{(0)}\right)_j = 0 \quad j = 1, 2, \ldots, p_Y.$$

ASYMPTOTIC DISTRIBUTION

Theorem 1 (Asymptotic Normality)**.** *Let the random matrices X and Y be defined as above. For any vector $\mathbf{v} \in \mathbb{R}^{p_Y}$, it holds that*

$$\sqrt{n}\left(\frac{1}{n}X^T Y\mathbf{v} - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Omega), \tag{9}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{p_X}$ has entries

$$\mu_i = \sum_{j=1}^{p_Y} v_j \rho_{i,j}^{X,Y}, \tag{10}$$

and where $\Omega \in \mathbb{R}^{p_X \times p_X}$ has entries

$$\omega_{i,j} = \left(\sum_{k=1}^{p_Y} v_k \rho_{i,k}^{X,Y}\right)\left(\sum_{k=1}^{p_Y} v_k \rho_{j,k}^{X,Y}\right) + \rho_{i,j}^{X} \mathbf{v}^T \Sigma_Y \mathbf{v}. \tag{11}$$

THE FDR-CORRECTED SPARSE CCA PROCEDURE

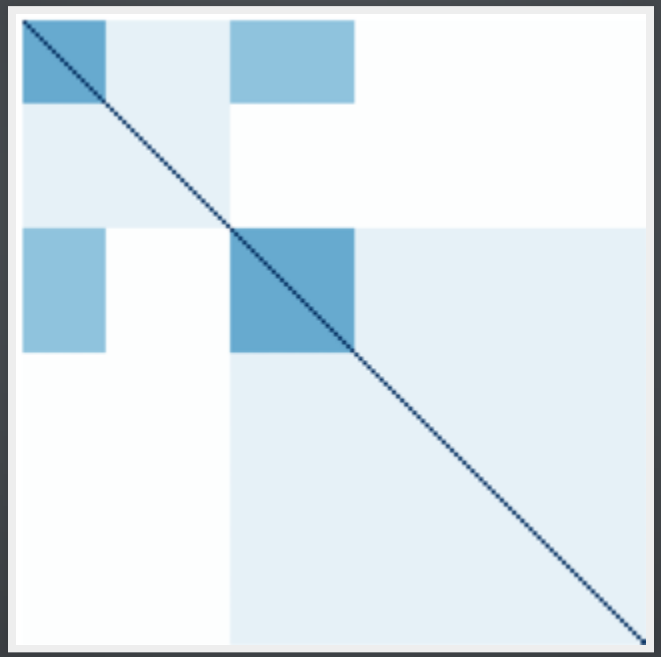
1. Divide each of X and Y into two subsets of sizes n_0 and n_1:
$$X = \begin{bmatrix} X^{(0)} \\ X^{(1)} \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Y^{(0)} \\ Y^{(1)} \end{bmatrix}.$$
2. Obtain preliminary sparse CCA estimates $\hat{u}^{(0)}$ and $\hat{v}^{(0)}$ on $X^{(0)}$ and $Y^{(0)}$. Additionally, use $X^{(0)}$ and $Y^{(0)}$ to obtain $\widehat{\Sigma}^{(0)}$, the ML estimate of $\mathrm{Cov}\left(\begin{bmatrix} X & Y \end{bmatrix}\right)$.
3. Obtain p-values using the asymptotic approximation (under the null)
$$\left(\frac{1}{\sqrt{n}} \left(X^{(1)}\right)^T Y^{(1)} \hat{v}^{(0)} \middle| \Sigma = \widehat{\Sigma}^{(0)} \right) \sim \mathcal{N}\left(0, \widehat{\Omega}^{(0)}\right),$$
where $\hat{\mu}^{(0)}$ and $\widehat{\Omega}^{(0)}$ are available in explicit form ($\hat{\mu}^{(0)} = 0$ under the null hypothesis).
4. Apply an FDR correcting procedure (such as BHq), and obtain the FDR-corrected estimates:
$$\hat{u}_i^{(1)} := \begin{cases} \left(X^T Y \hat{v}^{(0)}\right)_i, & \text{for any rejected } H_i^{(u)}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$
$$\hat{v}_j^{(1)} := \begin{cases} \left(Y^T X \hat{u}^{(0)}\right)_j, & \text{for any rejected } H_j^{(v)}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

SIMULATION RESULTS

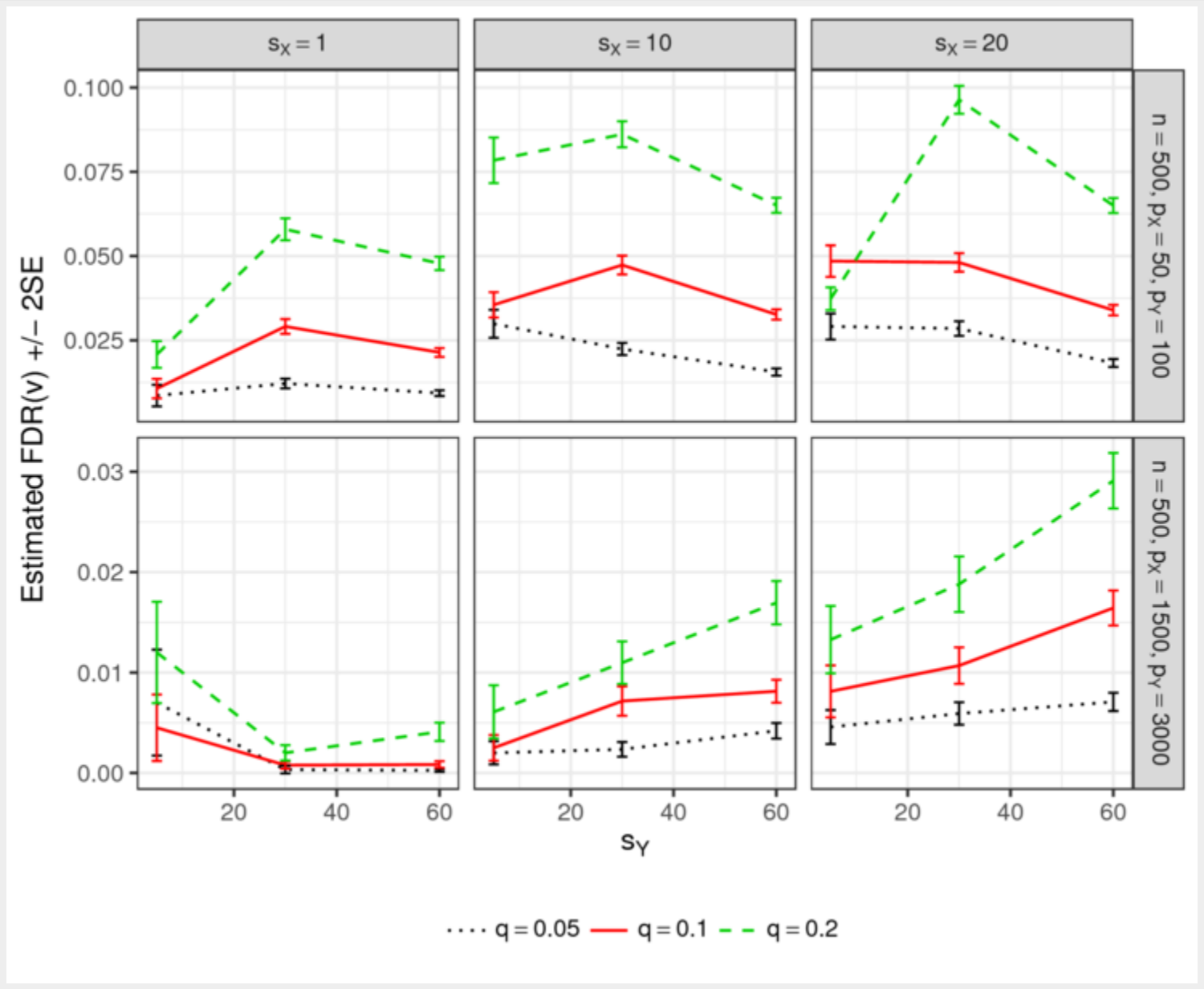
1. We show simulation results under **Gaussian scenarios**, in order to verify that the proposed procedure indeed controls the FDR under the assumptions that its derivation relies on.
2. We show simulation studies evaluating the performance on **non-Gaussian data**, which are generated based on real single-nucleotide polymorphism (SNP) data.

SIMULATION STUDY WITH GAUSSIAN DATA

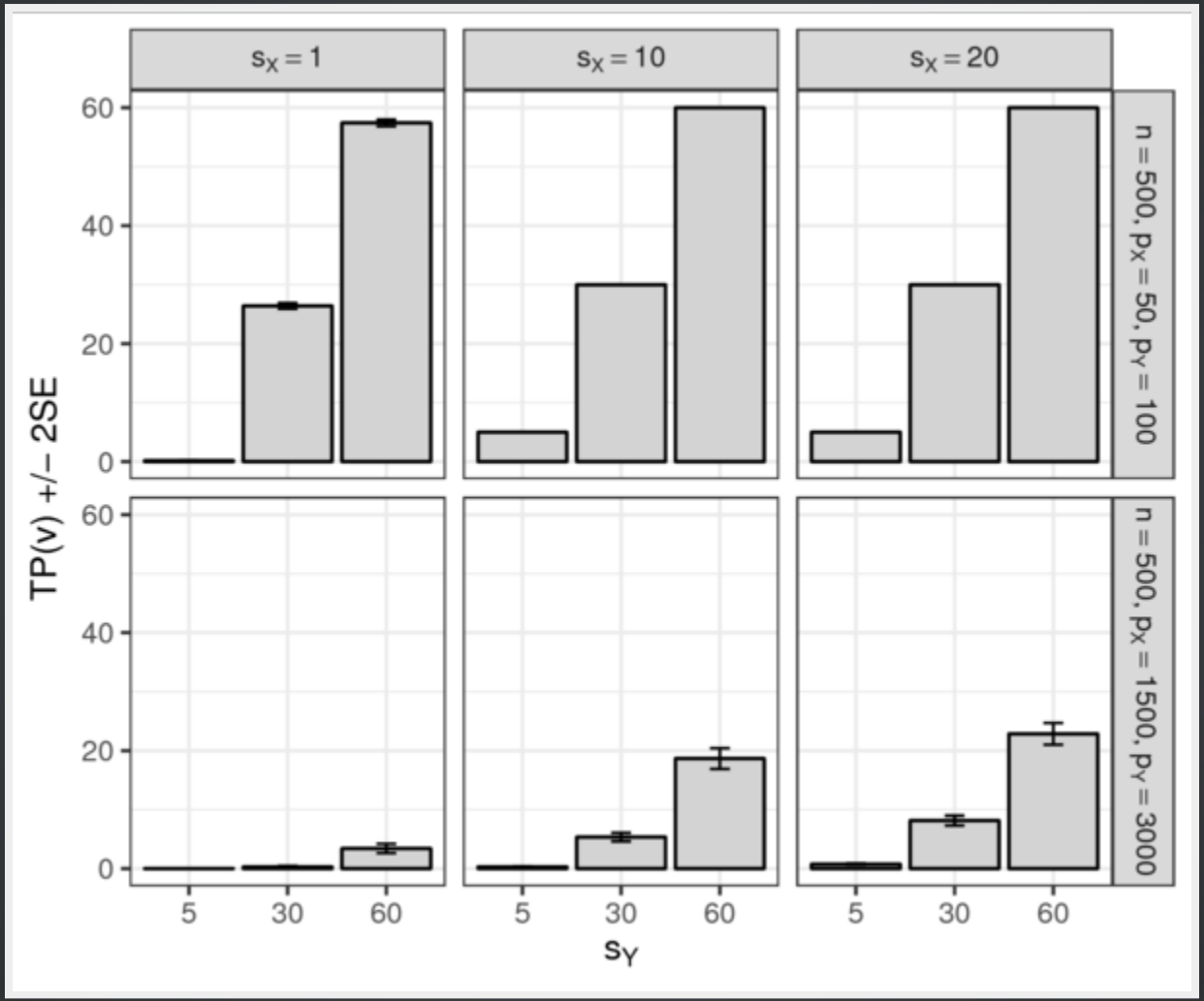
The data $\begin{bmatrix} X & Y \end{bmatrix}$ are generated from $\mathcal{N}(0, \Sigma)$, where Σ is blockwise constant.



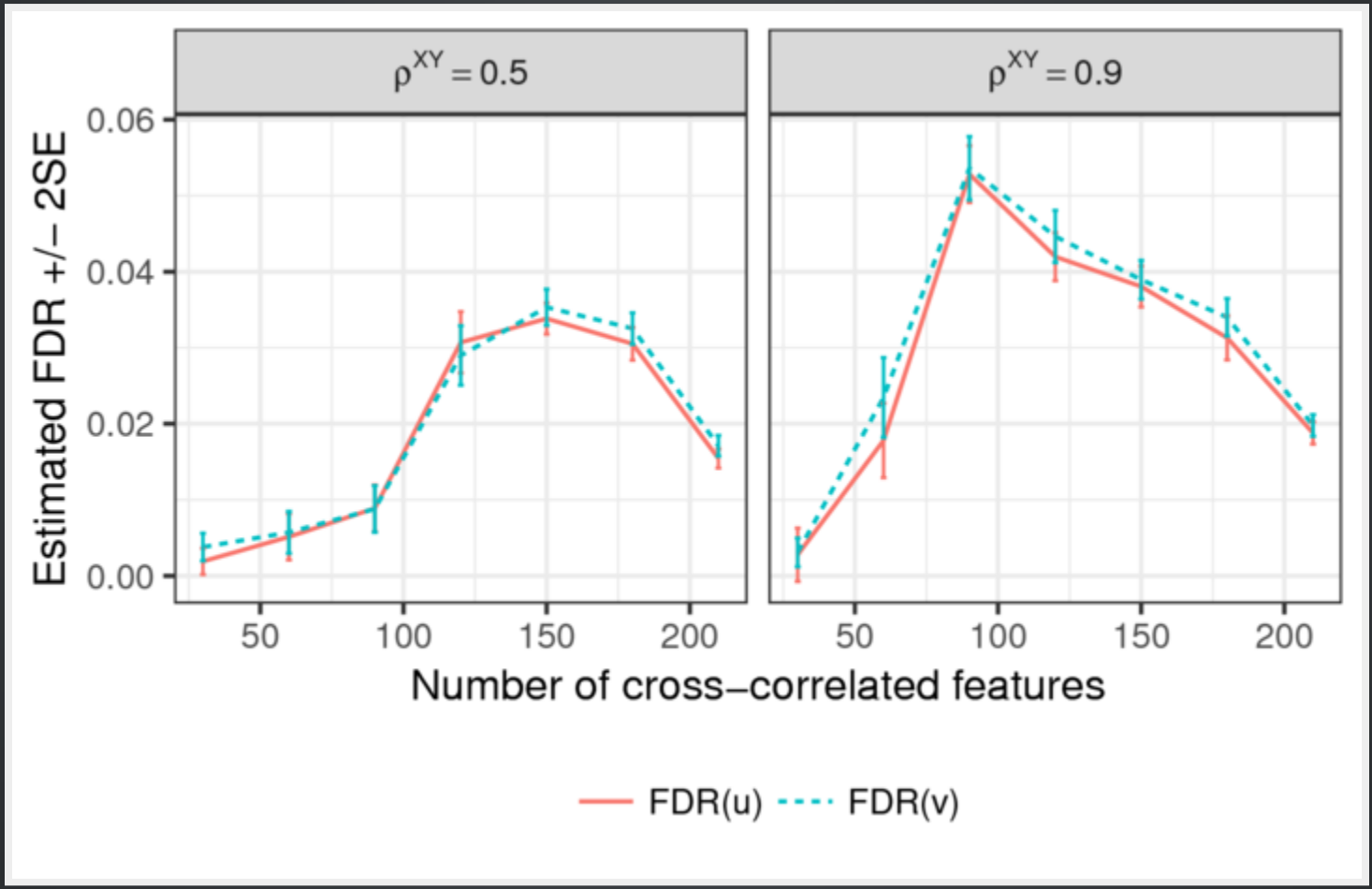
SIMULATION STUDY WITH GAUSSIAN DATA



SIMULATION STUDY WITH GAUSSIAN DATA



SIMULATION STUDY WITH NON-GAUSSIAN DATA (INVESTIGATING ROBUSTNESS TO DISTRIBUTIONAL ASSUMPTIONS)



FDR-CORRECTED SCCA APPLICATION TO IMAGING GENOMICS

APPLICATION TO IMAGING GENOMICS

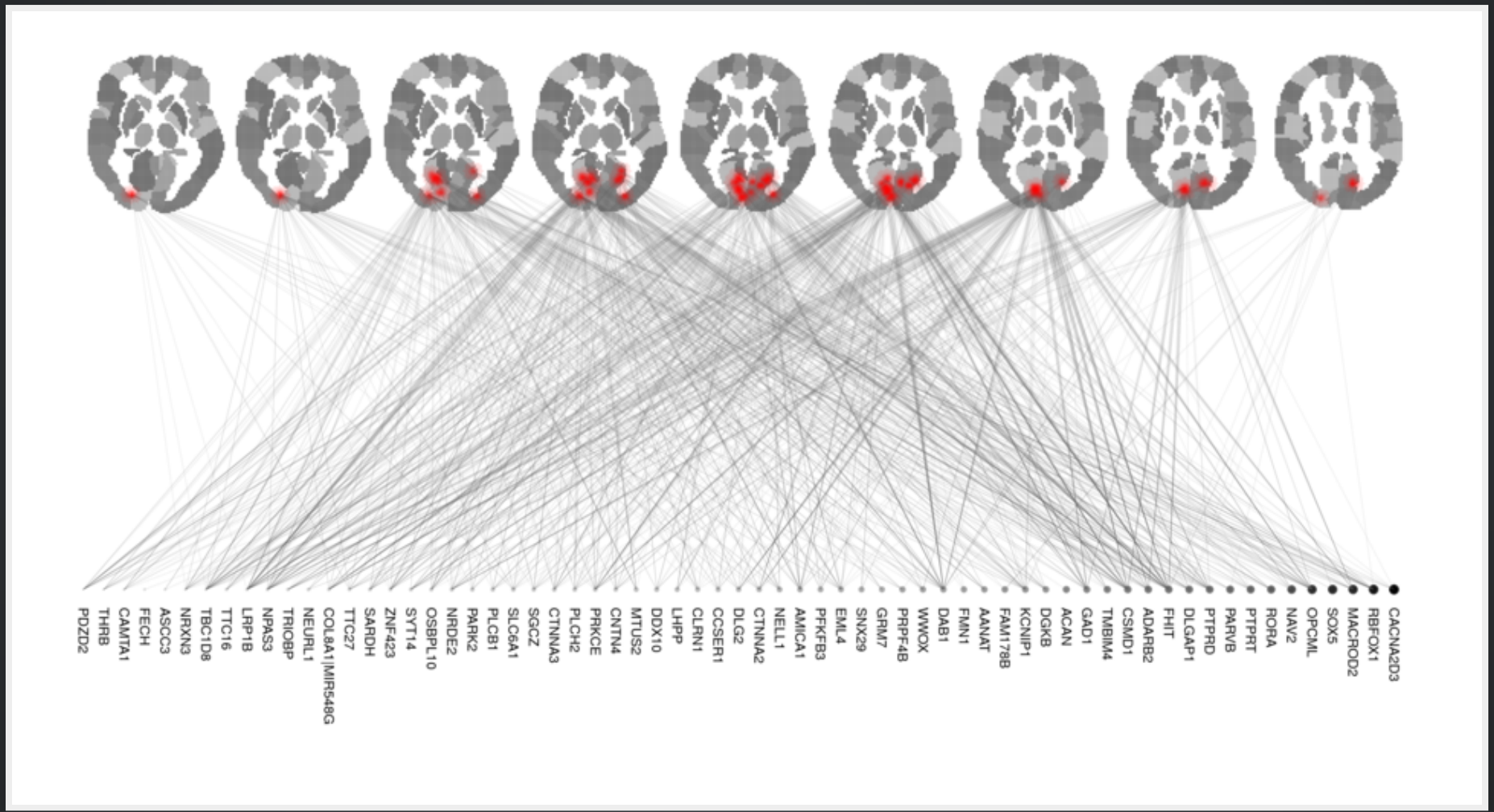
DATA

The Philadelphia Neurodevelopmental Cohort (PNC) is a large-scale collaborative study between the Brain Behaviour Laboratory at the University of Pennsylvania and the Children's Hospital of Philadelphia. It contains, among other modalities, a fractal n-back fMRI task, and SNP arrays for over 900 adolescents.

DATA

- The fractal n-back fMRI data were pre-processed using SPM12. Stimulus-on versus stimulus-off contrast maps were extracted for analysis. After discarding voxels with more than 1% missing data, the dataset consists of $85,796$ voxels.
- The SNP dataset contains $98,804$ SNPs (after pre-processing). PCA was performed within each gene to reduce dimensionality, resulting in $60,372$ genomic features.
- Our goal is to identify the essential regions of cross-correlation between the brain voxels and the genomic features.

RESULTS


CACNA2D3
RBFOX1
MACROD2
SOX5
OPCML
NAV2
RORA
PTPRT
PARVB
PTPRD
DLGAP1
FHIT
ADARB2
CSMD1
TMBIM4
GAD1
ACAN
DGKB
KCNIP1
FAM178B
AANAT
FMN1
DAB1
WWOX
PRPF4B
GRM7
SNX29
EML4
PFKFB3
AMICA1
NELL1
CTNNA2
DLG2
CCSER1
CLRN1
LHPP
DDX10
MTUS2
CNTN4
PRKCE
PLCH2
CTNNA3
SGCZ
SLC6A1
PLCB1
PARK2
NRDE2
OSBPL10
SYT14
ZNF423
SARDH
TTC27
COL8A1|MIR548G
NEURL1
TRIOBP
NPAS3
LRP1B
TTC16
TBC1D8
NRXN3
ASCC3
FECH
CAMTA1
THRB
PDZD2

RESULTS

- We group the selected voxels using the ROI definitions of the AAL parcellation. The most significant findings correspond to the *middle occipital gyri* (13 voxels). Additional selected voxels lie in the *left and right calcarine sulcus* (158 voxels), and *left cuneus* (3 voxels). Similar brain regions have been found in other fMRI studies of working memory.
- A literature search confirmed that a majority of the identified genes (at least 34 out of the 65) have been previously associated with various aspects of human cognitive function.

FDR-CORRECTED SCCA REFERENCES

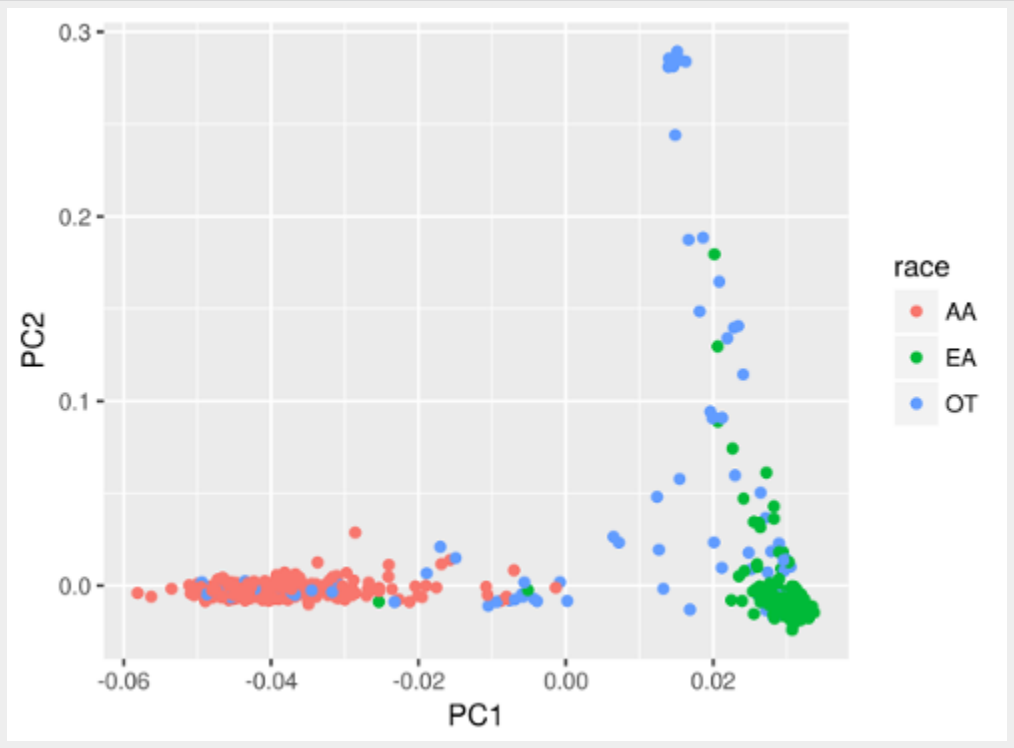
- Gossmann, A., Zille, P., Calhoun, V., & Wang, Y.-P. (2017). FDR-Corrected Sparse Canonical Correlation Analysis with Applications to Imaging Genomics. arXiv:1705.04312 [pdf] *(under review in IEEE/TMI)*
- Associated code: https://github.com/agisga/FDRcorrectedSCCA

FURTHER DEVELOPMENTS

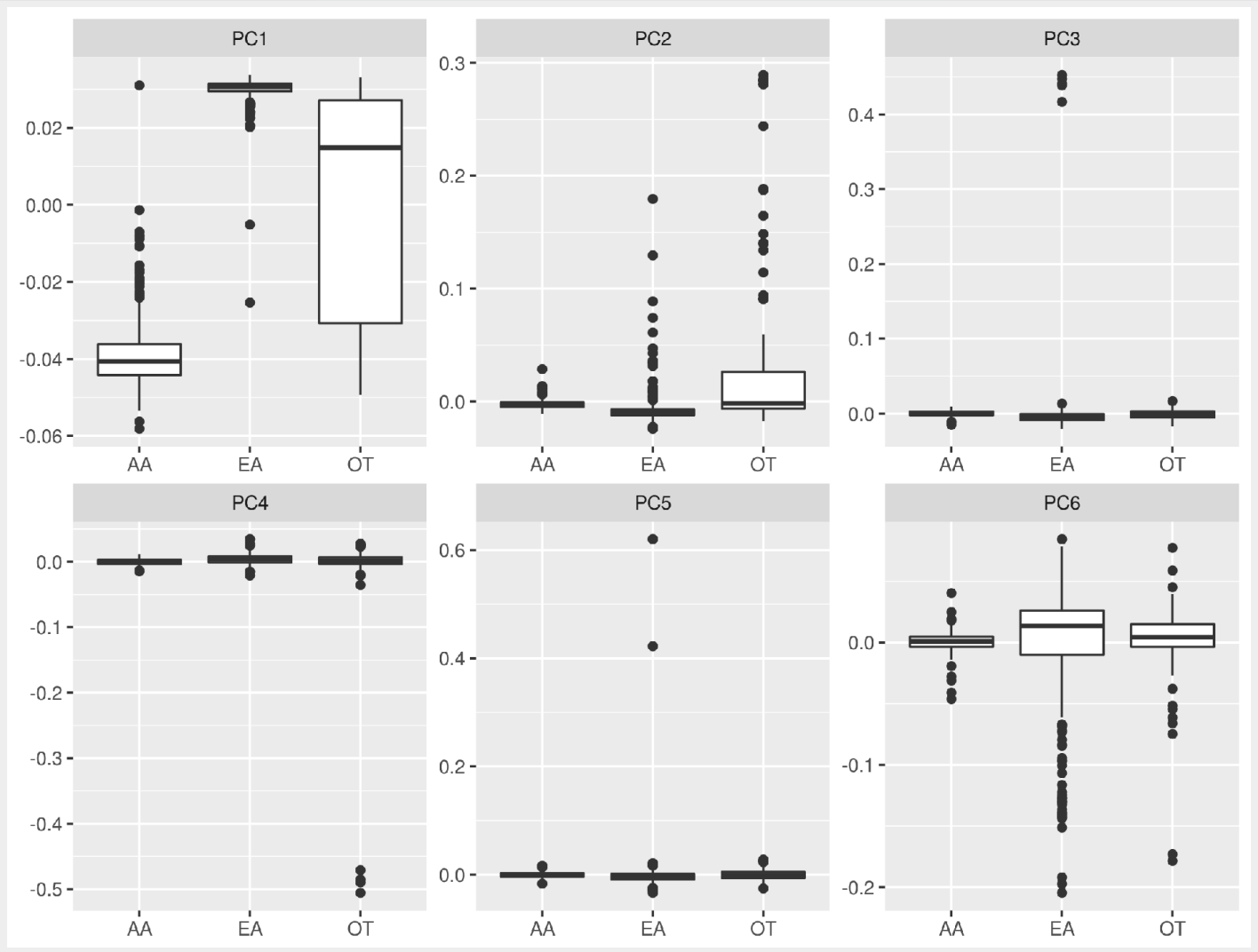
POPULATION STRATIFICATION AND CONFOUNDERS

Differences w.r.t. ethnicity and genotyping plattform appear to have an influence on the results.

⇝ EIGENSTRAT approach (Price et. al., 2006)



POPULATION STRATIFICATION AND CONFOUNDERS



APPLICATION TO FUNCTIONAL CONNECTIVITY NETWORKS

- SCCA can be applied to FCN data arising from different fMRI runs for the same set of subjects (e.g., NB vs. EM vs. Rest fMRI).
- Because FCN constitute a unique fingerprint for each subject, SCCA will pair the same connectome features across different modalities.
- Functional network connectivity patterns can also be calculated from group spatial ICA time courses.
- *Results:* Too many cross-correlation, not sufficiently sparse.

THE END

THANK YOU