CONTROLLING THE FALSE DISCOVERY RATE IN SPARSE AND HIGH-DIMENSIONAL STATISTICAL METHODS, WITH APPLICATIONS IN GENOMICS AND IMAGING

ALEXEJ GOSSMANN

TULANE UNIVERSITY

2017/05/21

BACKGROUND

THE MODEL SELECTION PROBLEM

- The simplest example: 👍 linear model.
- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where
  - $\mathbf{y} \in \mathbb{R}^n$ dependent variable (e.g., disease status/severity for n 😷),
  - $X \in \mathbb{R}^{n \times p}$ explanatory variables (for each 😷 record p features: 🚬, 💉, 💊, SNPs, fMRI, ...),
  - *Unknowns*: $\boldsymbol{\beta} \in \mathbb{R}^p$ (want to estimate), $\boldsymbol{\varepsilon}$ noise.
- **Prediction**: Find best predictions for $\mathbf{y}$.
- **Feature selection**: Find which β_i are non-zero.

...IN GENOMICS AND BRAIN IMAGING

..WHY WE CARE

🔬 + 💰

- Prediction of a disease phenotype based on a handful of features is needed for inexpensive diagnosis.
- Elimination of noisy or redundant features leads to more accurate prediction.
- "Data-generated hypotheses" lead to a better understanding of the underlying biology.

...IN GENOMICS AND BRAIN IMAGING

..CHALLENGES

- ~~*Possibly*~~ *Only slightly less often than always* $n \ll p$. 😓
- Curse of dimensionality (when $n < p$). 😰
- Overfitting. 😱
- Underfitting. 🙀

ℓ_1 REGULARIZATION (E.G. LASSO BY TIBSHIRANI, 1994)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1$$

- Yields a sparse $\hat{\boldsymbol{\beta}}$.
- Computationally efficient and very useful in practice.
- Problem 1: unclear how to select λ.
- Problem 2: unclear how to do statistical inference on $\hat{\boldsymbol{\beta}}$.

MULTIPLE HYPOTHESES TESTING PERSPECTIVE

Alternatively, feature selection can be regarded as testing the p hypotheses

$$H_i : \beta_i = 0, \quad i = 1, \dots, p.$$

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

THE MODEL SELECTION PROBLEM

Regression shrinkage and selection via the lasso

R Tibshirani - Journal of the Royal Statistical Society. Series B (..., 1996 - JSTOR Paperpile

We propose a new method for estimation in linear models. Thelasso'minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge ...

Cited by 21111 Related articles All 75 versions Import into BibTeX Save More

Controlling the false discovery rate: a practical and powerful approach to multiple testing

Y **Benjamini**, Y Hochberg - Journal of the royal statistical society. Series B (..., **1995** - JSTOR Paperpile

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected

Cited by 41325 Related articles All 50 versions Import into BibTeX Save More

SORTED L-ONE PENALIZED ESTIMATION (SLOPE, BOGDAN ET. AL., ANNALS APPL STAT, 2015)

$$\hat{\boldsymbol{\beta}}_{\text{SLOPE}} = \text{argmin}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \sum_{i=1}^{p} \lambda_i |\mathbf{b}|_{(i)},$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$; and $|b|_{(1)} \geq |b|_{(2)} \geq \ldots \geq |b|_{(p)}$ denotes the order statistic of the magnitudes of the vector $\mathbf{b} \in \mathbb{R}^p$.

$\rightsquigarrow$ Given $q \in (0, 1)$, there is a procedure to choose $\boldsymbol{\lambda}$ s.t. $\text{FDR}(\hat{\boldsymbol{\beta}}_{\text{SLOPE}}) \leq q$ is guaranteed, ...*if the explanatory variables have very small pair-wise correlations.*

GROUP SLOPE

GROUP SLOPE MOTIVATION

- Typically, genomic data are highly correlated.
- Often the data can be subdivided into groups with possibly a high within group correlation but a low between group correlation. ~~(Oh really?)~~
- *In case of biomedical data available prior knowledge often provides grouping structures naturally*. E.g., Genomic data: genes or genetic pathways; brain MRI data: anatomical atlases of brain regions; etc.

👍 Select or drop entire groups rather than individual variables. Redefine FDR w.r.t. groups (**gFDR**).

MODEL FORMULATION

- Let $X \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 I)$.
- The predictor variables $\boldsymbol{\beta}$ are divided into J groups of sizes $p_1, p_2, \cdots, p_J$, i.e. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_J^T)^T$ with $\boldsymbol{\beta}_i \in \mathbb{R}^{p_i}$.
- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \sum_{i=1}^J X_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}$.

GROUP SLOPE MODEL

FORMULATION 1 (GOSSMANN ET. AL. 2015)

$$\min_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \sum_{i=1}^{J} \lambda_i \sqrt{p_{(i)}}\|\mathbf{b}_{(i)}\|_2,$$

where

$$\sqrt{p_{(1)}}\|\mathbf{b}_{(1)}\|_2 \geq \sqrt{p_{(2)}}\|\mathbf{b}_{(2)}\|_2 \geq \dots \geq \sqrt{p_{(J)}}\|\mathbf{b}_{(J)}\|_2.$$

GROUP SLOPE MODEL

FORMULATION 2 (BRZYSKI ET. AL. 2016)

$$\min_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \sum_{i=1}^{J} \lambda_i \sqrt{p_{(i)}}\|X_{(i)}\mathbf{b}_{(i)}\|_2,$$

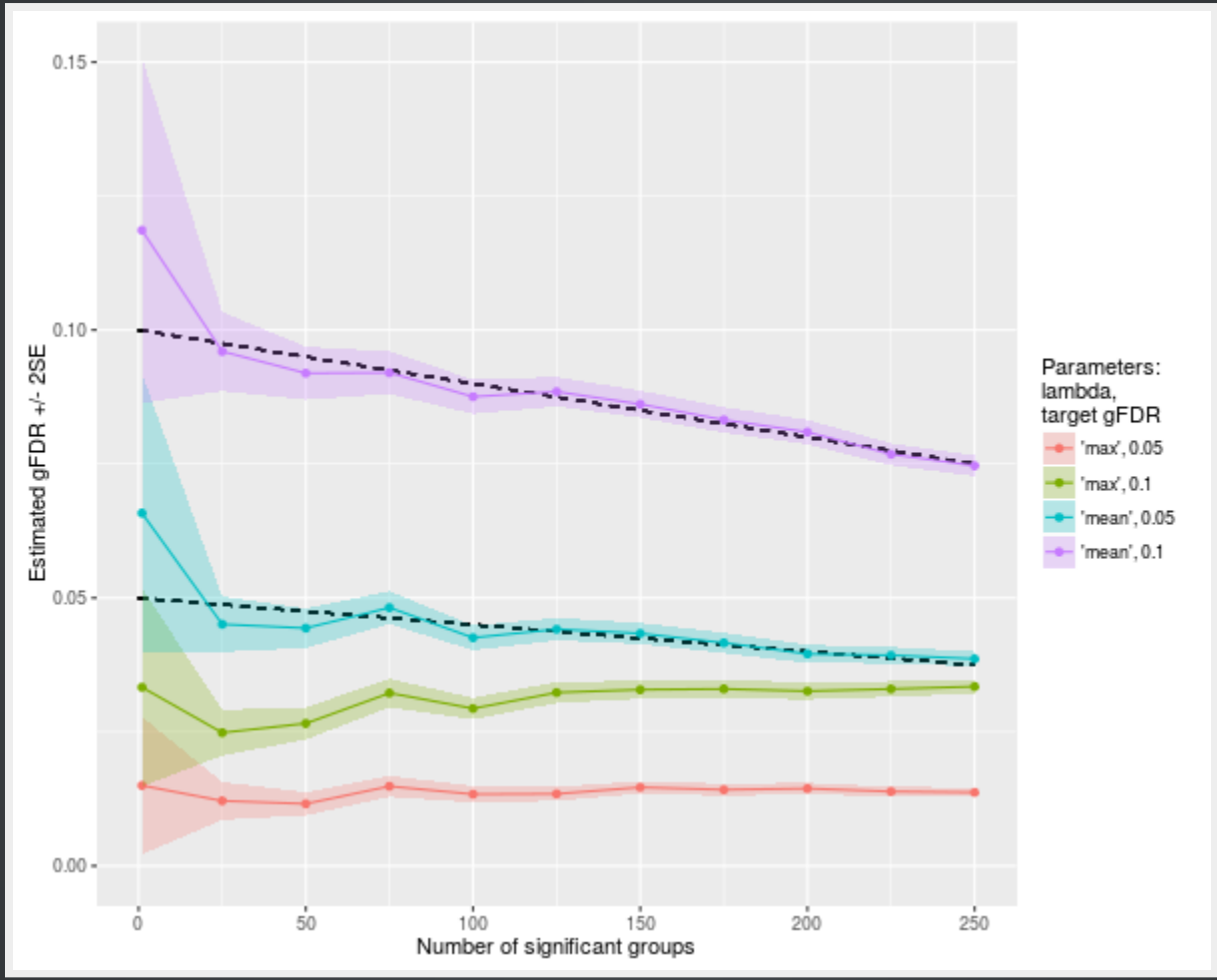
where

$$\sqrt{p_{(1)}}\|X_{(1)}\mathbf{b}_{(1)}\|_2 \geq \sqrt{p_{(2)}}\|X_{(2)}\mathbf{b}_{(2)}\|_2 \geq \ldots \geq \sqrt{p_{(J)}}\|X$$

GROUP SLOPE - THEORETICAL RESULTS

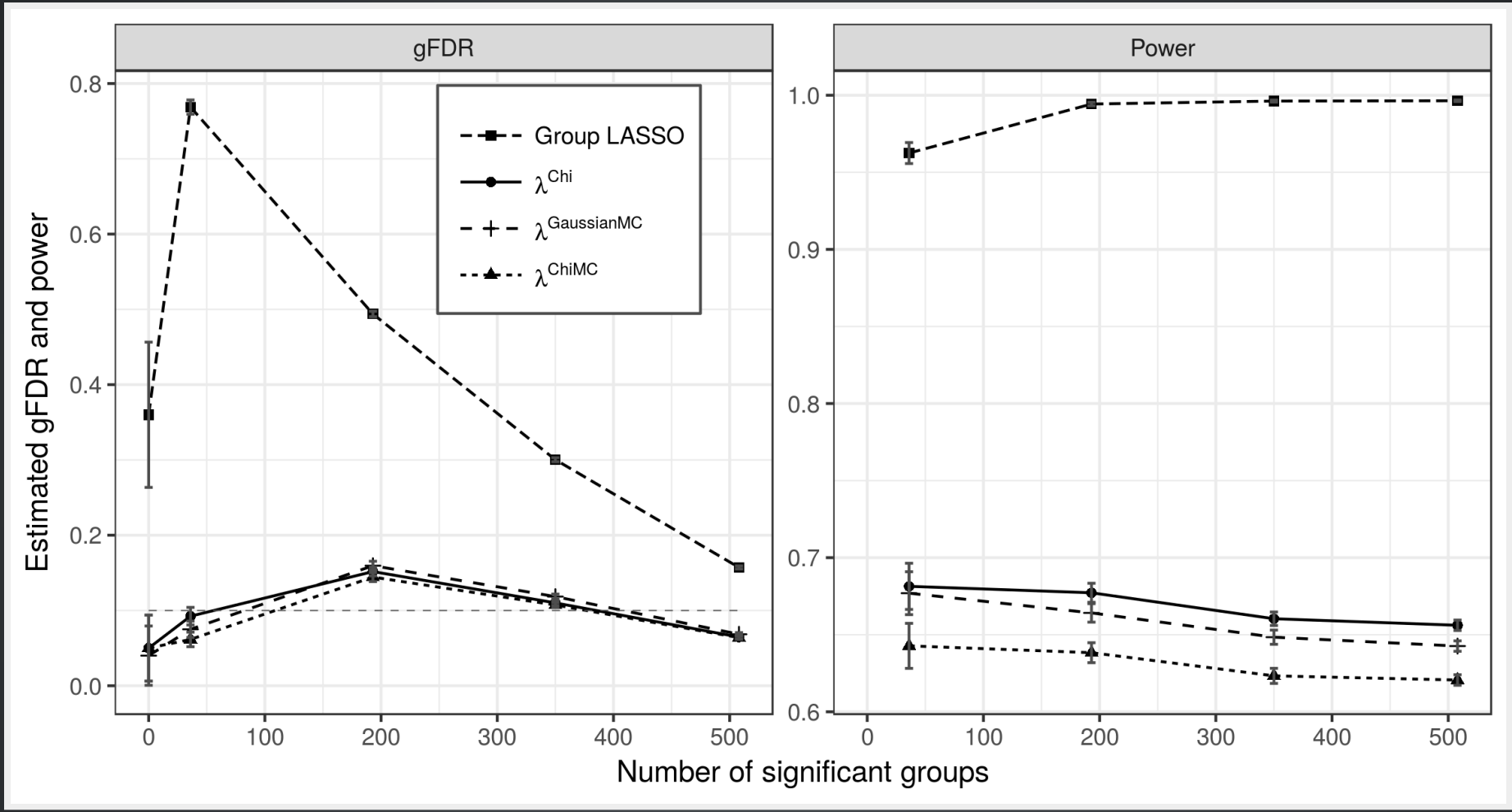
- Given a user-specified $q \in (0, 1)$, we came up with several procedures to select λ, such that we get $\mathrm{gFDR} \leq q$, if any two variables *from different groups* are nearly uncorrelated (Brzyski, Gossmann, et. al., 2016; Gossmann et. al., 2016).
- Under certain condition Group SLOPE enjoys some appealing estimation properties (asymptotically minimax, see Brzyski, Gossmann et. al., 2016).

SIMULATION WITH ORTHOGONAL GROUPS



$X \in \mathbb{R}^{5000\times5000}$; signal strength $\approx$ expected max. noise; 300 repitions at each sparsity level.

SIMULATION IN NON-ORTHOGONAL CASE



$X \in \mathbb{R}^{8915 \times 5976}$ contains real SNP data; 726 groups (mean size 8.23, median size 1); between-group corr. < 0.3; *simulated*

response $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, I)$.

GROUP SLOPE APPLICATION EXAMPLE

FRAMINGHAM COHORT

PREPROCESSING

- Exclusion of individuals or SNPs with more than 10% missing genotypes.
- Genotype imputation via IMPUTE2 based on the filtered data.
- The resulting preprocessed dataset consists of 8915 subjects' genotype data with 476907 annotated SNPs.
- Only 1771 subjects have corresponding spine BMD measurements.

FRAMINGHAM COHORT

CLUSTERING

- Only 1771 subjects have corresponding spine BMD measurements.
- We use the remaining over 7000 subjects to cluster the SNPs.
- Hierarchical clustering with an upper bound of 100 on cluster size, such that SNPs from different clusters have correlation < 0.3.

FRAMINGHAM COHORT

VARIABLE SCREENING (P-VALUE THRESHOLDING)

1. Obtain a p-value for each group of SPNs using an ordinary linear model and the F-test.
2. Retain only groups with p-value < 0.1.

$\Rightarrow$ Resulting X has dimensions 1771×117933, and consists of 6403 groups with average size equal to 18.42 (median size 2).

GROUP SLOPE RESULTS

- 40 SNPs were selected by Group SLOPE with target gFDR $q = 0.1$, and mapped to nearby genes.
- 15 genes have been found in previous studies to be associated with:
  - BMD (SMOC1, RPS6KA5, FGFR2, GAA, SCN1A, RAB5A, SOX1, and A2BP1),
  - osteoarthritis (A2BP1, ADAM12, MATN1),
  - lumbar disc herniation (KIAA1217),
  - osteopetrosis (VAV3),
  - biology of osteoclasts, osteoblasts and osteogenesis (VAV3, SLC7A7, ADAM12, PPARD, FGFR2, PTPRU, SMOC1).

GROUP SLOPE REFERENCES

1. Gossmann, A., Cao, S., & Wang, Y.-P. (2015). Identification of Significant Genetic Variants via SLOPE, and Its Extension to Group SLOPE. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, ACM BCB '15. DOI: 10.1145/2808719.2808743.
2. Gossmann, A., Cao., S., Brzyski, D., Zhao, L.-J., Deng, H.-W., & Wang, Y.-P. (2016). A sparse regression method for group-wise feature selection with false discovery rate control. *(Under review in IEEE/TCBB)*
3. Brzyski, D., Gossmann, A., Su, W., & Bogdan, M. (2016). Group SLOPE — adaptive selection of groups of predictors. arXiv:1610.04960. *(Under review in JASA)*
4. R packages:
    - cran.r-project.org/package=grpSLOPE
    - github.com/agisga/grpSLOPEMC

CONTROLLING FDR IN SPARSE CANONICAL CORRELATION ANALYSIS

SPARSE CANONICAL CORRELATION ANALYSIS

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

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 $x_1, \ldots, x_n \in \mathbb{R}^p$ be independent $\mathcal{N}(0, \Sigma_X)$,
$y_1, \ldots, y_n \in \mathbb{R}^q$ be independent $\mathcal{N}(0, \Sigma_Y)$,
$\mathrm{Cov}(x_k, y_k) = \Sigma_{XY} \in \mathbb{R}^{p \times q}$ for all $k \in \{1, \ldots, n\}$,
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

- Sparsity in the CCA solution can be achieved by utilizing penalty terms such as the ℓ_1-norm. Unique solution even when $p_X, p_Y \gg n$.
- Witten et. al. (2009):

$$\begin{aligned} &\text{maximize}_{u \in \mathbb{R}^p, v \in \mathbb{R}^q} \frac{1}{n} u^T X^T Y v, \\ &\text{subject to} \quad \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1, \\ &\text{and} \quad \|u\|_1 \leq c_1, \|v\|_1 \leq c_2. \end{aligned}$$

- *Selection of the sparsity parameters remains a challenging problem* (current options: cross-validation, AIC, permutation-based).
- Higher-order pairs of canonical vectors can be found by applying sparse CCA to a residual matrix, obtained from $X^T Y$ and the previously found canonical variates.

FDR-CORRECTION FOR SPARSE CCA

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

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

- 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 ar truly non-zero, test null hypotheses of the form

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

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

THE FDR-CORRECTED SPARSE CCA PROCEDURE

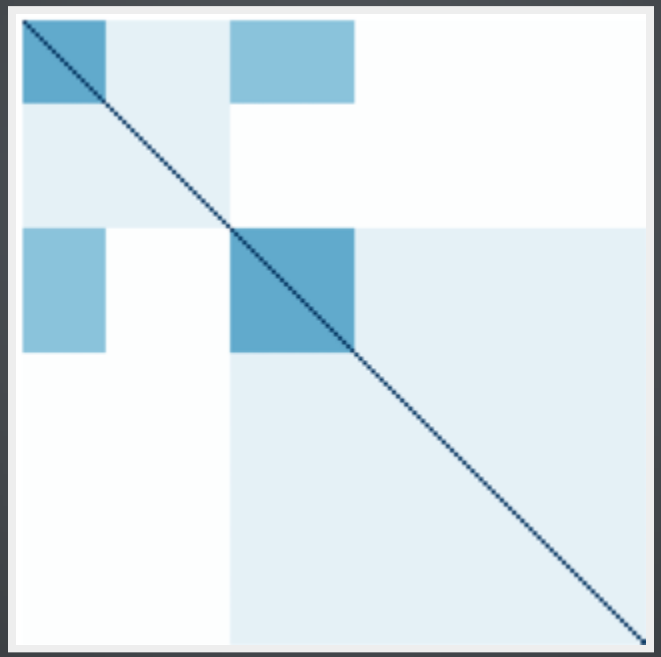
1. Divide each of X and Y into two subsets of sizes n_0 and n_1:
$$X = \begin{bmatrix} X^{(0)} \\ X^{(1)} \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Y^{(0)} \\ Y^{(1)} \end{bmatrix}.$$
2. Obtain preliminary sparse CCA estimates $\hat{u}^{(0)}$ and $\hat{v}^{(0)}$ on $X^{(0)}$ and $Y^{(0)}$. 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:
$$\hat{u}_i^{(1)} := \begin{cases} \left(X^T Y \hat{v}^{(0)}\right)_i, & \text{for any rejected } H_i^{(u)}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$
$$\hat{v}_j^{(1)} := \begin{cases} \left(Y^T X \hat{u}^{(0)}\right)_j, & \text{for any rejected } H_j^{(v)}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

SIMULATION RESULTS

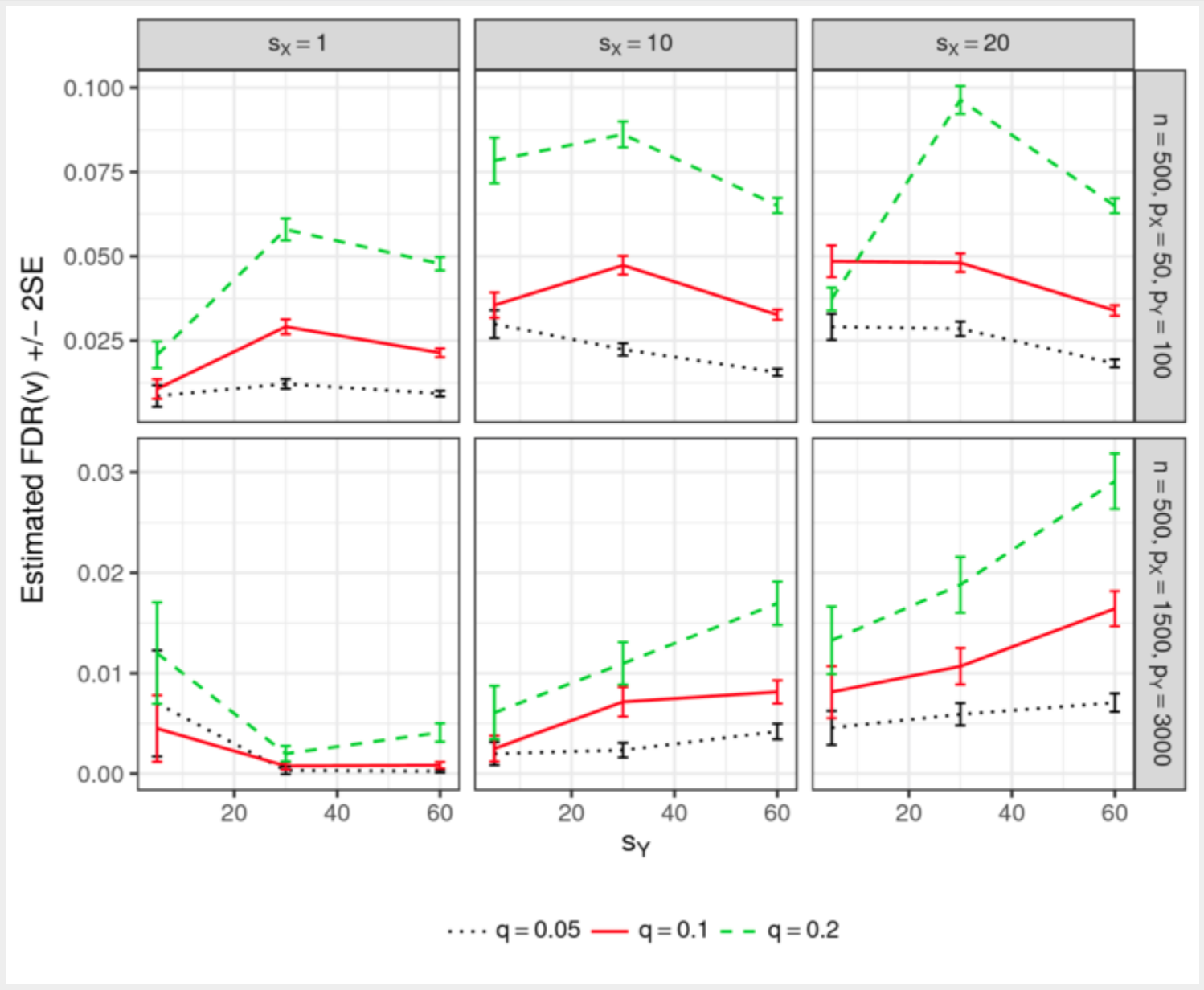
1. We show simulation results under **Gaussian scenarios**, in order to verify that the proposed procedure indeed controls the FDR under the assumptions that its derivation relies on.
2. We show simulation studies evaluating the performance on **non-Gaussian data**, which are generated based on real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

SIMULATION STUDY WITH GAUSSIAN DATA

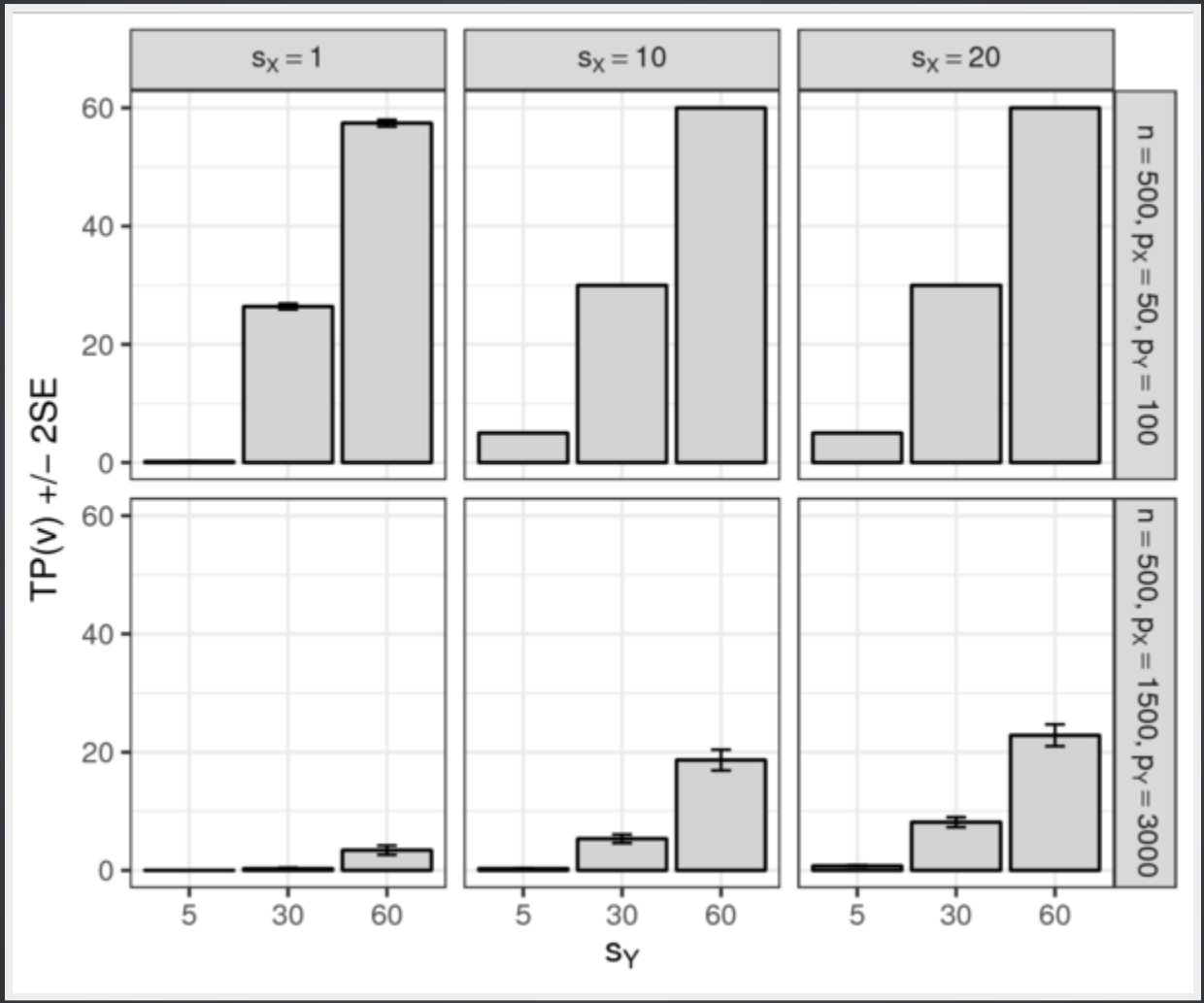
The data $\begin{bmatrix} X & Y \end{bmatrix}$ are generated from $\mathcal{N}(0, \Sigma)$, where Σ is blockwise constant.



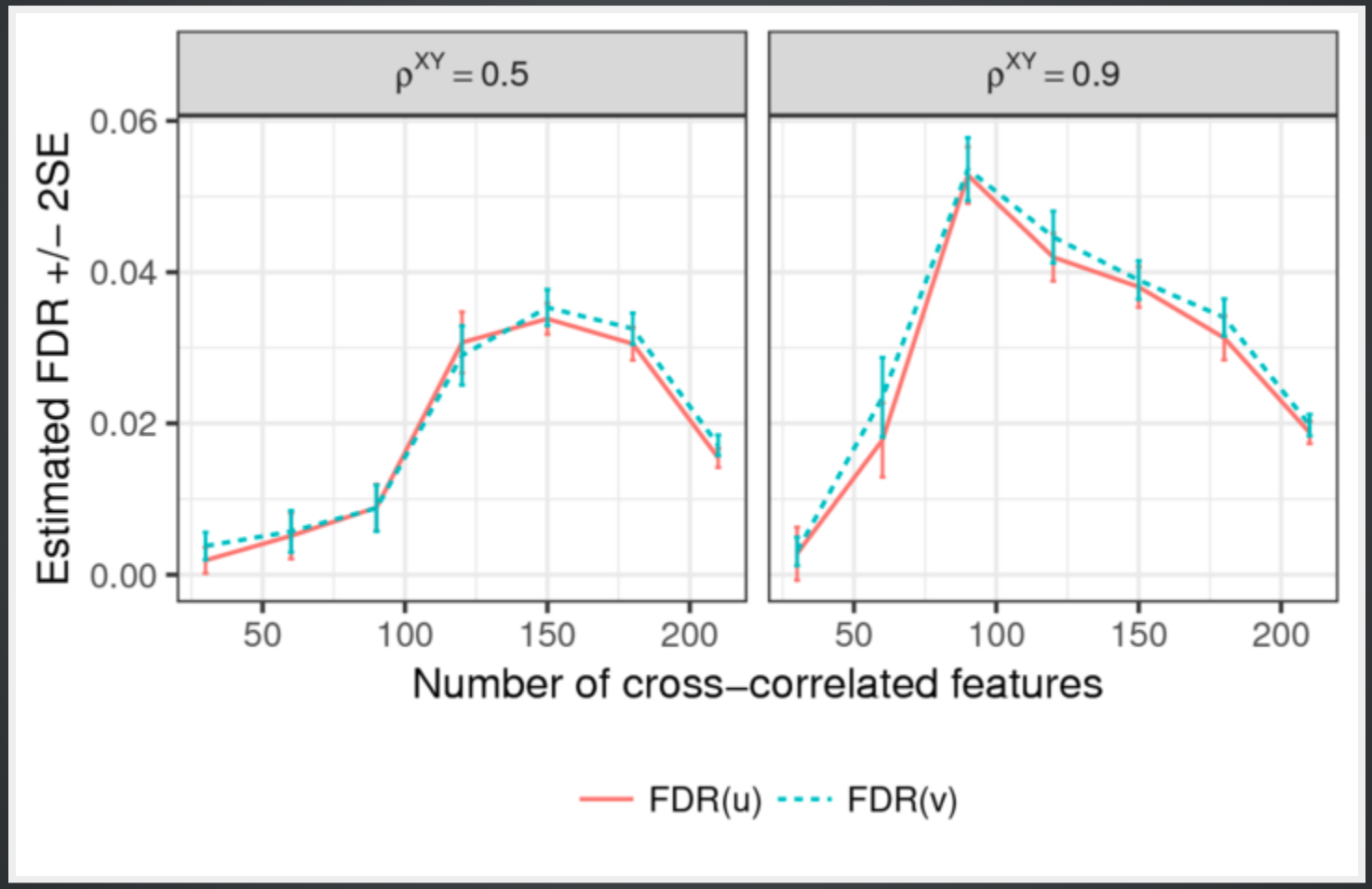
SIMULATION STUDY WITH GAUSSIAN DATA



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

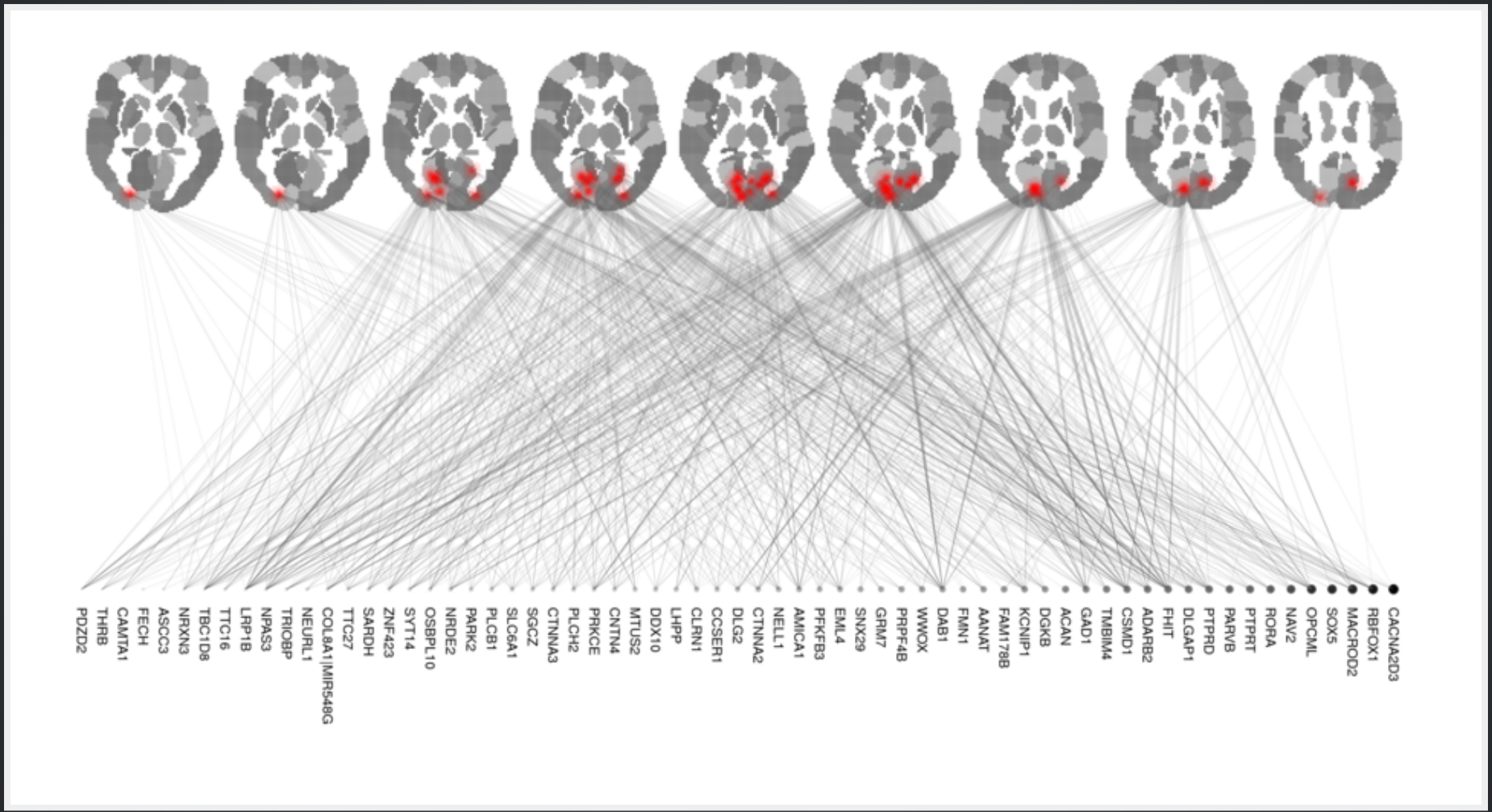
DATA

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.

FURTHER DEVELOPMENTS

- Addressing population stratification issues in the imaging genomics application example.
- Application to functional connectome data arising from different fMRI runs for the same set of subjects.
  - NB vs. EM vs. Rest.
  - Expecting results consistent with functional connectome individuality.
  - Functional network connectivity patterns can also be calculated from group spatial ICA time courses.

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

THE END

THANK YOU