SlopeCCA and gslopeCCA: sorted L1 penalized canonical correlation analysis

Alexej Gossmann and Yu-Ping Wang

Multiscale Bioimaging and Bioinformatics Laboratory, Tulane University, New Orleans, LA

Introduction

- Canonical correlation analysis (Hotelling, 1936), abbr. CCA, is a classical statistical technique, which can be used to make sense of the cross-correlation of two sets of measurements collected on the same set of samples (e.g., two genomic assays for the same set of cancer cells, or fMRI imaging and DNA sequencing data for the same mental illness patients).
- Sparse CCA (e.g., Parkhomenko et. al., 2009, or Witten et. al., 2009) extends the classical CCA to high-dimensional data by imposing a sparsity assumption on the CCA solution.
- Determination of the sparsity level, or the model tuning parameters, remains a challenging problem.
- We propose a definition of false discovery rate (FDR) for CCA, and (group) sorted ℓ_1 penalized CCA methods (slopeCCA and gslopeCCA). We show that slopeCCA and gslopeCCA are adaptive to the unknown sparsity level and keep the (group) FDR below a user-specified target threshold.

Assumed data structure

- Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independent $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_X)$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^q$ be independent $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_Y)$.
- Assume that $\mathrm{Cov}(\mathbf{x}_k, \mathbf{y}_k) = \boldsymbol{\Sigma}_{XY} \in \mathbb{R}^{p \times q}$ for all $k \in \{1, \ldots, n\}$, and that $\mathrm{Cov}(\mathbf{x}_k, \mathbf{y}_j) = \mathbf{0}$ whenever $k \neq j$.
- Define random matrices $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$

$$\boldsymbol{X} := \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \quad \boldsymbol{Y} := \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix}.$$

- We can think of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as two datasets containing respectively p and q features for the same n independent samples, where the features in the two datasets are cross-correlated with cross-covariance matrix $\boldsymbol{\Sigma}_{XY}$.

Canonical correlation analysis (CCA)

- CCA chooses $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^q$ to maximize the sample correlation between $\boldsymbol{X}\mathbf{u}$ and $\boldsymbol{Y}\mathbf{v}$; i.e., the CCA optimization problem is

$$\underset{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q}{\arg\max}\, \widehat{\mathrm{Cov}}(\boldsymbol{X}\mathbf{u}, \boldsymbol{Y}\mathbf{v}) = \underset{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q}{\arg\max}\, \frac{1}{n} \mathbf{u}^T \boldsymbol{X}^T \boldsymbol{Y} \mathbf{v},$$

  subject to

$$\widehat{\mathrm{Var}}(\boldsymbol{X}\mathbf{u}) = \frac{1}{n-1} \mathbf{u}^T \boldsymbol{X}^T \boldsymbol{X} \mathbf{u} = 1, \; \widehat{\mathrm{Var}}(\boldsymbol{Y}\mathbf{v}) = 1.$$

  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 $n \leq \max\{p, q\}$.

False discovery rate (FDR) for CCA

- For $i \in \{1, 2, \ldots, p\}$, we call the ith entry of the estimate $\hat{\mathbf{u}}$ a *false discovery*, if $\hat{u}_i \neq 0$ but $(\boldsymbol{\Sigma}_{XY})_{i,j} = 0$ for all $j \in \{1, 2, \ldots, q\}$.
- Let $\boldsymbol{V}_{\hat{\mathbf{u}}}$ denote the number of false discoveries in $\hat{\mathbf{u}}$.
- Let $\boldsymbol{R}_{\hat{\mathbf{u}}}$ be the total number of non-zero entries $\hat{\mathbf{u}}$.
- Analogously, define $\boldsymbol{R}_{\hat{\mathbf{v}}}$ and $\boldsymbol{V}_{\hat{\mathbf{v}}}$.

Definition (FDR)

Define the false discovery rate in $\mathbf{u}$ by

$$\mathbf{FDR}(\hat{\mathbf{u}}) := \mathrm{E}\left(\frac{\boldsymbol{V}_{\hat{\mathbf{u}}}}{\max\{\boldsymbol{R}_{\hat{\mathbf{u}}}, 1\}} \right),$$

and analogously define $\mathbf{FDR}(\hat{\mathbf{v}})$.

- $\mathbf{gFDR}$ is a generalization of $\mathbf{FDR}$ for the case when variables form groups (Brzyski et. al., 2016).

Sorted L1 norm (Bogdan et. al., 2015)

The sorted ℓ_1 norm in p dimensions is given by,

$$J_\lambda(\mathbf{u}) := \Sigma_{i=1}^p \lambda_i |u|_{(i)},$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ is a non-increasing regularizing sequence, and $|u|_{(1)} \geq |u|_{(2)} \geq \ldots \geq |u|_{(p)}$ is the order statistic of the magnitudes of the vector $\mathbf{u} \in \mathbb{R}^p$ (absolute values ranked in non-increasing order).

SlopeCCA and gslopeCCA

Inspired by SLOPE (Bogdan et. al., 2015) and Group SLOPE (Brzyski et. al., 2016, Gossmann et. al., 2015), we propose two novel sparse CCA methods with the goal of FDR and gFDR control.

Definition (slopeCCA: sorted L1 penalized CCA)

$$\mathbf{minimize}_{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q} \left\{ -\mathbf{u}^T \boldsymbol{X}^T \boldsymbol{Y} \mathbf{v} + \sqrt{n} J_{\lambda^u}(\mathbf{u}) + \sqrt{n} J_{\lambda^v}(\mathbf{v}) \right\},$$
$$\mathbf{subject\ to}\ \|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1.$$

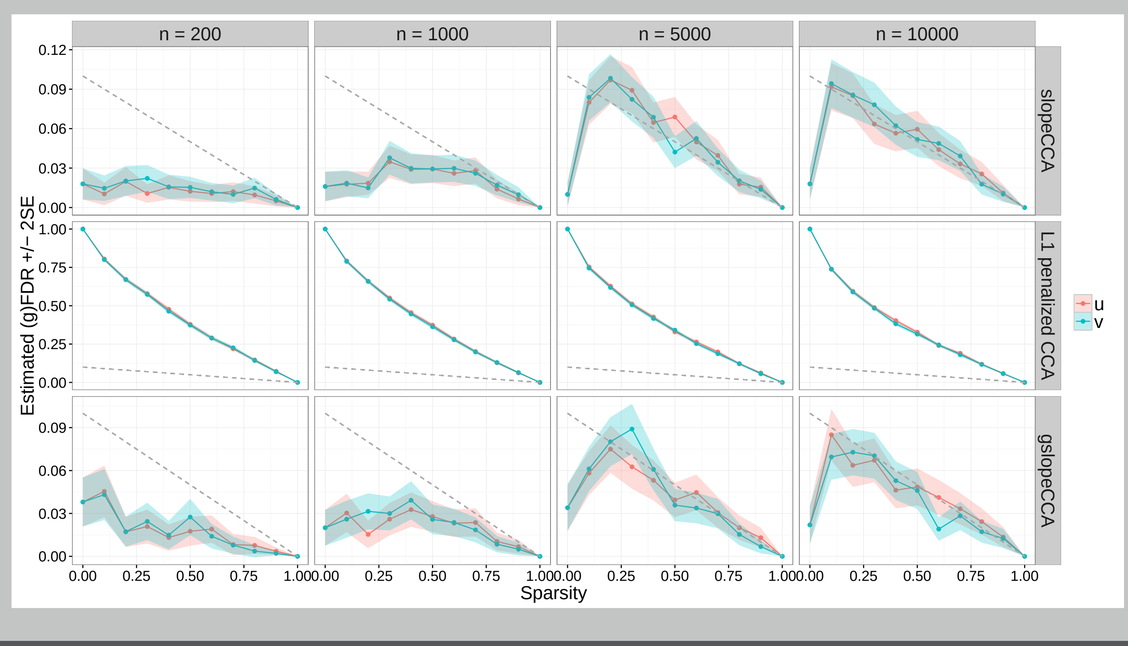
Definition (gslopeCCA: group sorted L1 penalized CCA)

$$\mathbf{minimize}_{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q} \Big\{ -\mathbf{u}^T \boldsymbol{X}^T \boldsymbol{Y} \mathbf{v} + \sqrt{n} J_{\lambda^u}\left((\|\mathbf{u}_1\|_2, \ldots, \|\mathbf{u}_{m_u}\|_2)^T \right) + \sqrt{n} J_{\lambda^v}\left((\|\mathbf{v}_1\|_2, \|\mathbf{v}_2\|_2, \ldots, \|\mathbf{v}_{m_v}\|_2)^T \right) \Big\},$$
$$\mathbf{subject\ to}\ \|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1.$$

Higher-order pairs of canonical vectors can be found by applying slopeCCA (or gslopeCCA) to a residual matrix, obtained from $\boldsymbol{X}^T\boldsymbol{Y}$ and the previously found canonical pairs. The slopeCCA and gslopeCCA optimization problems are biconvex and can be solved by alternating minimization algorithms.

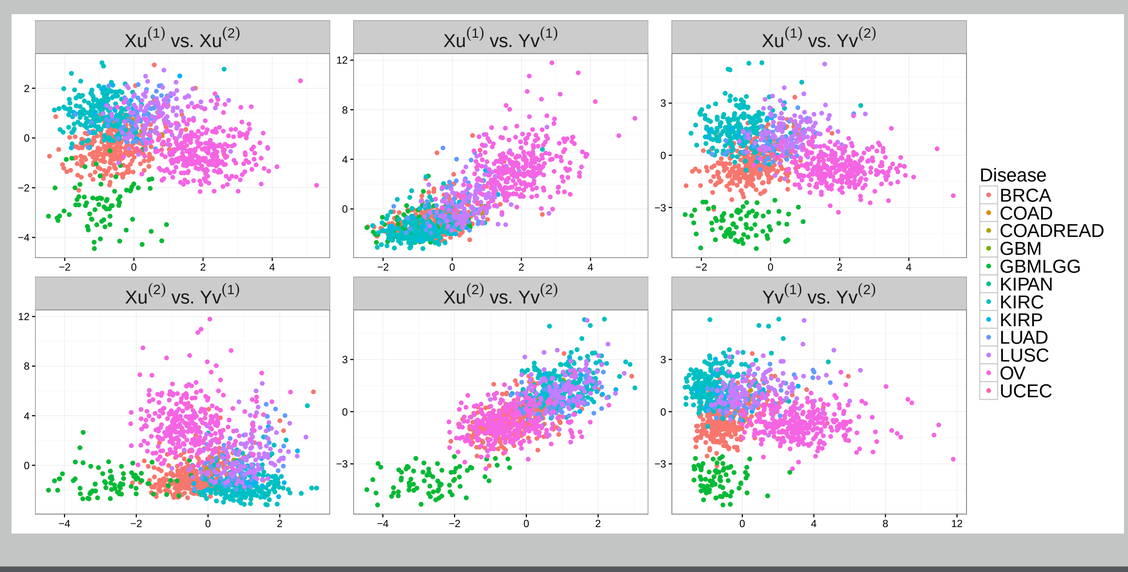
Asymptotic FDR properties of slopeCCA and gslopeCCA

- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$ are Gaussian with $p = q = 300$, varying n.
- $\boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_Y$ are *diagonal* (slopeCCA and PMA) or *block-diagonal* (gslopeCCA).
- 11 evenly spaced sparsity levels between 0 (i.e., $\boldsymbol{X}$ is uncorrelated with $\boldsymbol{Y}$) and 1 (i.e., every feature of $\boldsymbol{X}$ is correlated to some feature of $\boldsymbol{Y}$, and vice versa) were considered, where $(\boldsymbol{\Sigma}_{XY})_{ij} \in \{0, 0.5\}$.
- For comparison the sparse CCA of the widely used R package `PMA` (Witten et. al., 2009) is applied to the same data, with the tuning parameters determined by a permutation based approach (via `PMA::CCA.permute`).
- Dashed line represents the theoretical bound on FDR and gFDR for slopeCCA and gslopeCCA with target FDR of 0.1 (Theorem not shown).



Application to TCGA data

- We apply gslopeCCA to methylation and mRNA data for 12 diseases from The Cancer Genome Atlas (TCGA); $\boldsymbol{X} \in \mathbb{R}^{1436 \times 24981}$, $\boldsymbol{Y} \in \mathbb{R}^{1436 \times 20255}$.
- Canonical variates reveal differences between cancer types.
- A drawback: the obtained canonical vectors are not as sparse as desired, and therefore the results are difficult to interpret.



Acknowledgements

Our work is partially supported by NIH R01 GM109068, R01 MH104680, R01 MH107354.