A Simple Algorithm for Sparse Diagonal CCA

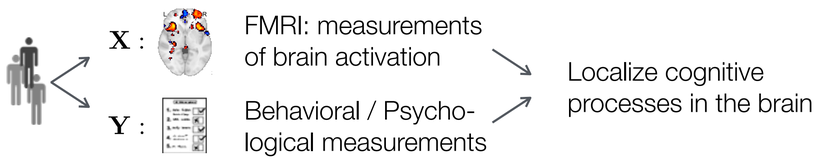
Megasthenis Asteris | Anastasios Kyrillidis | Oluwasanmi Koyejo | Russell Poldrack

The University of Texas at Austin — Stanford University

[The Problem]

Given *two sets of variables* derived from a *common set of samples*, find linear combinations of variables in each set, such that the induced canonical variables are maximally correlated.

- Cognitive neuroscience



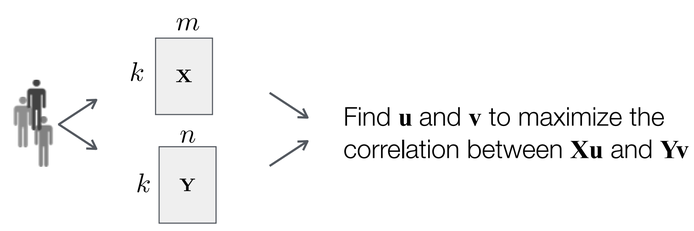
- Genetics / Molecular Biology

Input:

Two "views" of the same samples

Objective:

Find a pair of "canonical" vectors $\mathbf{u}, \mathbf{v}$ such that the projection of $\mathbf{X}$ on $\mathbf{u}$ and $\mathbf{Y}$ on $\mathbf{v}$ are maximally correlated.



As an optimization:

$$\max_{\mathbf{u},\ \mathbf{v}\neq 0} \frac{\mathbf{u}^\top \mathbf{\Sigma}_{\mathrm{XY}} \mathbf{v}}{\left(\mathbf{u}^\top \mathbf{\Sigma}_{\mathrm{XX}} \mathbf{u}\right)^{1/2} \left(\mathbf{v}^\top \mathbf{\Sigma}_{\mathrm{YY}} \mathbf{v}\right)^{1/2}}.$$

Here:

- **Sparse:** Extracted canonical vectors must be sparse; limited number of nonzero entries.
- **Diagonal:** Covariance matrices in each dataset treated as identity matrices. I.e., variables are standardized and uncorrelated.

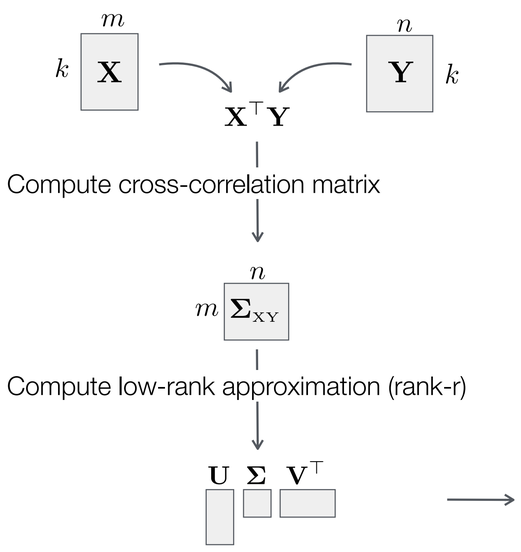
$$\max_{\mathbf{u},\ \mathbf{v}} \frac{\mathbf{u}^\top \mathbf{\Sigma}_{\mathrm{XY}} \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$

$$\mathbf{u} \in \mathbb{R}^m : \|\mathbf{u}\|_0 \leq s_x$$

$$\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\|_0 \leq s_y$$

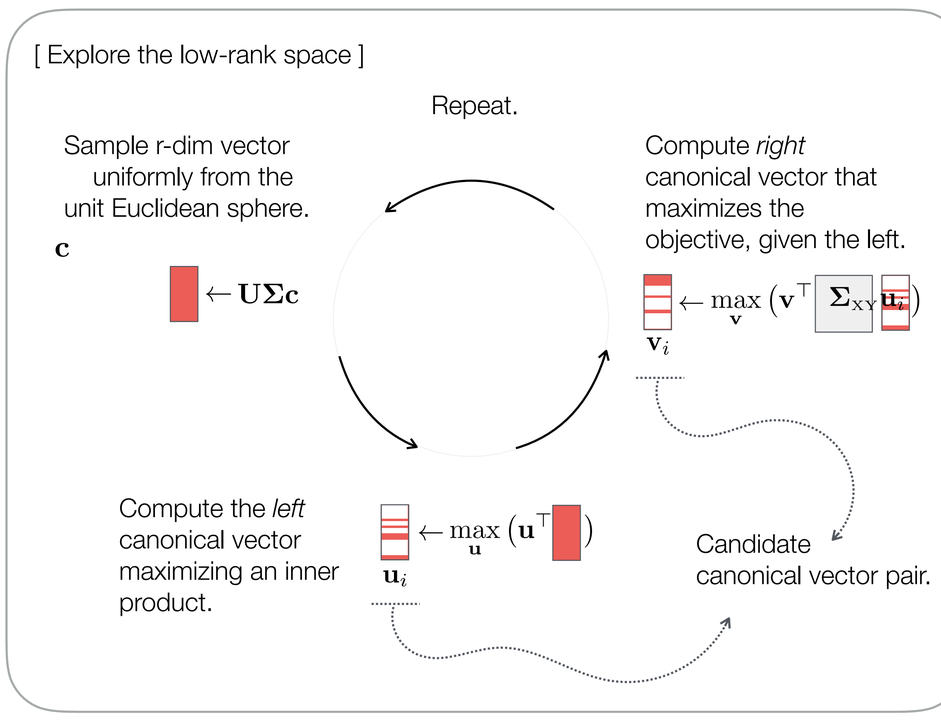
[Algorithm]

Input: Two datasets (views of the same set of observations.



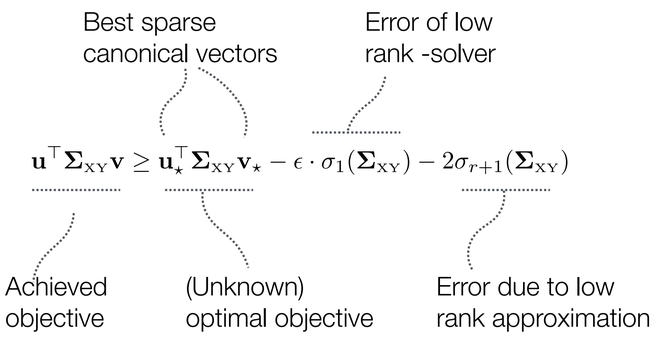
Output: The pair $\mathbf{u}_i, \mathbf{v}_i$ of sparse canonical vectors that maximizes the objective function.

[Explore the low-rank space]



[Guarantees]

Data dependent spectral guarantees on the output.



(Better guarantees if constraints are one-sided.)

Time complexity:

- Polynomial (almost linear) in dimensions m, and n.
- Polynomial in the target sparsity parameter.
- **Bottleneck:** Sampling
  The number of iterations T of the algorithm depends exponentially on the approximation rank r; higher approximation rank gives better results, but requires exponentially higher computational complexity.

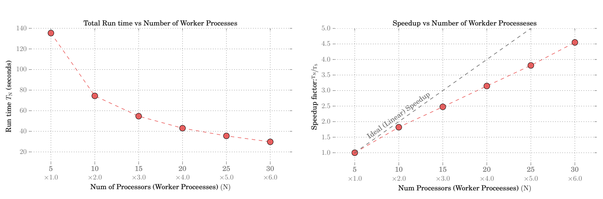
[Beyond Sparsity]

Sparsity only comes in the picture when generating the canonical pair; it is a 'projection' subroutine.

- We can apply any structure for which an 'efficient' projection step exists; e.g., group-sparsity, nonnegativity, etc.
- Theoretical guarantees extend to other types of constraints.

[In Practice]

- Configuration
  Disregarding theoretical guarantees, we can select an arbitrary configuration of the `accuracy` parameters (rank of approximation and number of iterations).
- Parallelization
  The main loop of the algorithm is trivially parallelizable: we can achieve an almost linear speedup in the number of available CPUs



[Experiments]

Breast cancer dataset (Chin et al., 2006): *Gene expression and DNA copy number measurements on a set of 89 tissue samples.*

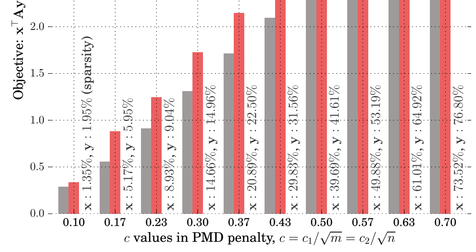
- X dataset: 89 × 2149 matrix (DNA) with CGH spots for each sample.
- Y dataset: 89 × 19672 matrix (RNA) of genes, along with information for the chromosomal locations of each CGH spot and each gene.

Objective:

Identify genes whose expression is correlated with a set of chromosomal gains or losses. Perform a quantitative comparison with (Witten et al., 2009); state-of-the art for sparse (diagonal) CCA.

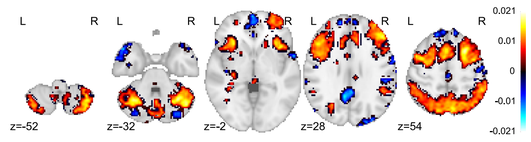
Output / Observations:

Higher objective (correlation) values achieved for several target sparsity values.



FMRI and behavioural measurements on 497 subjects available from the Human Connectome Project (HCP) (Van Essen et al., 2013).

- X dataset (497 × 65598): FMRI data upon standard preprocessing, general linear model analysis, resampling, and masking non-grey matter regions.
- Y dataset (497 × 38): scores from psychological tests, physiological measurements, and self reported behavior questionnaires.



Left component

Objective:

Investigate the shared co-variation between patterns of brain activity as measured by the experimental tasks, and behavioral variables

Behavioral Factor & Weight	
PMAT24_A_CR	0.487
PicVocab_AgeAdj	0.448
ReadEng_AgeAdj	0.440
PicVocab_Unadj	0.433
ReadEng_Unadj	0.426

Right component

Output / Observations:

Left: Identified fronto-parietal regions known to be involved in executive function and working memory and deactivation in the default mode areas, which is also associated with engagement of difficult cognitive functions.

Right: The behavioral variables identified to be positively correlated with the activation of this network are all related to various aspects of intelligence.