Stay on path: PCA along graph paths

Megasthenis Asteris, Anastasios Kyrillidis, Alex Dimakis, Han-Gyol Yi, Bharath Chandrasekaran

The University of Texas at Austin

[Sparse PCA]

Find direction of maximum variance (similar to vanilla PCA), but Extracted feature is sparse: a linear combination of a few variables.

Observe n samples in $\mathbb{R}^p$ and solve:

$$\max_{\mathbf{x}} \quad \mathbf{x}^\top \widehat{\mathbf{\Sigma}}\, \mathbf{x}$$

$$\text{subject to} \;\; \|\mathbf{x}\|_2 = 1, \|\mathbf{x}\|_0 = k$$

Empirical covariance matrix

NP-hard

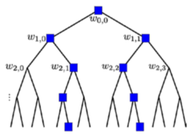
Why? [Engineer] Extracted feature is more interpretable.

[Statistician] Hope for recovery of "true" PC in high-dimensions.

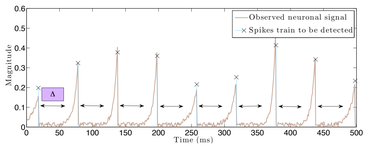
[More structure...]

Sparsity is Structure (a 0th order approximation)

What if we know more (e.g., group sparsity, tree structures,…)?



Wavelets of natural images: coefficients lie on a tree



Neuronal spike train with refractoriness period Δ

[Baraniuk et al., 2008; Kyrillidis et al., 2014, Friedman et al., 2010, …]

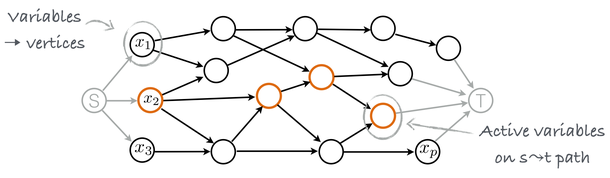
Structure in sparse PCA [Jenatton et al., 2010]

[(Sparse) PCA on graph path]

Idea: structure captured by underlying graph.

Here: structure = graph path

- Underlying directed acyclic graph G on p vertices.
- Desired PC supported on variables that lie along a path.



- Motivated by a neuroscience problem
- Bonus: *Multiple Choice* PCA
  Variables divided in multiple groups; one active variable per group.

NP-Hard, in general

Graph Path PCA

$$\max_{\mathbf{x}} \quad \mathbf{x}^\top \widehat{\mathbf{\Sigma}}\, \mathbf{x}$$

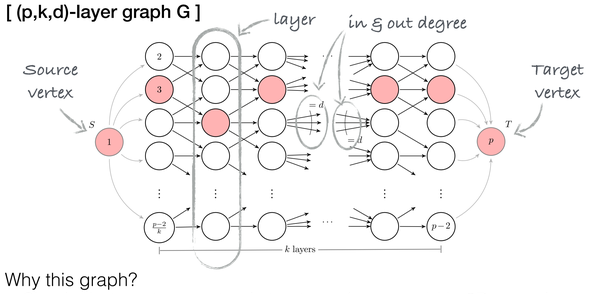
$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}(G)$$

$$\mathcal{X}(G) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_2 = 1, \text{supp}(\mathbf{x}) \subset \text{path of } G\}$$

known graph

[Data model]

[(p,k,d)-layer graph G]



Why this graph?

- Simple, but captures natural structures (e.g., a grid in 3D space).
- Non-trivial: the quadratic maximization is NP-hard.

[Spike along a path]

Data samples generated according to the *spiked covariance* model, but signal supported supported along a path of G.

$$\mathbf{y}_i = \sqrt{\beta} \cdot u_i \cdot \mathbf{x}_\star + \mathbf{z}_i$$

Observe n i.i.d. samples $\mathbf{\Sigma} = \mathbf{I}_p + \beta \cdot \mathbf{x}_\star \mathbf{x}_\star^\top$

$u_i \sim N(0,1)$ i.i.d

Gaussian noise (i.i.d)

SNR

Signal, supported on path of G.

[Results]

Theorem 1: *Lower Bound*

G : given (p,k,d)-layer graph. (known)

There exists signal $\mathbf{x}_\star$ supported on an st-path of G, such that:

If $\mathbf{y}_1, \dots, \mathbf{y}_n$ is a sequence of i.i.d. samples drawn from

$$\mathcal{D}_p(\mathbf{x}_\star) = \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p + \beta \cdot \mathbf{x}_\star \mathbf{x}_\star^\top\right)$$

Then, for any estimator $\widehat{\mathbf{x}}$

$$\mathbb{E}_{\mathcal{D}_p^n(\mathbf{x}_\star)} \left(\|\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathbf{x}_\star \mathbf{x}_\star^\top\|_F\right) \;\geq\; O\left(\sqrt{\tfrac{1+\beta}{\beta^2} \cdot \tfrac{1}{n}\left(\log \tfrac{p}{k} + k \log d\right)}\right)$$

Minimax errors bounded away from 0, unless $\Omega\left(\log \frac{p}{k} + k \log d\right)$.

Theorem 2: *Upper Bound*

G : given (p,k,d)-layer graph. (known)

$\mathbf{x}_\star$: signal support on st-path of G. (unknown)

Observe sequence $\mathbf{y}_1, \dots, \mathbf{y}_n$ of i.i.d. samples from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} \succeq \mathbf{0}$ with eigenvalues $\lambda_1 > \lambda_2 \geq \dots$ and principal eigenvector $\mathbf{x}_\star$.

Compute estimator $\widehat{\mathbf{x}}$ by solving the constrained quadratic maximization on the empirical covariance $\widehat{\mathbf{\Sigma}}$.

Then,

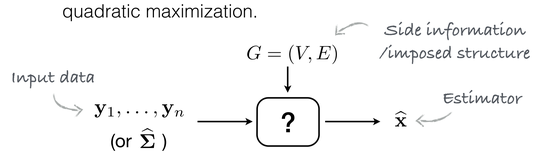
$$\mathbb{E}\left(\|\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathbf{x}_\star \mathbf{x}_\star^\top\|_F\right) \leq C \cdot \frac{\lambda_1}{\lambda_1 - \lambda_2} \cdot \frac{1}{n} \cdot \max\left\{\sqrt{nA},\, A\right\},$$

where $A = O\left(\log \frac{p}{k} + k \log d\right)$.

Compare with $O\left(k \log \frac{p}{k}\right)$ for the simple sparse PCA case

[Algorithms]

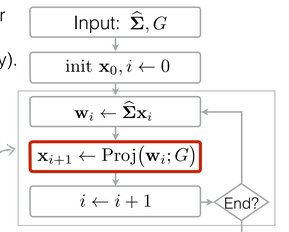
Goal: Approximate the solution the NP-Hard constrained quadratic maximization.



[Algorithm 1] *"A Power Method-based approach."*

Similar to Truncated Power Method for sparse PCA:

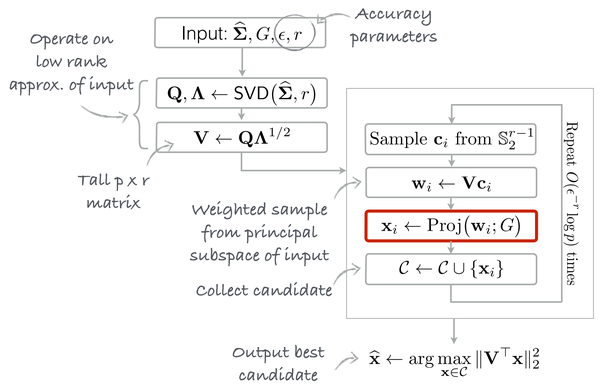
- Initialize an estimate (e.g., randomly).
- Multiply estimate with cov. matrix.
- Project on feasible set to get new estimate.
- Repeat.



Power Iteration with projection step.

[Algorithm 2] *"Subspace sample & project."*

- Compute a low-rank principal subspace of input data.
- Sample points from that subspace with appropriate weights.
- Project each sample on the feasible set, and get a candidate solution.
- Compare candidate solutions and output the best!



[Projection step]

- Project a p-dimensional vector on the feasible set; i.e., the set of unit-norm vectors supported along an st-path of G.
- Common step in both algorithms.

$$\text{Proj}(\mathbf{w}; G) = \arg\min_{\mathbf{x} \in \mathcal{X}(G)} \|\mathbf{x} - \mathbf{w}\|_2$$

Due to the constraints.

$$\arg\max_{\mathbf{x} \in \mathcal{X}(G)} \left(\mathbf{x}^\top \mathbf{w}\right)^2$$

Due to Cauchy-Schwarz

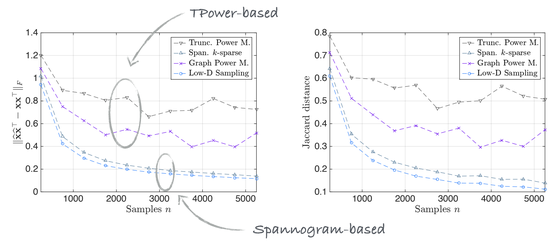
$$\arg\max_{\text{path } \pi} \sum_{i \text{ on } \pi} w_i^2$$

Longest (weighted) path problem on G, with special weights!

G acyclic; $O(p + |E|)$

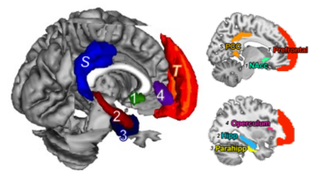
[Experiments]

[Synthetic] Data generated according to the (p,k,d)-layer graph model. *(p=1000, k=50, d=10 , 100 MC iterations)*



[Neuroscience Data] (Human Connectome Project)

- Single-session/single-participant resting state fMRI dataset.
- Variables: p = 111 Regions of Interest (HarvardOxford Atlas).
- Measurements: time series of n = 1200 points.



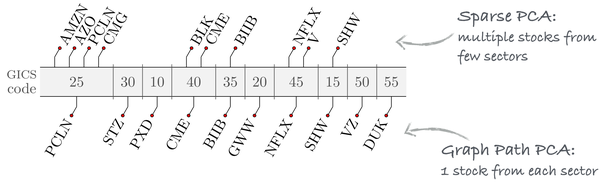
Construct Layered graph:

- Based on Euclidean dist. of ROI mass centers.
- Chose S (posterior cingulate cortex) and T (prefrontal cortex).
- Each layer is fully connected with its previous one.

[Multiple Choice PCA] One non-zero variable from each group.

Example: S&P 500 Index

- Variables: stocks, conceptually divided into 10 business sectors (GICS)
- Measurements: prices over a period of 1259 days (5 years)



Form (p,k,d) layer graph:

- Layers correspond to GICS sectors. Arbitrary order.
- All variables of a layer connected to all variables of next layer.

Compute PC along graph path!