Gradient-based Sparse Principal Component Analysis

Yixuan Qiu Jing Lei Kathryn Roeder

Department of Statistics and Data Science

Carnegie Mellon University

The 11th ICSA International Conference

Outline

Motivation

A New Efficient Sparse PCA Algorithm

Numerical Experiments and Application

Motivation

Overview of Sparse PCA

- Sparse PCA = PCA + Sparsity
- Factor loadings are sparse

Overview of Sparse PCA

- Sparse PCA = PCA + Sparsity
- Factor loadings are sparse
- Why sparse?
  - PCA may be inconsistent in high dimensions (Johnstone and Lu, 2009; Jung and Marron, 2009)
    - Sparsity $\Rightarrow$ denoising
  - Each principal component only depends on a small number of variables
    - Sparsity $\Rightarrow$ better interpretation

Sparse PCA Formulations

- Many different formulations
- Nonconvex objective functions
  - The lasso approach in PCA (Jolliffe, Trendafilov, and Uddin, 2003)
  - Regression-based (Zou, Hastie, and Tibshirani, 2006)
  - Penalized matrix decomposition (Witten, Tibshirani, and Hastie, 2009)
  - Generalized power method (Journée et al., 2010)
  - Iterative thresholding method (Shen and Huang, 2008; Ma, 2013)
  - ...
- Convex objective functions
  - DSPCA (d'Aspremont et al., 2005)
  - Fantope projection and selection (Vu et al., 2013)

Computation of Sparse PCA

- Nonconvex methods
  - Fast
  - Little global convergence guarantee
  - Heavily relies on model assumptions and initial values
- Convex methods
  - Global convergence
  - Weak assumptions
  - Slow

A New Efficient Sparse PCA Algorithm

Model Setting

- Convex formulation proposed by Vu et al. (2013)

$$\begin{aligned} \min_{X} \quad & -\operatorname{tr}(SX) + \lambda \|X\|_1 \\ \text{s.t.} \quad & O \preceq X \preceq I \text{ and } \operatorname{tr}(X) = d \end{aligned}$$

- $S_{p\times p}$: sample covariance matrix
- λ: sparsity parameter
- d: number of components
- $X_{p\times p}$: estimator of the projection matrix $\Pi = \Gamma\Gamma^{\mathrm{T}}$
- $\Gamma_{p\times d}$: factor loading matrix (eigenvectors)

Intuition

- Convex formulation

$$\begin{aligned} \max_{X} \quad & \operatorname{tr}(SX) - \lambda \|X\|_1 \\ \text{s.t.} \quad & O \preceq X \preceq I \text{ and } \operatorname{tr}(X) = d \end{aligned}$$

- Nonconvex counterpart

$$\begin{aligned} \max_{\Gamma} \quad & \operatorname{tr}(\Gamma^{\mathrm{T}} S \Gamma) - \lambda d \|\Gamma\|_{2,0}^2 \\ \text{s.t.} \quad & \Gamma^{\mathrm{T}} \Gamma = I_d \end{aligned}$$

- $\operatorname{tr}(\Gamma^{\mathrm{T}} S \Gamma) = \operatorname{tr}(S \Gamma \Gamma^{\mathrm{T}}) = \operatorname{tr}(SX)$: explained variance
- $\|\Gamma\|_{2,0}^2$: number of nonzero rows
- $X = \Gamma \Gamma^{\mathrm{T}}$ and $\Gamma^{\mathrm{T}} \Gamma = I_d$ share the same nonzero eigenvalues

Existing Computation Method

- ADMM algorithm

$$X_{k+1} = \mathcal{P}_{\mathcal{F}^d}(Y_k - U_k + \alpha S)$$
$$Y_{k+1} = \mathcal{S}_{\alpha\lambda}(X_{k+1} + U_k)$$
$$U_{k+1} = U_k + X_{k+1} - Y_{k+1}$$

- $\mathcal{S}_{\alpha\lambda}$: soft-thresholding operator, easy
- $\mathcal{P}_{\mathcal{F}^d}$: projection operator onto $\mathcal{F}^d = \{X : O \preceq X \preceq I \text{ and } \mathrm{tr}(X) = d\}$, hard
  - Requires a full eigen decomposition in each iteration
  - $\mathcal{O}(p^3)$ complexity

Optimization on Intersection of Convex Sets

- Let $f(X) = -\mathrm{tr}(SX) + \lambda\|X\|_1$, then the solution is

$$X_* = \arg\min_{X \in \mathcal{F}^d} f(X)$$

- A constrained problem on the intersection of convex sets $\mathcal{F}^d = C_1 \cap G_1 \cap G_2$, where
- $C_1 = \{X : \mathrm{tr}(X) = d\}$
- $G_1 = \{X : g_1(X) \leq 0\}$, $g_1(X) = \theta_{\max}(X) - 1$
- $G_2 = \{X : g_2(X) \leq 0\}$, $g_2(X) = -\theta_{\min}(X)$

A Projection-free Algorithm

- Let $\mathcal{L}(X) = f(X) + \mu\left(d_{C_1}(X) + r_1[g_1(X)]_+ + r_2[g_2(X)]_+\right)$
- $d_{C_1}(X)$: distance between X and C_1
- $[x]_+ = \max\{x, 0\}$
- An unconstrained problem $\min_{X\in\mathcal{X}} \mathcal{L}(X)$
- Projection onto $\mathcal{X} = \{X : \|X\|_F \leq \sqrt{d}\}$ is trivial

Theorem

If $\mu \geq (\|S\|_F + \lambda p + 1)\sqrt{p/(d+1)}$, $r_1 = \sqrt{d(d+1)}$, $r_2 = \sqrt{p(d+1)}$, *then* $\min_{X\in\mathcal{F}^d} f(X) = \min_{X\in\mathcal{X}} \mathcal{L}(X)$.

Convergence Analysis

- Many different algorithms to solve $\min_{X \in \mathcal{X}} \mathcal{L}(X)$
  - Subgradient descent
  - Proximal-proximal gradient method (Ryu and Yin, 2017)
- For the proximal-proximal gradient method, after T iterations,

$$\mathcal{L}(\hat{X}) \leq \min_{X \in \mathcal{X}} \mathcal{L}(X) + \frac{C}{T} \quad \text{and} \quad d_{\mathcal{F}^d}(\hat{X}) \leq \frac{C}{T},$$

  where C is some constant.

Statistical Property

- Assumptions
  - Sparsity: the factor loading matrix has at most s nonzero rows
  - Identifiability: the d-th eigengap $\delta_d = \theta_d(\Sigma) - \theta_{d+1}(\Sigma) > 0$
  - Sub-exponential elements:
    $\max_{i,j} P(|S_{ij} - \Sigma_{ij}| \geq u) \leq 2\exp(-4nu^2/\sigma^2)$ for all $u \leq \sigma$

Theorem
Take $\lambda = \sigma\sqrt{\log(p)/n}$, *and then with probability at least* $1 - 2/p^2$,

$$\|\hat{X} - \Pi\|_F \leq \frac{4\sigma s\sqrt{\log(p)}}{\delta_d\sqrt{n}} + \frac{\sqrt{2C/\delta_d}}{\sqrt{T}} + \frac{C}{T}.$$

- Interpretation: statistical error + optimization error + feasibility error

Extensions - Online Learning

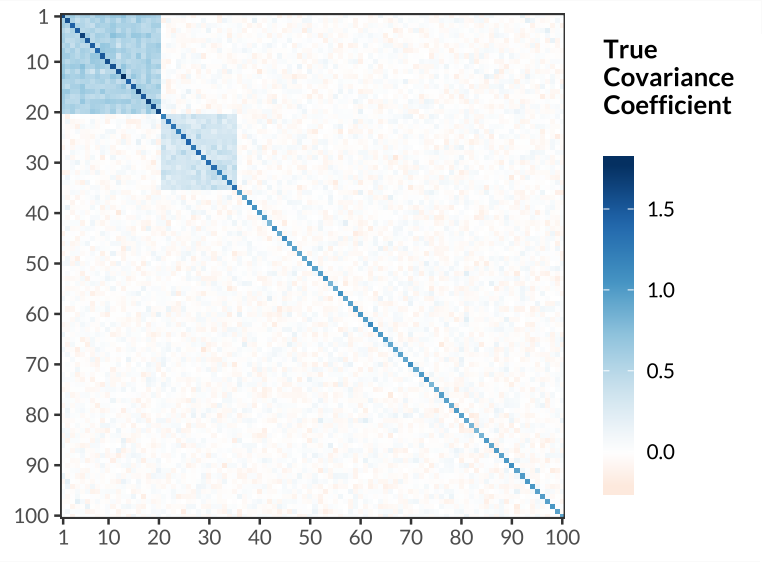
- Streaming data: $Z_1, Z_2, \ldots \in \mathbb{R}^p$ with $E(Z_t) = 0$ and $Cov(Z_t) = E(Z_t Z_t^{\mathrm{T}}) = \Sigma$, $t \geq 1$
- Time-t loss function: $\ell_t(X_t) = -Z_{t+1}^T X_t Z_{t+1} + \lambda \|X_t\|_{1,1} + \nu d_{\mathcal{F}^d}(X_t)$
- Total loss (regret): $\mathcal{R}(\{X_t\}, T) = \sum_{t=1}^T \ell_t(X_t) - \sum_{t=1}^T \ell_t(\Pi)$

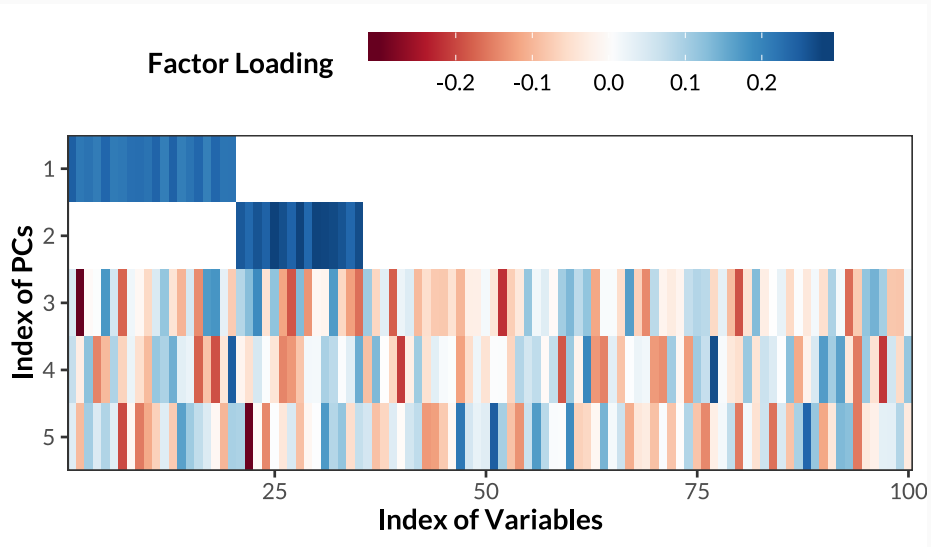
- Thanks to the gradient-based and projection-free formulation, many online learning algorithms can be applied here
  - Stochastic (sub)gradient descent
  - Generalized online mirror descent (Orabona et al., 2015)
- Details in arXiv: 1911.08048

Numerical Experiments and Application

Computational Efficiency

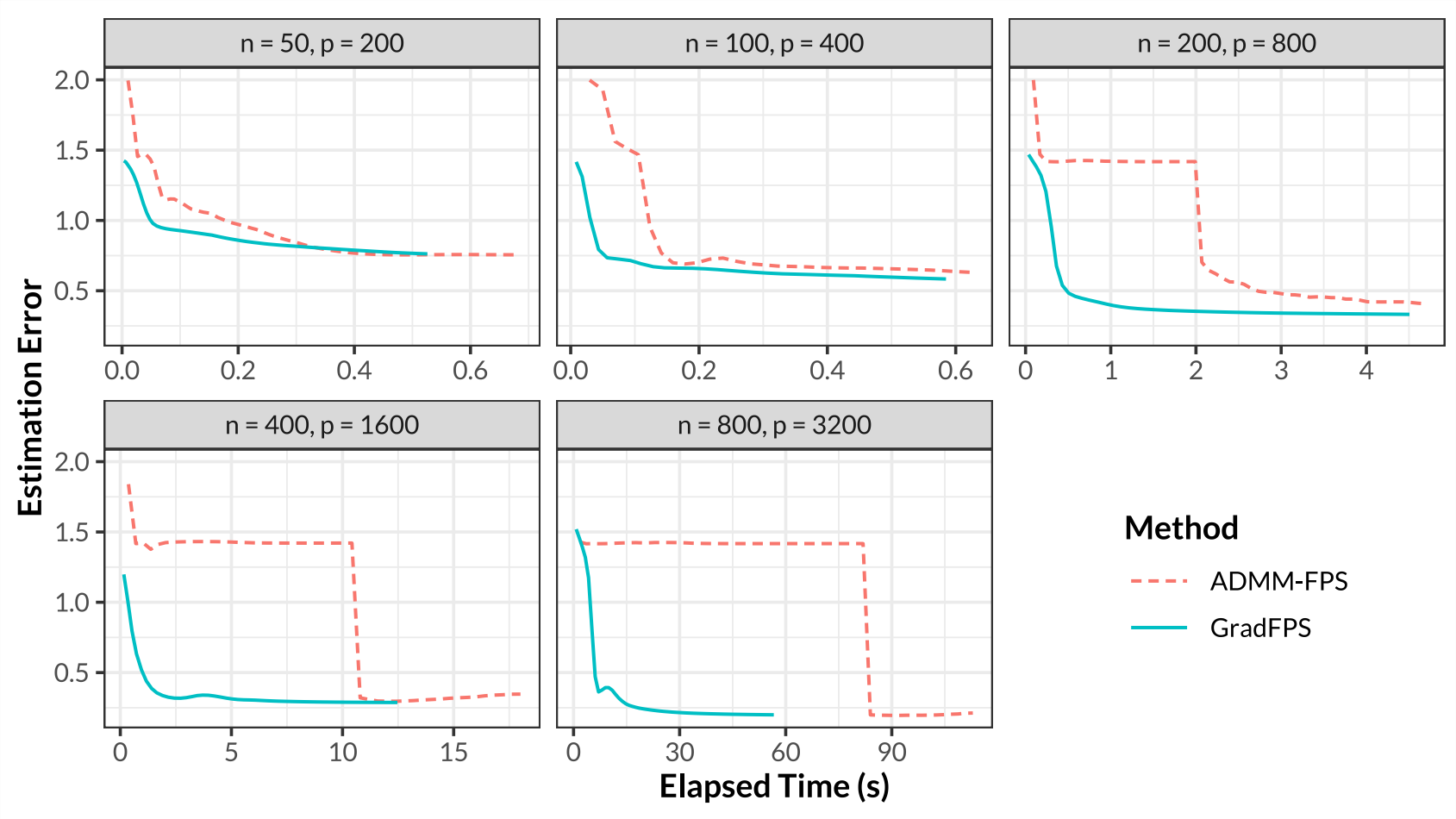
- Model setting





Computational Efficiency

- Comparing with the existing ADMM algorithm

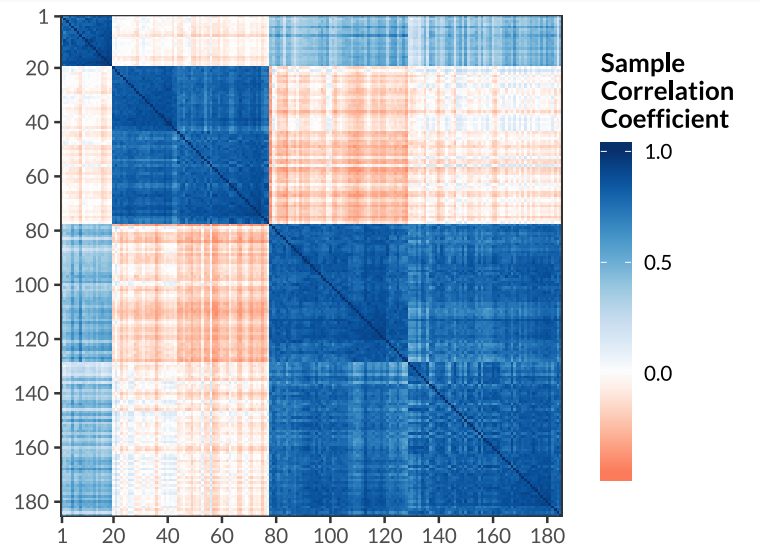


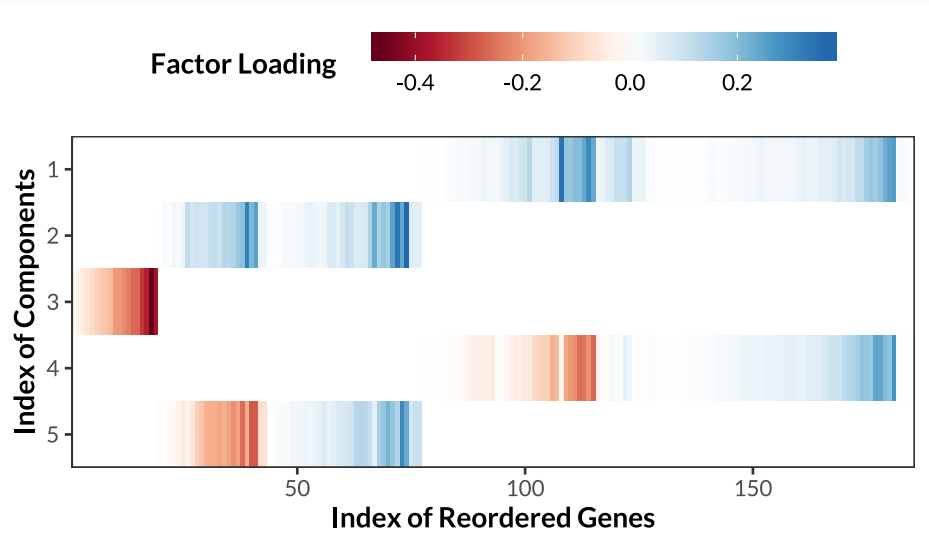
Gene Co-expression Network

- Brain gene expression data collected by the CommonMind Consortium (Fromer et al., 2016)
- To detect groups of genes such that genes in the same group have high mutual correlations
- $p = 16,423$ genes from 258 schizophrenia subjects and 279 control subjects
- Compute $d = 5$ sparse principal components
- Cluster selected genes into $k = 5$ groups based on the factor loadings

Gene Co-expression Network

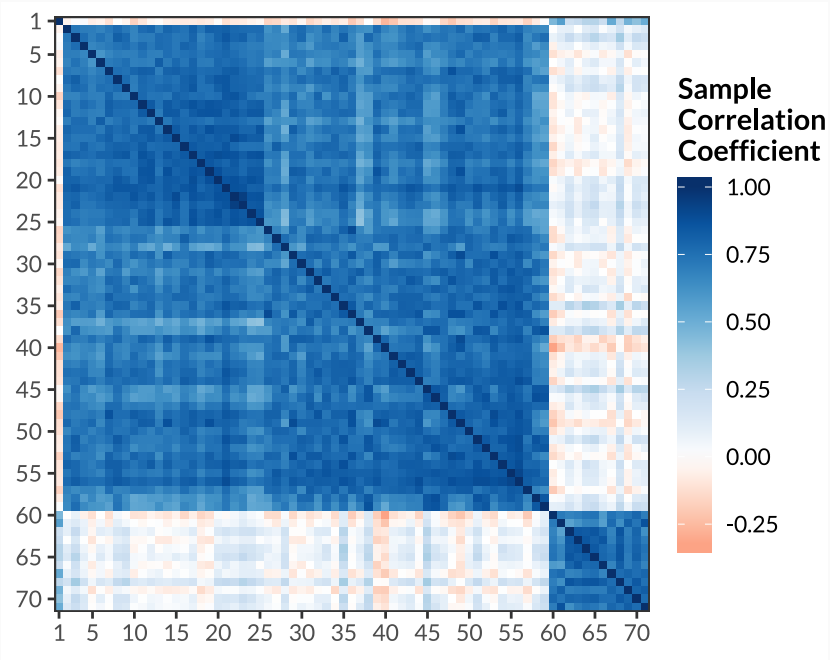
- Clustering result

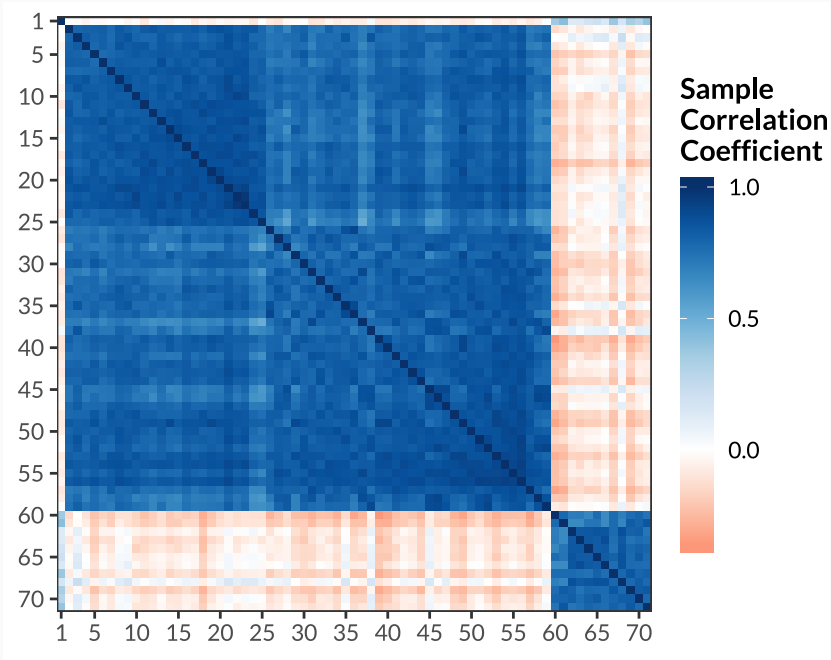




Gene Co-expression Network

- Differential analysis, control vs schizophrenia





THANK
YOU!