A new basis for sparse PCA

Fan Chen

Department of Statistics, UW-Madison

May 21, 2020
Preliminary Exam

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Background: Project road map

I study **social media** and **genomics** data.

Social media:

- Sampling of Twitter accounts (*JRSS-B*)
- Clustering of Twitter accounts (**This talk**)
- # of clusters & tweet analysis (Ongoing)

Genomics:

- Integrative analysis of transcriptome data (*Genome Biology*)
- Multivariate extensions (Ongoing)

Background: Clustering and PCA

- For clustering Twitter accounts, we took a **network-based** approach.
- This approach has been shown advantageous over some alternatives (Zhang, C, Rohe).
- Spectral clustering is a popular and effective method.
- It is essentially **principal component analysis** (PCA) + clustering.
- Can we get clusters directly from a sparse version of PCA?

Background: Sparse PCA in a nutshell

- Data matrix $X_{n \times p}$
- PCA finds the linear combination of columns, Xy, such that the most variance is kept,

$$\max_y \mathrm{Var}(Xy) \quad \text{s.t.} \quad \|y\|_2 = 1.$$

  Here, y contains the PC *loadings*.
- The elements in y are usually non-zero.
- *Sparse* PCA seeks “sparse” loadings.

Background: The plethora of available methods

A very short list of previous proposes:

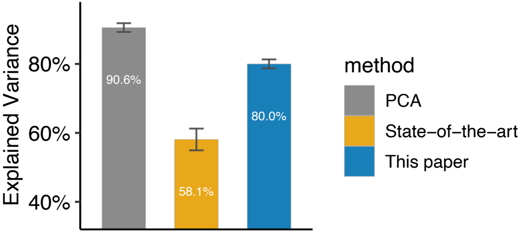
- the iconic regression-based approach (Zou '06)
- a convex relaxation via semidefinite programming (d'Aspremont '05)
- the penalized matrix decomposition framework (Witten '09)
- the generalized power method (Journée '10)

⋮

Theoretical developments are extensive, e.g., consistency, minimaxity, and statistical-computational trade-offs under **certain conditions**.

Background: An enigma of sparse PCA

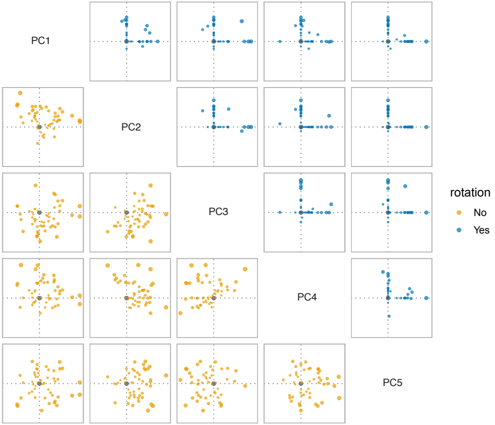
- Big loss of explained variance/information in the data.



- Better sparse loadings exist, if we use a new basis.

Background: An enigma of sparse PCA

- All assumes the eigen basis is sparse for high-dimensional data.
- But, PCs are rarely sparse.



- Unless we rotate them.

Method: A new formulation

- We propose to consider a **rotated basis** for sparse PCA.
- Consider the *matrix reconstruction error* minimization problems
  - Classic sparse PCA

$$\begin{aligned} \min \quad & \|X - Z\mathbf{D}Y^{\mathsf{T}}\|_{\mathsf{F}} \\ \text{s.t.} \quad & \|Y\|_1 \le \gamma \\ & Z^{\mathsf{T}}Z = Y^{\mathsf{T}}Y = I_k \\ & \mathbf{D} \text{ is diagonal} \end{aligned}$$

  - New sparse PCA

$$\begin{aligned} \min \quad & \|X - Z\mathbf{B}Y^{\mathsf{T}}\|_{\mathsf{F}} \\ \text{s.t.} \quad & \|Y\|_1 \le \gamma \\ & Z^{\mathsf{T}}Z = Y^{\mathsf{T}}Y = I_k \end{aligned}$$

- Does the middle $\mathbf{B}$ matrix allow orthogonal rotations on Y (or Z)?
- Yes! Suppose the SVD of $\mathbf{B}$ is $O\mathbf{D}R^{\mathsf{T}}$, then $ZBY^{\mathsf{T}} = (ZO)\mathbf{D}(YR)^{\mathsf{T}}$.

Method: Two interpretations of the formulation

Proposition (Orthogonal rotations can only help.)

If D is diagonal, then for any Z and Y,

$$\min \|X - Z\boldsymbol{B}Y^{\mathsf{T}}\|_{\mathsf{F}} \leq \min \|X - Z\boldsymbol{D}Y^{\mathsf{T}}\|_{\mathsf{F}}.$$

Proposition (A useful transformation for the algorithm.)

The new sparse PCA formulation is equivalent to a maximization problem,

$$\min \|X - Z\boldsymbol{B}Y^{\mathsf{T}}\|_{\mathsf{F}} \quad \Leftrightarrow \quad \max \|Z^{\mathsf{T}}XY\|_{\mathsf{F}}$$

subject to the same constraints and $\boldsymbol{B} = Z^{\mathsf{T}}XY$.

Algorithm: iteratively update Z and Y fixing one another.

Method: How to update Y fixing Z?

$$\max \quad \|Z^{\mathsf{T}} X Y\|_{\mathsf{F}} \quad \text{s.t.} \quad Y^{\mathsf{T}} Y = I_k, \ \|Y\|_1 \leq \gamma$$

1 First, consider only $Y^{\mathsf{T}} Y = I_k$.
One maximizer is the right singular vectors of $Z^{\mathsf{T}} X$ $\rightarrow \tilde{Y}$

2a The objective function is rotation **invariant**.
For any orthogonal matrix R, $\tilde{Y} R$ is also a maximizer.

2b Let's find the rotation that minimizes $\|\tilde{Y} R\|_1$. $\rightarrow Y^*$
(More on orthogonal rotations next up.)

3 Finally, consider the sparsity constraint, $\|Y\|_1 \leq \gamma$, and
"soft-threshold" the elements of Y^*. $\rightarrow \hat{Y}$

Method: Update Y fixing Z in three steps

Algorithm 1: Polar-Rotate-Shrink (PRS)

Input: matrix $A = X^\top Z$

Procedure $\texttt{PRS}(A)$**:**

$\tilde{Y} \leftarrow$ left singular vectors of A `// polar`

$Y^* \leftarrow$ rotate $\tilde{Y}$ with *varimax* † `// rotate`

$\hat{Y} \leftarrow$ soft-threshold Y^* `// shrink`

Output: $\hat{Y}$

†: Up next

Method: Why the varimax rotation?

Let $Y = \tilde{Y}R$ be the rotated matrix for some orthogonal R.

- $\|Y\|_1 = \sum_{i,j} |Y_{ij}|$ is *not* a smooth function of Y if it contains zero.
- Instead, minimize a smoother objective: $\|Y\|_{4/3}$
- Further, Hölder's inequality says that (with the conjugates $4/3$ and 4)

$$\|Y\|_{\frac{4}{3}} \geq \frac{\sqrt{k}}{\|Y\|_4}$$

  Hence, we maximize $\|Y\|_4 = \sum_{i=1}^{p} \sum_{j=1}^{k} y_{ij}^4$.
- When $Y^{\mathsf{T}} Y = I_k$, this is actually the varimax rotation (Kaiser '58). This technique has been popular in the psychology literature. In R, the base function `varimax` computes this.

Method: The SCA algorithm

Algorithm 2: Sparse Component Analysis (SCA)

Input: data matrix X and the number k of PCs

Procedure: $\texttt{SCA}(X, k)$:

initialize $\hat{Z}$ and $\hat{Y}$ with the top k singular vectors of X

repeat

$\hat{Z} \leftarrow \text{polar}(X\hat{Y})$

$\hat{Y} \leftarrow \texttt{PRS}(X^{\mathsf{T}}\hat{Z})$

until *convergence*

Output: sparse loadings $\hat{Y}$

Method: Two-way data analysis

- Sparse PCA reduces column dimensionality of X.
- The framework naturally generalizes to a two-way analysis for simultaneously row and column dimensionality reductions.
  - *Sparse matrix approximation* (SMA):

$$
\begin{array}{ll}
\min & \|X - Z\mathbf{B}Y^{\mathsf{T}}\|_{\mathsf{F}} \\
\text{s.t.} & \|Z\|_1 \le \gamma_z \\
 & \|Y\|_1 \le \gamma_y \\
 & Z^{\mathsf{T}}Z = Y^{\mathsf{T}}Y = I_k
\end{array}
$$

- For example, if X is the adjacency matrix of a bipartite graph, the SMA estimates the PCs for both sets of nodes.

Results: Overview

- Simulation studies:
  - explain more variance in the data (∗)
  - converge faster
  - more robust against the changes of parameters
- Data examples:
  - sparse coding of images
  - blind source separation
  - analysis of single-cell gene expression (∗)
  - clustering of Twitter accounts (∗)

∗ : this talk

Results: Capture more variance in the data

- Simulate data $X_{100\times100}$ from a low-rank model $SY^{\mathsf{T}} + E$, where
  - $S_{100\times16}$ contains the scores,
  - $Y_{100\times16}$ is sparse,
  - $E_{100\times100}$ is some noise.
- Impose the same ℓ_1-norm constraint on loadings.
- Assess the proportion of variance explained (PVE),

$$\|X_Y\|_{\mathrm{F}}^2, \quad \text{where} \quad X_Y = XY(Y^{\mathsf{T}}Y)^{-1}Y^{\mathsf{T}}.$$

Results: Capture more variance in the data

- SCA explains significantly more variance.

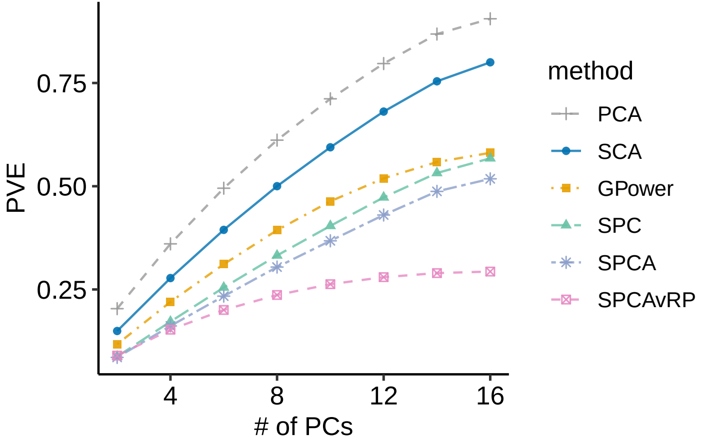


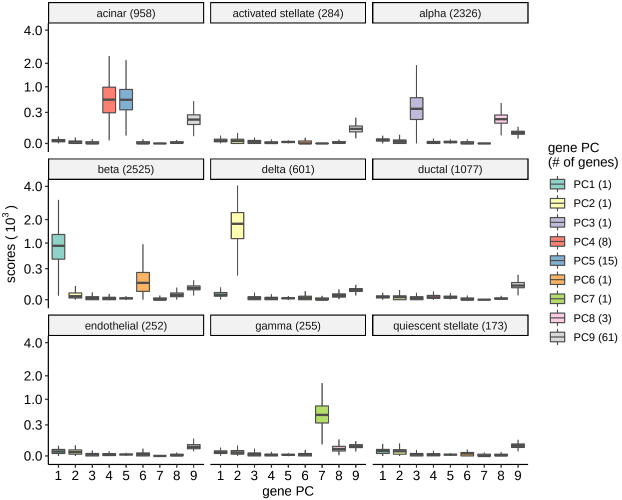
Figure: Comparison of the PVE by PCs.

Results: Analysis of scRNA-seq data

- scRNA-seq profiles the amount of gene expression for individual cells.
- For example, a human pancreatic islet cell data contains
  - $p = 17499$ genes
  - $n = 8451$ cells (with 9 cell types)
  - X_{ij} is the expression of gene j in sample i
- **Task**: extract the sparse gene PCs that characterize the cell types (without supervision).

Results: Analysis of scRNA-seq data

- SCA finds gene markers of cell types (PVE $= 94.34\%$).



Results: Clustering of Twitter accounts

- We collected a targeted sample of Twitter friendship network: (C, Zhang, Rohe, *JRSS-B* '20)
  - $n = 193,120$ Twitter accounts
  - who follow a total of $p = 1,310,051$ accounts
  - $A \in \{0,1\}^{n \times p}$ with $A_{ij} = 1$ if and only if account i follows account j.
- **Task**: find clusters of Twitter accounts.

Results: Clustering of Twitter accounts

- As a result, we observed that the clusters of Twitter accounts form homogeneous, connected, and stable social groups (Zhang, C, Rohe).
- The row and column clusters are matched.

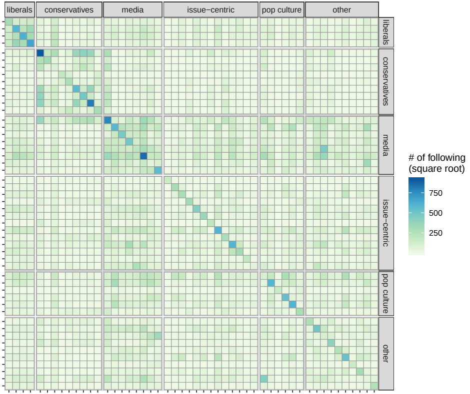


Figure: Heat map of friend counts between clusters of Twitter accounts.

Summary: This talk

- We introduce a new method for sparse PCA with an extension for two-way data analysis.
- The methods differ from the previous because of the orthogonally rotated basis.
- This approach is particularly useful when a data matrix is presumed low-rank but its singular vectors are **not** readily sparse.

Summary: Next talk...

1. Spectral technique for text (e.g., tweets) searching applications. (with Rohe)
2. Spectral method for estimating graph dimensionality (# of clusters). (with Roch & Rohe)
3. Sparse partial least square regression for transcriptome data analysis. (with Keleș)

Thank you!

Advisors	Karl Rohe and Sündüz Keleș
Committee	Po-Ling Loh, Michael Newton, Sébastien Roch (Math)

Sparse PCA and independent component analysis

- Similarities:
  - For sparse signals, $\text{SCA}^{\mathsf{T}} \approx \text{ICA}$.
  - Both are related to kurtosis (fourth-moment statistics).
- Nuances:
  - ICA also extracts non-sparse signals, while sparse PCA does not.
  - ICA presumes no or very little noise in X, in order for estimating guarantees.
  - Sparse PCA tackles high-dimensional regimes.

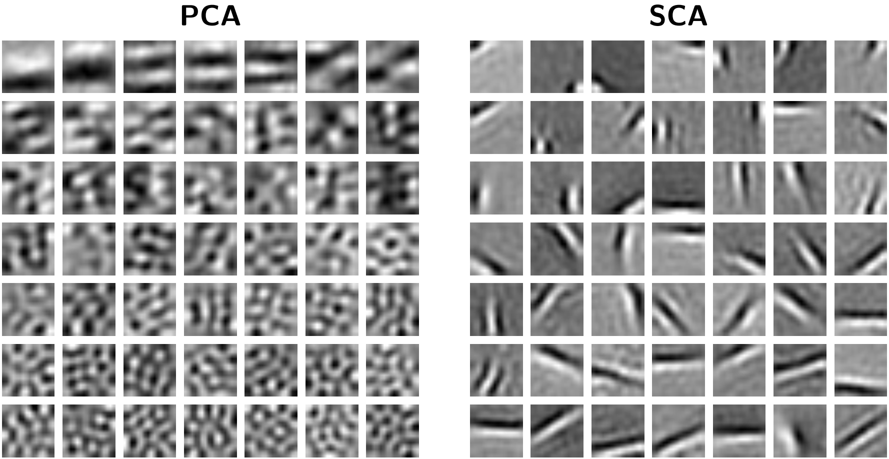
Results: SCA is more robust and stable

Figure: Heat maps of the sparse PC loadings returned by SCA and SPC, with three different sparsity parameters ($\gamma = 24, 36, 48$)

Results: SCA is capable of blind source separation

- Blind source separation finds source signal from mixing observational data.

Results: Sparse coding of images



Sparse image encoding using traditional PCA (left) and sparse PCA (right).