A New Basis for Sparse PCA

Fan Chen (fanci@google)

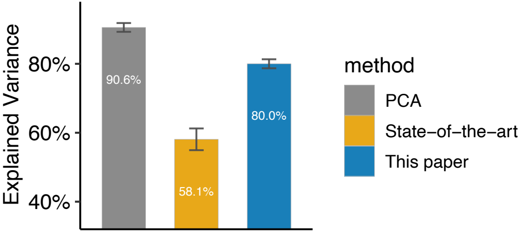
(with Karl Rohe)
Statistics @ UW-Madison

August 14, 2020
Statistics Journal Club, Google

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

In this talk, you will hear about:

- (Why) A number of sparse PCA methods perform poorly.



- A new basis of sparse PCA and a beautiful world (with examples).

We study public opinions on Twitter

Public opinions on Twitter — `murmuration.wisc.edu`

- Sampling of Twitter accounts (*JRSS-B*)
- Clustering of Twitter accounts (*submitted*, **This talk**)
- Networked public opinions (*submitted*)
- # of clusters & tweet analysis (ongoing...)

Sparse PCA in a nutshell

- Data matrix $X_{n \times p}$ (centered).
- PCA finds k linear combinations of columns, XY, such that the most variance is kept,

$$\max_{Y} \|XY\|_2 \quad \text{s.t.} \quad Y^T Y = I_k.$$

  Here, $Y \in \mathbb{R}^{p \times k}$ contains the PC *loadings*.
- The elements in Y are usually non-zero.
- Sparse PCA seeks "*sparse*" loadings.

The plethora of available methods

A very short list of previous proposes:

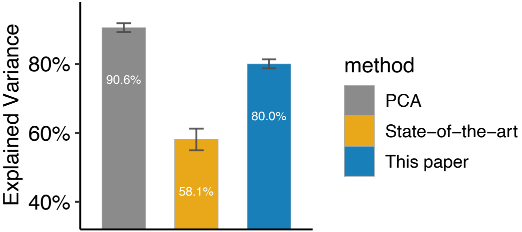
- the iconic regression-based approach (Zou '06)
- a convex relaxation via semidefinite programming (d'Aspremont '05)
- the penalized matrix decomposition framework (Witten '09)
- the generalized power method (Journée '10)

⋮

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

An enigma of sparse PCA

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



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

A stereotype formulation of sparse PCA

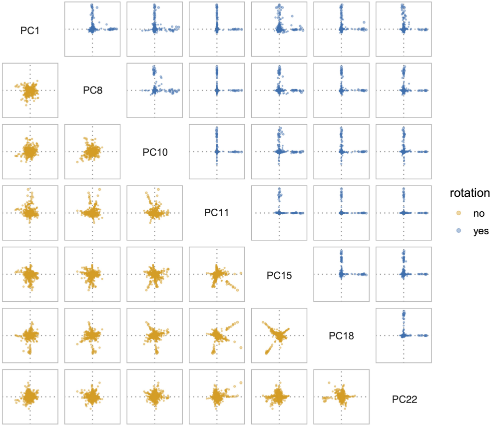
- Consider the *matrix reconstruction error* minimization problems
  - Classic sparse PCA

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

- **Implicative assumption**: The singular vectors were readily sparse.

Singular vectors are not readily sparse.

- But, PCs are rarely sparse in high-dimensional data.



- They can be sparse, if we rotate them.

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

Two interpretations of the formulation

Proposition (Orthogonal rotations can only help.)

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

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

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.

How to update Y fixing Z?

$$\max \quad \|Z^\top XY\|_\mathrm{F} \quad \text{s.t.} \quad Y^\top Y = I_k,\ \|Y\|_1 \le \gamma$$

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

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, $\|Y\|_1 \le \gamma$, and "soft-threshold" the elements of Y^*. $\rightarrow \hat{Y}$

Update Y fixing Z in three steps

Algorithm 1: Polar-Rotate-Shrink (PRS)

Input: matrix $A = X^\mathsf{T} Z$

Procedure 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: $\hat{Y}$

†: Invented by Kaiser (1958)

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.

Results: A beautiful world

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

$*$: this talk

Sparse coding of images

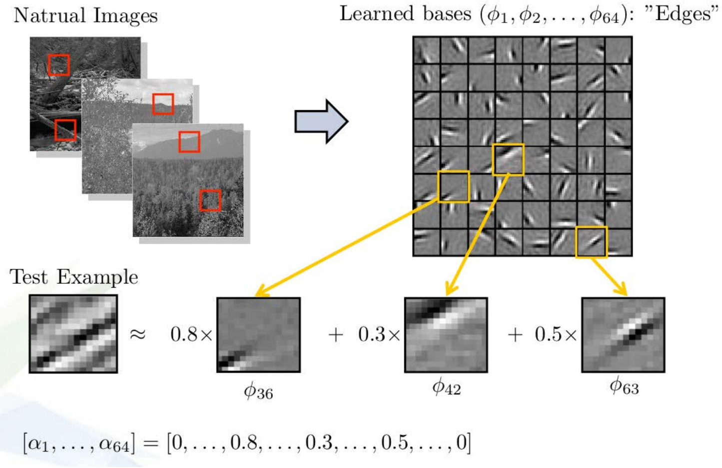
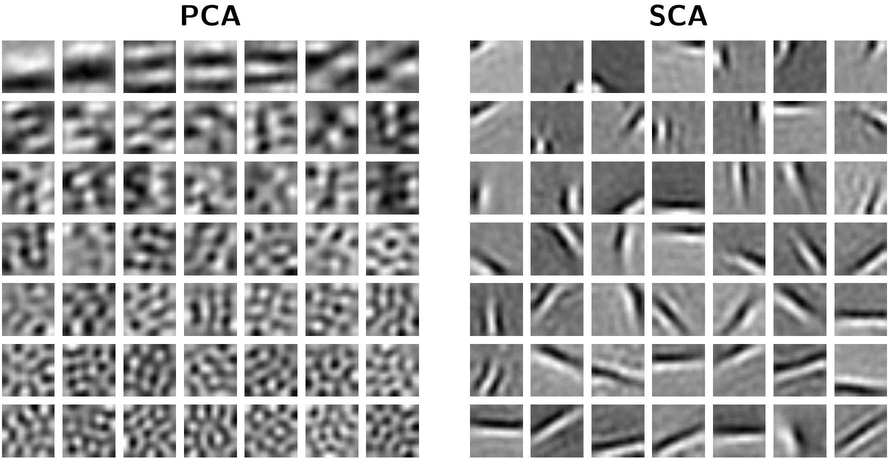


Figure by Brian Booth (2013).

- Can sparse PCA find these “Edges” too?

Sparse coding of images



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

Clustering of Twitter accounts: Setup

- **Prior work**: We collected a targeted sample of politics-related from Twitter accounts (C, Zhang, Rohe, *JRSS-B*, 2020)
- **Data**: Twitter friendship network
  - $n = 193,120$ Twitter accounts
  - $p = 1,310,051$ accounts being followed
- Adjacency matrix $A \in \{0,1\}^{n \times p}$ with

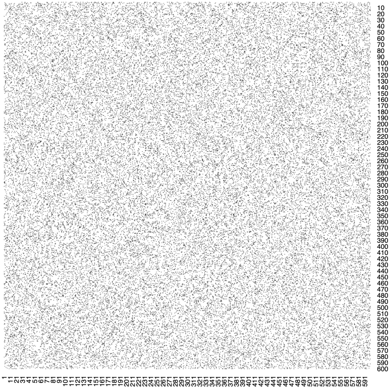
$$A_{ij} = 1, \text{ if } i \text{ follows } j$$

- **Task**: find k clusters of (n or p) Twitter accounts with A.

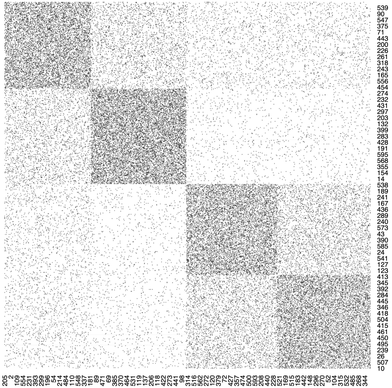
Clustering of Twitter accounts: Toy example

Example: 600 nodes and 4 clusters.

What we observe:



What we want:



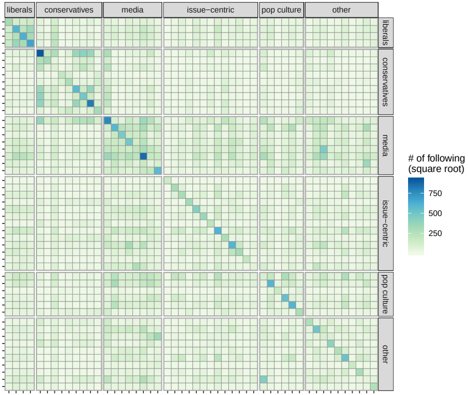
When we cluster rows and columns, we see blocks.

Clustering of Twitter accounts: Algorithm

- **Idea**: Treat the users being followed (i.e., columns of A) as variables.
- **Recall**: Loadings delineates PCs by original variables.
- **Solution**:
  1. Find k sparse PCs of A (or its normalized version).
  2. Cluster users with sparse PC loadings.

Clustering of Twitter accounts: Results

- As a result, we observed that the clusters of Twitter accounts form homogeneous, connected, and stable social groups (Zhang, C, Rohe).
- Recall: we want to see diagonal blocks.



- Enriched friendship within each clusters of Twitter accounts.

This talk

- Introduced a new method of finding **sparse** signals in data.
- The key advance is the orthogonal **rotation**.
- This approach is particularly **useful** when a data matrix is presumed low-rank but its singular vectors are not readily sparse.

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

    $\hat{Z} \leftarrow$ right singular vectors of $X\hat{Y}$

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

  until *convergence*

Output: sparse loadings $\hat{Y}$

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):

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

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

Discussion: Sparse PCA and ICA

- 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.

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),

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

Capture more variance in the data

- SCA explains significantly more variance.

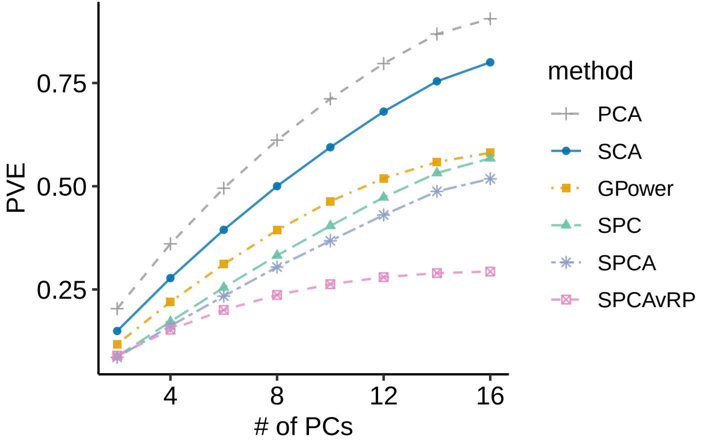


Figure: Comparison of the PVE by PCs.

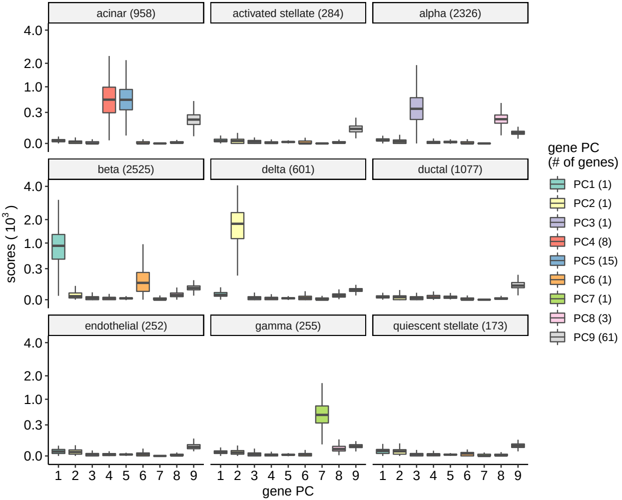
SCA is more robust and stable

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

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).

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



SCA is capable of blind source separation

- **Task**: Extract the source signals/images, only seeing the mixed ones.