Sparse PCA

Archer Yang

McGill University

October 3, 2024

- **Reading assignment:** Zou (2006); Zou (2018)

Precursor

Motivation of Sparse PCA

Pitprops Data: $N = 180$ and $p = 13$. SPCA generated very **sparse** loading structures *without losing much variance*.

	PCA			SPCA		
	PC1	PC2	PC3	PC1	PC2	PC3
topdiam	-.404	.218	-.207	-.477		
length	-.406	.186	-.235	-.476		
moist	-.124	.541	.141		.785	
testsg	-.173	.456	.352		.620	
ovensg	-.057	-.170	.481	.177		.640
ringtop	-.284	-.014	.475			.589
ringbut	-.400	-.190	.253	-.250		.492
bowmax	-.294	-.189	-.243	-.344	-.021	
bowdist	-.357	.017	-.208	-.416		
whorls	-.379	-.248	-.119	-.400		
clear	.011	.205	-.070			
knots	.115	.343	.092		.013	
diaknot	.113	.309	-.326			-.015
variance	32.4	18.3	14.4	28.0	14.0	13.3

SCoTLASSO

- Jolliffe, Trendafilov, and Uddin (2003) proposed **SCoTLASS**, which successively maximizes

$$
\max_{||\boldsymbol{b}_k||=1} \boldsymbol{b}_k^\top \boldsymbol{S} \boldsymbol{b}_k
$$

$$
\text{s.t.} \qquad \boldsymbol{b}_j^\top \boldsymbol{b}_k = 0, \quad h < k, \quad ||\boldsymbol{b}_k||_1 \leq t
$$

- Nonconvex optimization problem, high computational cost.

INSIGHT

PCA – equivalent form

- SPCA extends the reconstruction view of the PCA to derive sparse PCs. Recall that the first PC $\boldsymbol{v}_1$ can be defined as

$$\boldsymbol{v}_1 = \arg\min_{\boldsymbol{b}} \frac{1}{N} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{b}\boldsymbol{b}^\top \boldsymbol{x}_n \right\|^2 \quad \text{subject to } \|\boldsymbol{b}\|^2 = 1 \tag{1}$$

- We reformulate (1) as

$$\boldsymbol{v}_1 = \arg\min_{\boldsymbol{a},\boldsymbol{b}} \frac{1}{N} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{a}\boldsymbol{b}^\top \boldsymbol{x}_n \right\|^2 \quad \text{subject to } \|\boldsymbol{a}\|^2 = 1 \text{ anp } \boldsymbol{a} = \boldsymbol{b}$$

PCA – equivalent form

- The following theorem says that we can drop $\boldsymbol{a} = \boldsymbol{b}$ and still recover $\boldsymbol{v}_1$ exactly.

Theorem 1

For any $\lambda > 0$, let

$$(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \arg\min_{\boldsymbol{a},\boldsymbol{b}} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{a}\boldsymbol{b}^\top \boldsymbol{x}_n \right\|^2 + \lambda ||\boldsymbol{b}||^2$$

$$\text{subject to } ||\boldsymbol{a}||^2 = 1$$

Then $\hat{\boldsymbol{b}} \propto \boldsymbol{v}_1$

PCA – equivalent form

Theorem 2

Let $\widetilde{\boldsymbol{B}} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K)$

$$\widetilde{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_n \right\|^2$$
$$\text{subject to } \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_k$$

One can show that, for any $\lambda > 0$, let $\widehat{\boldsymbol{A}} = (\hat{\boldsymbol{a}}_1, \cdots, \hat{\boldsymbol{a}}_k) \in \mathbb{R}^{p \times k}$ and $\widehat{\boldsymbol{B}} = (\hat{\boldsymbol{b}}_1, \cdots, \hat{\boldsymbol{b}}_k)$

$$(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}) = \arg\min_{\boldsymbol{A}, \boldsymbol{B}} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{x}_n \right\|^2 + \lambda \sum_{j=1}^{k} \|\boldsymbol{b}_j\|^2$$
$$\text{subject to } \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_k$$

Then $\hat{\boldsymbol{a}}_j = \boldsymbol{v}_j$ and $\hat{\boldsymbol{b}}_j \propto \boldsymbol{v}_j$ for $j = 1, 2, \cdots, k$.

Proof

- Use the notation $\boldsymbol{A} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_k) \in \mathbb{R}^{p\times k}$ and $\boldsymbol{B} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_k) \in \mathbb{R}^{p\times k}$.
- Let $\boldsymbol{A}_\perp$ be any orthonormal matrix such that $\overline{\boldsymbol{A}} = [\boldsymbol{A}, \boldsymbol{A}_\perp]$ is $p \times p$ square orthonormal.
- For any matrix $\boldsymbol{X}$ and orthogonal matrices $\boldsymbol{A}$ and $\boldsymbol{A}_\perp$:

$$
\begin{aligned}
\|\boldsymbol{X}\boldsymbol{A}\|_F^2 + \|\boldsymbol{X}\boldsymbol{A}_\perp\|_F^2 &= \|\boldsymbol{X}\overline{\boldsymbol{A}}\|_F^2 = \mathrm{tr}(\overline{\boldsymbol{A}}^\top \boldsymbol{X}^\top \boldsymbol{X}\overline{\boldsymbol{A}}) \\
&= \mathrm{tr}(\overline{\boldsymbol{A}\boldsymbol{A}}^\top \boldsymbol{X}^\top \boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) \\
&= \|\boldsymbol{X}\|_F^2
\end{aligned}
$$

Proof

- We have

$$
\begin{aligned}
\sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{x}_n\|^2 &= \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\|_F^2 \\
&= \mathrm{tr}((\boldsymbol{X}^\top - \boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top)(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top)) \\
&= \mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X} - 2\boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top) \\
&= \|\boldsymbol{X}\boldsymbol{A}_\perp\|_F^2 + \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A} - 2\boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X} \\
&\qquad + \boldsymbol{B}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}) \\
&= \|\boldsymbol{X}\boldsymbol{A}_\perp\|_F^2 + \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A} - 2\boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X} \\
&\qquad + \boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}) \\
&= \|\boldsymbol{X}\boldsymbol{A}_\perp\|_F^2 + \|\boldsymbol{X}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{B}\|_F^2
\end{aligned}
$$

- $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X})$, since $\boldsymbol{A}\boldsymbol{B}^\top$ is a square matrix.

Proof

- Let
$$C_\lambda(\boldsymbol{A}, \boldsymbol{B}) = \sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{x}_n\|^2 + \lambda \sum_{j=1}^{k} \|\boldsymbol{b}_j\|^2$$

- Hence, with $\boldsymbol{A}$ fixed, solving

$$\arg\min_{\boldsymbol{B}} C_\lambda(\boldsymbol{A}, \boldsymbol{B})$$

is equivalent to solving the series of ridge regressions

$$\arg\min_{\boldsymbol{B}=(\boldsymbol{b}_j)_{j=1}^{k}} \sum_{j=1}^{k} \|\boldsymbol{X}\boldsymbol{a}_j - \boldsymbol{X}\boldsymbol{b}_j\|^2 + \lambda \|\boldsymbol{b}_j\|^2$$

Proof

- It is easy to show that

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{A}$$

- Then using Lemma 1

$$C_\lambda(\boldsymbol{A}, \widehat{\boldsymbol{B}}) = \mathrm{tr}((\boldsymbol{X}\boldsymbol{A})^\top (\boldsymbol{I} - \boldsymbol{S}_\lambda)(\boldsymbol{X}\boldsymbol{A}))$$

- Rearranging the terms, we get

$$C_\lambda(\boldsymbol{A}, \widehat{\boldsymbol{B}}) = \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) - \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{S}_\lambda \boldsymbol{X} \boldsymbol{A})$$

  which must be minimized with respect to $\boldsymbol{A}$ with $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$.

- This is equivalent to

$$\min_{\boldsymbol{A}} C_\lambda(\boldsymbol{A}, \boldsymbol{B}) \Longleftrightarrow \max_{\boldsymbol{A}} \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{S}_\lambda \boldsymbol{X} \boldsymbol{A})$$

Proof

- Hence $\boldsymbol{A}$ should be taken to be the largest k eigenvectors of $\boldsymbol{X}^\top \boldsymbol{S}_\lambda \boldsymbol{X}$.
- If the SVD of $\boldsymbol{X}$ is

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$$

- It is easy to show that $\boldsymbol{X}^\top \boldsymbol{S}_\lambda \boldsymbol{X}$ has eigen-decomposition

$$\boldsymbol{X}^\top \boldsymbol{S}_\lambda \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{D}^2\boldsymbol{V}^\top$$

  hence

$$\widehat{\boldsymbol{A}} = \boldsymbol{V}[,1:k] = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k)$$

Proof

- Now plugging $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ into

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}$$

$$= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}^2\boldsymbol{V}^\top\boldsymbol{I}_p[,1:k]$$

- Therefore $\widehat{\boldsymbol{B}} \propto \boldsymbol{V}[,1:k]$

Lemma 1

Consider the ridge regression criterion

$$C_\lambda(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||^2$$

Then if $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} C_\lambda(\boldsymbol{\beta})$

$$C_\lambda(\hat{\boldsymbol{\beta}}) = \boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{S}_\lambda)\boldsymbol{y}$$

where $\boldsymbol{S}_\lambda$ is the ridge operator

$$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top$$

See proof in Zou (2006)

Sparse PCA

Sparse PCA

The sparse PCA criterion for the first k sparse principal components is defined as

$$(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}) = \arg\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{x}_n \right\|^2 + \lambda \sum_{j=1}^{k} \|\boldsymbol{b}_j\|^2 + \lambda_1 \sum_{j=1}^{k} \|\boldsymbol{b}_j\|_1$$

subject to $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_k$

Computation of Sparse PCA

- If $\boldsymbol{A}$ is fixed, the optimization problem of $\boldsymbol{B}$ is

$$\boldsymbol{B}^{+} = \arg\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{B}^{\top}\boldsymbol{x}_n \right\|^2 + \lambda \sum_{j=1}^{k} \|\boldsymbol{b}_j\|^2 + \lambda_1 \sum_{j=1}^{k} \|\boldsymbol{b}_j\|_1$$

- For $\boldsymbol{B}^{+} = (\boldsymbol{b}_1^{+}, \cdots, \boldsymbol{b}_k^{+})$, this can be decomposed into

$$\boldsymbol{b}_j^{+} = \arg\min_{\boldsymbol{A},\boldsymbol{B}} \left\| \boldsymbol{X}\boldsymbol{a}_j - \boldsymbol{X}\boldsymbol{b}_j \right\|^2 + \lambda \sum_{j=1}^{k} \|\boldsymbol{b}_j\|^2 + \lambda_1 \sum_{j=1}^{k} \|\boldsymbol{b}_j\|_1$$

- The above is an elastic net penalized least square problem

Computation of Sparse PCA

- If $\boldsymbol{B}$ is fixed, the optimization problem of $\boldsymbol{A}$ is

$$\boldsymbol{A}^{+} = \arg\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{B}^{\top}\boldsymbol{x}_n \right\|^2$$

$$\text{subject to } \boldsymbol{A}^{\top}\boldsymbol{A} = \boldsymbol{I}_k$$

- We can see that when $k = p$, it is the Procrustes rotation problem.
- Analytical solution: compute the SVD $(\boldsymbol{X}^{\top}\boldsymbol{X})\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}$ and set

$$\boldsymbol{A}^{+} = \boldsymbol{U}\boldsymbol{V}^{\top}$$

Computation of Sparse PCA

- The SPCA algorithm iterates between the elastic net regression step and the SVD step till convergence.
- The output is the normalized $\boldsymbol{B}$ matrix for $j = 1, \cdots, k$

$$\boldsymbol{v}_j = \hat{\boldsymbol{b}}_j / ||\hat{\boldsymbol{b}}_j||$$

Example: Pitprops Data

Table 1. Pitprops Data: Loadings of the First Six Principal Components

Variable	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
topdiam	−0.404	0.218	−0.207	0.091	−0.083	0.120
length	−0.406	0.186	−0.235	0.103	−0.113	0.163
moist	−0.124	0.541	0.141	−0.078	0.350	−0.276
testsg	−0.173	0.456	0.352	−0.055	0.356	−0.054
ovensg	−0.057	−0.170	0.481	−0.049	0.176	0.626
ringtop	−0.284	−0.014	0.475	0.063	−0.316	0.052
ringbut	−0.400	−0.190	0.253	0.065	−0.215	0.003
bowmax	−0.294	−0.189	−0.243	−0.286	0.185	−0.055
bowdist	−0.357	0.017	−0.208	−0.097	−0.106	0.034
whorls	−0.379	−0.248	−0.119	0.205	0.156	−0.173
clear	0.011	0.205	−0.070	−0.804	−0.343	0.175
knots	0.115	0.343	0.092	0.301	−0.600	−0.170
diaknot	0.113	0.309	−0.326	0.303	0.080	0.626
Variance (%)	32.4	18.3	14.4	8.5	7.0	6.3
Cumulative variance (%)	32.4	50.7	65.1	73.6	80.6	86.9

Example: Pitprops Data

Table 2. Pitprops Data: Loadings of the First Six Modified PCs by SCoTLASS. Empty cells have zero loadings.

$t = 1.75$ *Variable*	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
topdiam	0.664			−0.025	0.002	−0.035
length	0.683	−0.001		−0.040	0.001	−0.018
moist		0.641	0.195		0.180	−0.030
testsg		0.701	0.001			−0.001
ovensg					−0.887	−0.056
ringtop		0.293	−0.186		−0.373	0.044
ringbut	0.001	0.107	−0.658		−0.051	0.064
bowmax	0.001			0.735	0.021	−0.168
bowdist	0.283					−0.001
whorls	0.113		−0.001	0.388	−0.017	0.320
clear						−0.923
knots		0.001		−0.554	0.016	0.004
diaknot			0.703	0.001	−0.197	0.080
Number of nonzero loadings	6	6	6	6	10	13
Variance (%)	19.6	16.0	13.1	13.1	9.2	9.0
Adjusted variance (%)	19.6	13.8	12.4	8.0	7.1	8.4
Cumulative adjusted variance (%)	19.6	33.4	45.8	53.8	60.9	69.3

Example: Pitprops Data

Table 3. Pitprops Data: Loadings of the First Six Sparse PCs by SPCA. Empty cells have zero loadings.

Variable	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
topdiam	−0.477					
length	−0.476					
moist		0.785				
testsg		0.620				
ovensg	0.177		0.640			
ringtop			0.589			
ringbut	−0.250		0.492			
bowmax	−0.344	−0.021				
bowdist	−0.416					
whorls	−0.400					
clear				−1		
knots		0.013			−1	
diaknot			−0.015			1
Number of nonzero loadings	7	4	4	1	1	1
Variance (%)	28.0	14.4	15.0	7.7	7.7	7.7
Adjusted variance (%)	28.0	14.0	13.3	7.4	6.8	6.2
Cumulative adjusted variance (%)	28.0	42.0	55.3	62.7	69.5	75.8

Example: Pitprops Data

- The SPCA algorithm iterates between the elastic net regression step and the SVD step till convergence.
- The output is the normalized $\boldsymbol{B}$ matrix for $j = 1, \cdots, k$

$$\boldsymbol{v}_j = \hat{\boldsymbol{b}}_j / ||\hat{\boldsymbol{b}}_j||$$

Example: Pitprops Data

- The SPCA algorithm iterates between the elastic net regression step and the SVD step till convergence.
- The output is the normalized $\boldsymbol{B}$ matrix for $j = 1, \cdots, k$

$$\boldsymbol{v}_j = \hat{\boldsymbol{b}}_j / ||\hat{\boldsymbol{b}}_j||$$

Example: Pitprops Data

- The SPCA algorithm iterates between the elastic net regression step and the SVD step till convergence.
- The output is the normalized $\boldsymbol{B}$ matrix for $j = 1, \cdots, k$

$$\boldsymbol{v}_j = \hat{\boldsymbol{b}}_j / ||\hat{\boldsymbol{b}}_j||$$