SPARSE PRINCIPAL COMPONENT ANALYSIS

Hui Zou, Trevor Hastie & Robert Tibshirani

STA315 - Advanced Statistical Learning
Presented by Mohammed Yusuf Shaikh

PRINCIPAL COMPONENT ANALYSIS

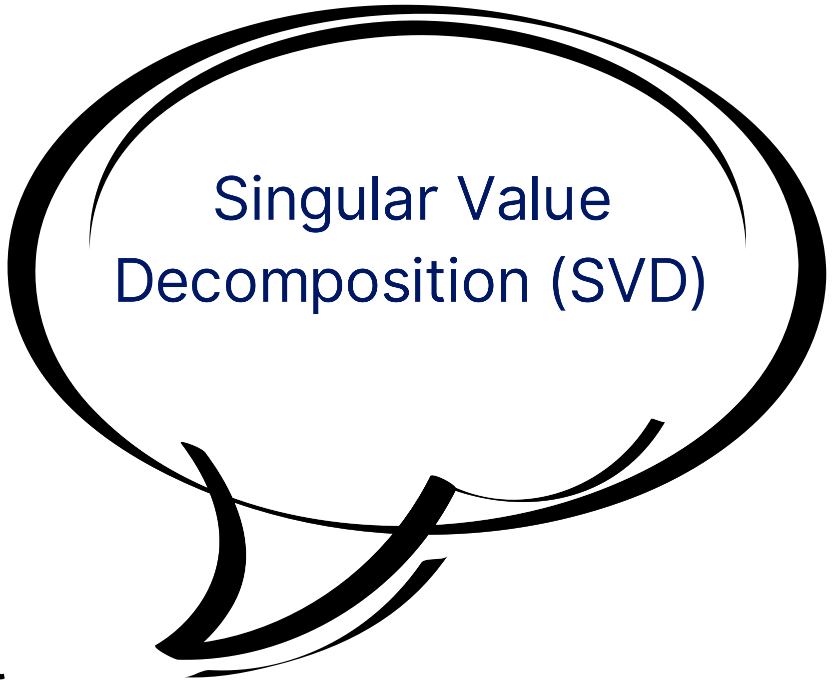
- Data Processing & Dimension Reduction
- **Goal:** Finding Orthogonal variables (Principal Component) that captures maximum variance
- Compressing data while retaining important information

IMAGE RECOGINTION

GENE EXPRESSION ANALYSIS

How to compute PCA?

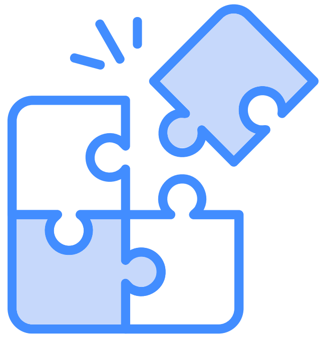
$$X = UDV^\top$$



Let $X \in \mathbb{R}^{n \times p}$ (centered: $\mathbb{E}[X] = 0$)

- **Z = UD** is the matrix of principal component scores. The data points in the reduced-dimensional space.
- **U** contain ***eigenvectors (directions)***
- **D** contains ***eigenvalues* (amount of variance captured by PC)**
- Columns of **V** are the loadings how much each original variable contributes to each principal component.

Limitations of PCA



- Lack of Sparsity - doesn’t do variable selection
- The loadings are usually nonzero, making it hard to interpret which variables that are truly important.
- Does not capture non-linear relationship
- Outliers can affect direction of principal components

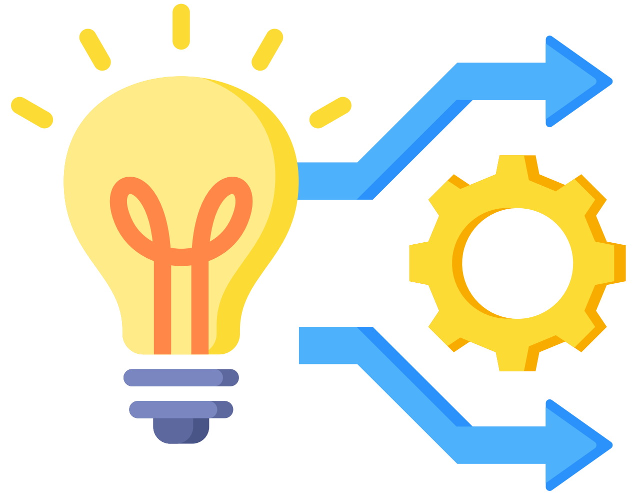
Overcoming PCA Limitation

Rotation Techniques

- Makes it easier to interpret each PC as being associated with a smaller subset of variables

Vines Method

- Restricting loadings to a small set of values like 0, 1, and −1 to simplify interpretation.

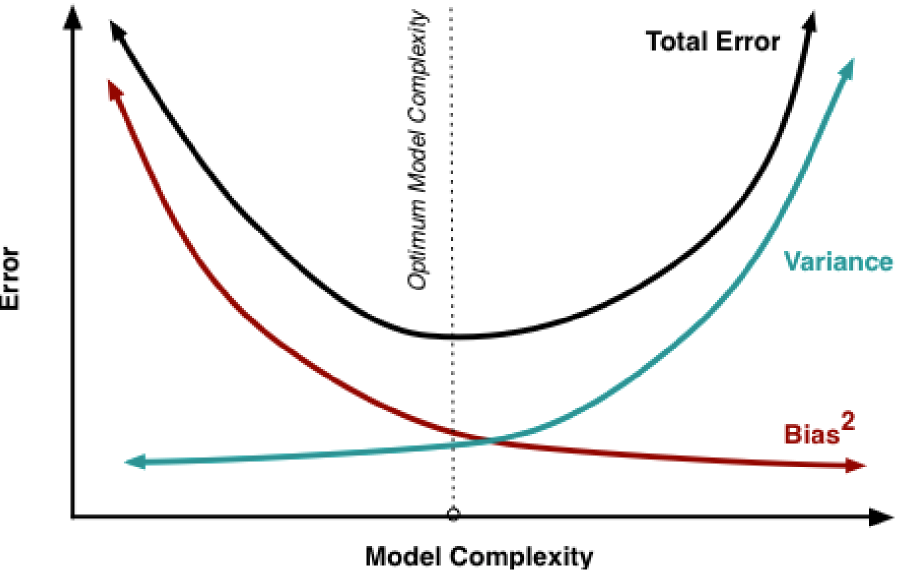


Thresholding Approach

- Set the loadings with absolute values smaller than a threshold to zero
- Can be potentially misleading in various respects

Lasso Regression: Shrinking Towards Sparsity

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$



Pros

- Selects the most important variables
- Sets less relevant coefficients to exactly zero, when λ is large
- Helps in model interpretability through the bias-variance trade-off
- Works well when n > p

Cons

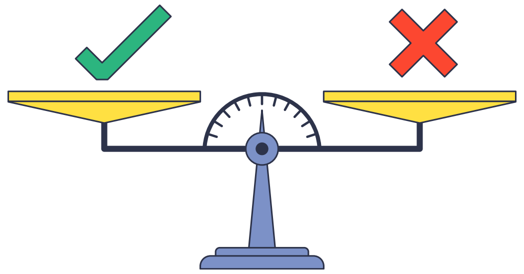
- When p > n, lasso can select at most n variables.

Elastic Net Regression

$$\hat{\beta}_{\text{en}} = (1 + \lambda_2) \left\{ \arg\min_{\beta} \|Y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\}.$$

- Elastic Net extends lasso to overcome its LASSO drawbacks while preserving all properties
- Is a combination of the ridge (L2) and lasso (L1) penalties.
- When $\lambda_2 = 0 \rightarrow$ Elastic Net = Lasso
- When $p > n$, using $\lambda_2 > 0$ allows more than n variables to be selected

Sparse Component Technique using LASSO



Pros

- An early attempt to introduce sparsity in PCA by imposing L1 constraint
- Tuning parameter ***t***, sufficiently small ***t*** forces some loading to be exactly 0

Cons

- There is no clear method to determine an optimal t, requiring multiple trial values.
- High computational cost
- Fails to produce sufficiently sparse loadings while maintaining a high percentage of explained variance.

From PCA to Sparse PCA via Regression

Theorem 1: PCA as Ridge Regression

- $Z_i = U_i D_{ii}$: the i-th principal component
- Ridge Regression formulation:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \|Z_i - X\beta\|^2 + \lambda\|\beta\|^2$$

Where, $\hat{v} = \dfrac{\hat{\beta}_{ridge}}{\|\hat{\beta}_{ridge}\|} = V_i$

Why Ridge is Needed?

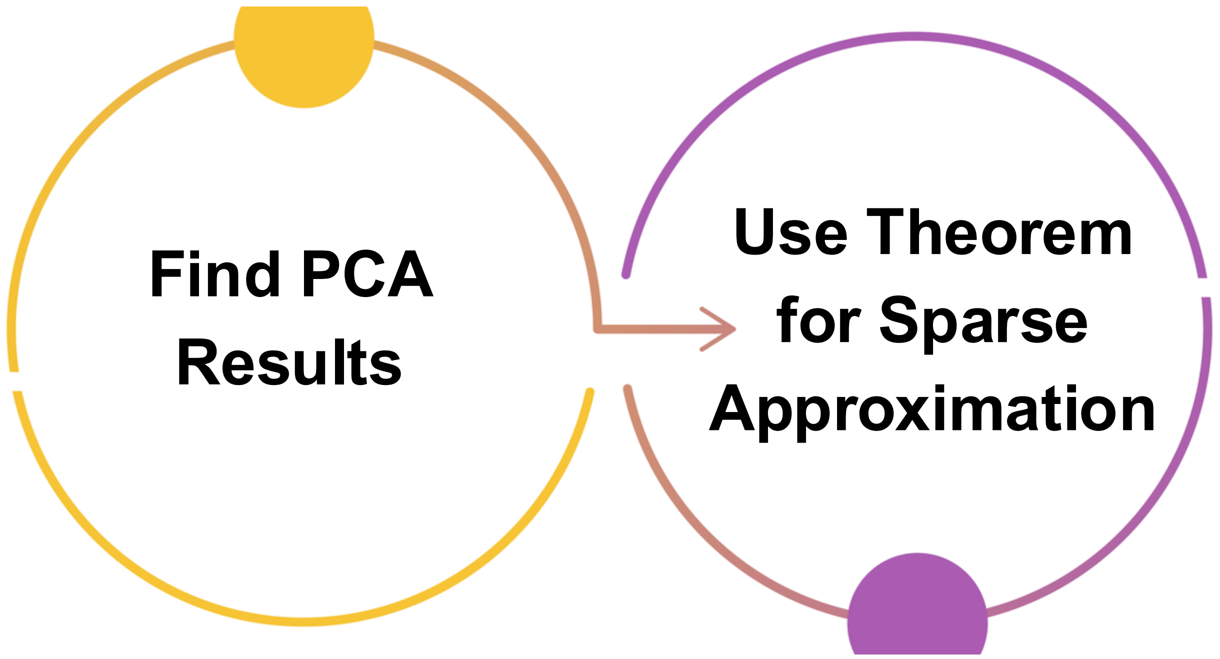
- If p > n, OLS has no unique solution
- PCA is always have unique solution while utilizing SVD
- Ridge penalty removes indeterminacy and ensures stability
- After normalization, the result is independent of λ

Adding Lasso for Sparsity

$$\hat{\beta} = \arg\min_{\beta} \|Z_i - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2 + \lambda_1\|\beta\|_1, \quad \text{Where} \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \qquad \widehat{V}_i = \frac{\hat{\beta}}{\|\hat{\beta}\|}$$

- Adding L1 Penalty
- This is called **Naive Elastic Net**, differs from elastic net by scaling factor **(1+λ)**
- A large enough $\lambda 1$ → gives sparse β, and thus sparse V*i*
- Provides a flexible and efficient way to obtain sparse approximation

Two Stage Analysis →



→ MOTIVATION

SPCA – Criterion and Algorithm

Estimate matrices $A, B \in \mathbb{R}^{p\times k}$:

$$\arg\min_{A,B} \|X - XBA^T\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j}\|\beta_j\|_1 \quad \text{s.t. } A^TA = I_k$$

$A \in \mathbb{R}^{p\times k}$ Orthonormal matrix of projection directions

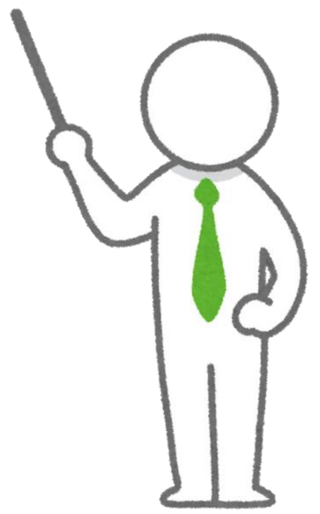
$B \in \mathbb{R}^{p\times k}$ Matrix of sparse loadings (β_j)
Determines which variables influence each component

1. **Initialization**:
   Let A start at $V[:, 1:k]$, the loadings of the first k ordinary principal components.
2. **Elastic Net Step (Fix A)**:
   Given $A = [\alpha_1, \dots, \alpha_k]$, solve the elastic net problem for $j = 1, 2, \dots, k$:

$$\beta_j = \arg\min_{\beta} \left\|X^TX(\alpha_j - \beta)\right\|^2 + \lambda\|\beta\|^2 + \lambda_{1,j}\|\beta\|_1$$

3. **Procrustes Step (Fix B)**:
   For fixed $B = [\beta_1, \dots, \beta_k]$, compute the SVD of $X^TXB = UDV^T$, then update $A = UV^T$
4. **Repeat** Steps 2–3 until convergence.
5. **Normalize** loadings:

$$\hat{V}_j = \frac{\beta_j}{\|\beta_j\|}, \quad j = 1, \dots, k$$



SPCA: Pros & Cons

Pros

- Produces sparse loadings → easier to identify important variables
- Dimension Reduction, retains most variance using fewer variables
- Flexibility: Controls sparsity via tuning (lasso) and stability via ridge
- Efficient Algorithms: Solved via alternating updates + Elastic Net
- Extends PCA: Reduces to standard PCA when sparsity penalty → 0
- Better suited for High-Dimensional Data

Cons

- Loss of Orthogonality: Sparse PCs are not guaranteed to be uncorrelated
- Requires careful choice of λ and λ_1,j
- Conceptually and computationally more involved than PCA

Adjusted Total Variance in SPCA

Issue: In SPCA, PCs may be correlated, so it overestimates true variance

$$\text{Adjusted variance} = \sum_{j=1}^{k} R_{jj}^2$$

Solution: Use adjusted total variance, corrects for overestimation due to correlated components

- Using **QR decomposition**: ***Z* = *QR***, where ***Q*** is orthonormal, ***R*** is upper-triangular.

Pitprops Dataset Example – Tables

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

PCA Results

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

SCoTLASS Results

Pitprops Dataset – Tables

SPCA Results

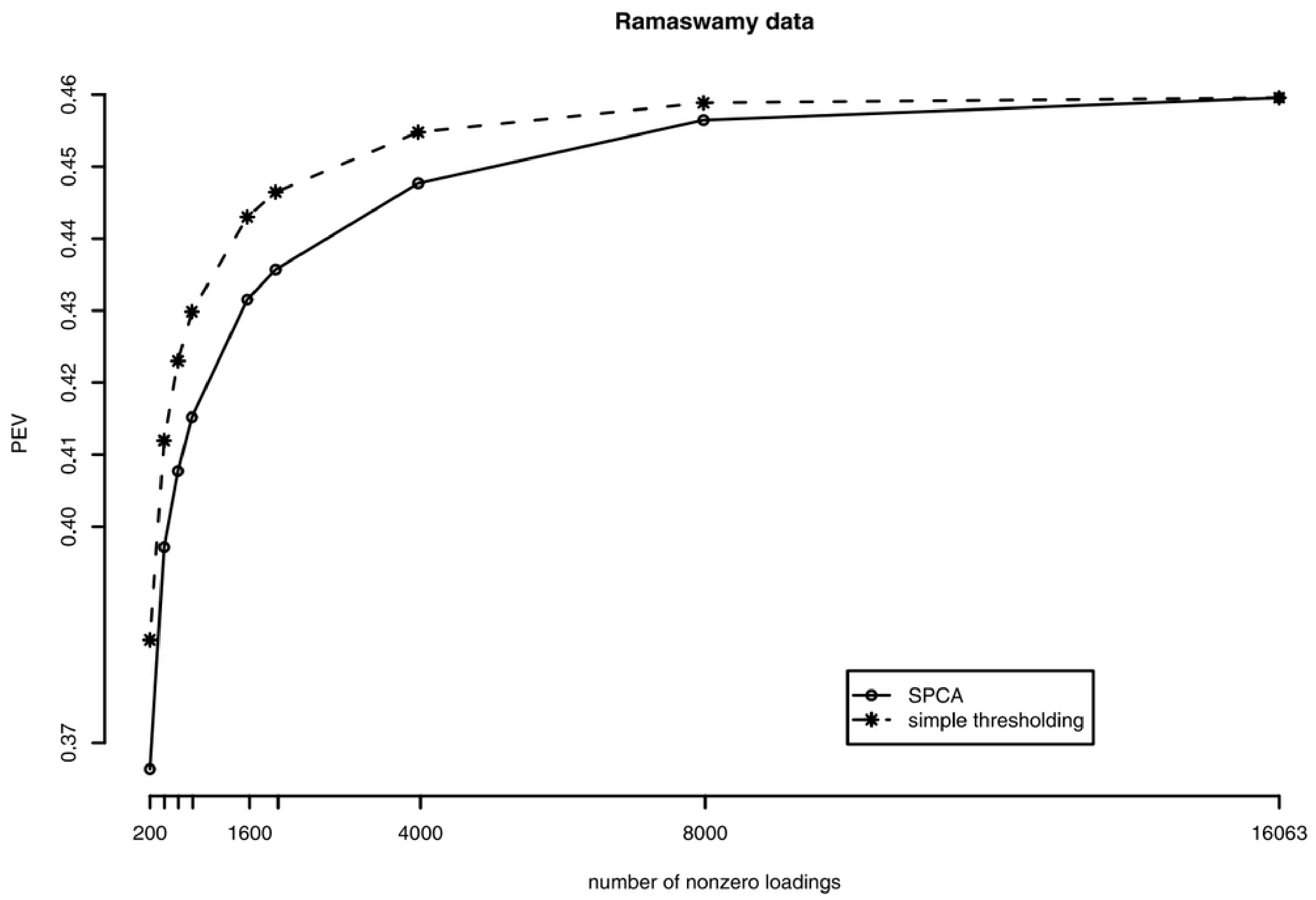
Table 3. Pitprops Data: Loadings of the First Six Sparse PCs by SPCA. Empty cells have zero loadings.

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

Pitprops Data: Result Summary

SPCA	SCoTLASS	PCA
75.8% Variance Explained	69.3% Variance Explained	Total variance explained by first 6 PCs
High sparsity	Sparse loadings	No sparsity

Gene Selection from Ramaswamy Microarray Data



- Dataset: 16,063 genes, 144 samples
- Goal: Identify genes that best explain gene expression variance
- SPCA and Simple Thresholding Methods
- Only ~2.5% of genes (≈400 genes) are enough to retain ~40% variance
- Original first PC explains 46% of variance
- SPCA trades slight variance loss for much higher interpretability

SPCA: A Principled Solution for Sparse Dimension Reduction

- Based on regression-type optimization
- Reduces to PCA when penalty vanishes
- Flexible control over sparsity
- Efficient algorithms for all data sizes
- High explained variance
- Better variable identification
- SPCA is implemented in R - *elasticnet*


Q&A