Introduction of Sparsity in Principal Components Analysis[1]

A. Bhattacharjee* R. Mondal* R. Vasishtha* S. S. Banerjee*

*Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

May 9, 2022

[1]Main References: Zou et al. (2006), Leng and Wang (2009)

Contents

Motivation of SPCA

Properties of Ordinary PCA

- Dimension reduction.
- Minimum loss of information.

Drawback of Ordinary PCA

- Each PC is a linear combination of all the p variables and the loadings are non-zero.

LASSO and Elastic Net

- Consider a regression model with n observations and p regressors. $\mathbf{Y}_{n\times 1}$ is the response vector. $\mathbf{X}_{n\times p}$ is the design matrix.

- **Lasso** estimate of regression parameter is given by,

$$\hat{\beta}_L = \arg\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} \mid \beta_j \mid$$

- **Elastic Net** estimate of regression parameter is given by,

$$\hat{\beta}_E = (1+\lambda_2)\left\{ \arg\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda_1 \sum_{j=1}^{p} \mid \beta_j \mid + \lambda_2 \sum_{j=1}^{p} \mid \beta_j \mid^2 \right\}$$

PCA through SVD

- $\mathbf{X}$ is an $n \times p$ data matrix.
- Without loss of generality it can be assumed that the column means of $\mathbf{X}$ are zero.
- Suppose that the SVD of $\mathbf{X}$ is given as.

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

- $\mathbf{Z} = \mathbf{U}\mathbf{D}$ are the Principal Components.
- The columns of $\mathbf{V}$ are the corresponding loadings of the PCs.

Direct Sparse Approximation I

Theorem (1)

For each i denote the i-th PC by $Z_i = \mathbf{UD}_i$ *Consider a positive* λ *and the ridge estimate is given by,*

$$\hat{\beta}_R = \arg\min_{\beta} ||Z_i - \mathbf{X}\beta||^2 + \lambda||\beta||^2 \tag{1}$$

Let $\hat{v} = \frac{\hat{\beta}_R}{||\hat{\beta}_R||}$, *then* $\hat{v} = \mathbf{V}_i$.

Here $\mathbf{D}_i$ is the i-th column of $\mathbf{D}$ and and $\mathbf{V}_i$ is the i-th column of $\mathbf{V}$.

Direct Sparse Approximation II

- Theorem (1) establishes the connection between PCA and the regression method.
- It is possible to get sparse PCs by considering the following minimization problem,

$$\hat{\beta} = \arg\min_{\beta} (\mathbf{Z_i} - \mathbf{X}\beta)^T (\mathbf{Z_i} - \mathbf{X}\beta) + \lambda ||\beta||^2 + \lambda_1 ||\beta||_1 \tag{2}$$

- Theorem (1) depends on the results of PCA and so it is not an alternative procedure.

SPCA Criterion I

Theorem (2)

Suppose we are considering the first k PCs. Let $\mathbf{A}_{p\times k} = [\alpha_1, \dots \alpha_k]$ *and* $\mathbf{B}_{p\times k} = [\beta_1, \dots \beta_k]$*. Then for any* $\lambda > 0$ *let,*

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i||^2 + \lambda \sum_{i=1}^{k} ||\beta_j||^2 \tag{3}$$

$$\textit{subject to } \mathbf{A}^T\mathbf{A} = I_{k\times k}$$

Then $\hat{\beta}_j \propto V_j$ *for* $j = 1, 2, \dots, k$.

SPCA Criterion III

Adding LASSO penalty to (3) and considering the following optimization problem,

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i||^2 + \lambda \sum_{i=1}^{k} ||\beta_j||^2 + \sum_{j=1}^{k} \lambda_{1,j} ||\beta_j||_1 \tag{4}$$

$$\text{subject to} \mathbf{A}^T\mathbf{A} = I$$

we can carry on the connection between PCA and regression using the LASSO approach to produce sparse loading. (4) is referred to as the SPCA criterion hereafter.

Numerical Solution

We discuss an algorithm to minimize the SPCA criterion function (4).
We note that (4) can be re-written as:

$$tr(\mathbf{X}^T\mathbf{X}) + \sum_{j=1}^{k} \left(\beta_j^T (\mathbf{X}^T\mathbf{X} + \lambda) \beta_j^T - 2\alpha_j^T \mathbf{X}^T\mathbf{X}\beta_j + \lambda_{1,j} |\beta_j|_1 \right)$$

Thus given $\mathbf{A}$, it is basically k independent elastic net problems. (4) can also be rewritten as:

$$tr(\mathbf{X}^T\mathbf{X}) - 2tr(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{B}) + tr\mathbf{B}^T(\mathbf{X}^T\mathbf{X} + \lambda)\mathbf{B} + \sum_{j=1}^{k} \lambda_{1,k} |\beta_j|_1$$

Thus if $\mathbf{B}$ is fixed, we should maximize $tr(\mathbf{A}^T(\mathbf{X}^T\mathbf{X})\mathbf{B}$ subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$.

Numerical Solution

Theorem

Let $\mathbf{A}$ and $\mathbf{B}$ be $p \times k$ matrices and $\mathbf{B}$ has rank k. Consider the constrained maximization problem,

$$\hat{\mathbf{A}} = \arg\max_{\mathbf{A}} tr(\mathbf{A}^T\mathbf{B}) \text{ subject to } \mathbf{A}^T\mathbf{A} = \mathbf{I}_k$$

Suppose the SVD of $\mathbf{B}$ is $\mathbf{B} = UDV^T$, then $\hat{\mathbf{A}} = UV^T$.

General SPCA Algorithm

Step 1: Initialize A as $V[,1:k]$, the loadings of first k *ordinary principal components*.

Step 2: Given fixed A, solve the following “naive” elastic net problem for $j = 1, \ldots, k$

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

Step 3: For each fixed $\mathbf{B}$, find SVD of $\mathbf{X}^T\mathbf{X}\mathbf{B} = UDV^T$. Then update $\mathbf{A} = UV^T$.

Step 4: Repeat steps 2-3 until $\mathbf{B}$ converges.

Step 5: Normalization: $\hat{V}_j = \beta_j/|\beta_j|,\ j = 1, \ldots, k$

Adjusted total variance

- The ordinary principal components are uncorrelated and their loadings are orthogonal, i.e., if $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}$, then $\mathbf{V}^T\mathbf{V} = \mathbf{I}_k$ and $\mathbf{V}^T\hat{\Sigma}\mathbf{V}$ is diagonal.
- PCs obtained by SPCA are not necessarily uncorrelated.
- Suppose $\hat{Z}$ be the modified PCs. If they are correlated, then $tr(\hat{Z}^T\hat{Z})$ does not yield the correct total variance explained by $\hat{Z}$.

Adjusted total variance

- We define $\hat{Z}_{j\cdot 1,\ldots,j-1}$ as the reminder of $\hat{Z}_j$ after adjusting the effects of of the remaining PCs, i.e.

$$\hat{Z}_{j\cdot 1,\ldots,j-1} = \hat{Y}_j - H_{1,\ldots,j-1}\hat{Y}_j$$

- Then the adjusted variance of $\hat{Z}_j$ is $|\hat{Z}_{j\cdot 1,\ldots,j-1}|^2$

- To easily calculate the adjusted variance easily, we use QR decomposition. Let $\hat{Z} = QR$, where Q is orthonormal and R is upper triangular, then

$$|\hat{Z}_{j\cdot 1,\ldots,j-1}|^2 = R_{j,j}^2$$

- Clearly the explained total variance is equal to $\sum_{j=1}^{k} R_{j,j}^2$.

Problem with SPCA: Using Adaptive LASSO

- **Problem:** When $p \ll n$, the excessive shrinkage equally applied by lasso to each coefficient seems to be problematic, at least in the least-squares setting (Zou (2006)).

- **Solution:** Modify the lasso penalty so that different shrinkage can be used for different coefficients, leading to a consistent selection of the important coefficients with high efficiency. (**Adaptive LASSO**, Zou (2006))

GAS-PCA

- SPCA is improved upon by modifying (4) in the following two ways:
  1. LASSO method is replaced by Adaptive LASSO.
  2. The least-squares objective function in S-PCA is replaced by a generalized least-squares objective function.

- **Intuitive Justifications:**
  1. Using generalized least squares allows incorporates a broader class of estimators.
  2. If more shrinkage is used for the zero coefficients with less shrinkage for the nonzero ones, an estimator with higher efficiency may be obtained.

GAS-PCA criterion

- Minimize the following general least-squares objective function:

$$\sum_{j=1}^{d_0} \left\{ (\alpha_j - \beta_j)' \tilde{\Omega} (\alpha_j - \beta_j) + \sum_{k=1}^{d} \lambda_{jk} |\beta_{jk}| \right\}, \tag{5}$$

  where $\tilde{\Omega}$ is a positive definite matrix with a probabilistic limit Ω, a positive definite matrix, referred to as the *kernel matrix*.

- **BIC criterion:**

$$BIC_{\lambda j} = (\alpha_j - \beta_j)' \tilde{\Omega} (\alpha_j - \beta_j) + df_{\lambda j} \times \frac{\log n}{n}.$$

  Here $df_{\lambda j}$ is the number of nonzero coefficients identified in $\hat{\beta}_{\lambda j}$

Choice of $\tilde{\Omega}$: LSA

- **LSA:** Estimator produced by minimizing the following least-squares–type objective function (Wang and Leng (2007)):

$$(\hat{\theta}-\theta)' \hat{cov}(\hat{\theta})(\hat{\theta}-\theta) + \sum_{k=1}^{d} \lambda_k |\theta_k|.$$

- **Choice of $\tilde{\Omega}$:** $cov^{-1}(\tilde{\beta}_j)$.

- No simple formula exists for $cov^{-1}(\tilde{\beta}_j)$.

- $\hat{cov}(\tilde{\beta}_j) = cov_s(\hat{\beta}_j^{boot})$, where $\hat{\beta}_j^{boot}$ are bootstrap samples drawn from $\mathcal{N}(0, \tilde{\Sigma})$.

Theoretical Results: Some Notations

- $a_n = \{\lambda_{jk} : \beta_{jk} \neq 0 : 1 \leq j \leq d_0, 1 \leq k \leq d\}$
- $b_n = \{\lambda_{jk} : \beta_{jk} = 0 : 1 \leq j \leq d_0, 1 \leq k \leq d\}$
- We fix $\hat{\alpha}_{\lambda j}$ to be fixed at $\bar{\alpha}_j \in \mathbb{R}^d$
- $\bar{\beta}_{\lambda j} = \text{argmin}_{\beta_j}\{(\bar{\alpha}_j - \beta_j)'\tilde{\Omega}(\bar{\alpha}_j - \beta_j) + \sum_{k=1}^{d} \lambda_{jk}|\beta_{jk}|\}$
- $s_j = \{1 \leq k \leq d : \beta_{jk} \neq 0\}$
- $\hat{s}_j^{BIC} = \{1 \leq k \leq d : \bar{\beta}_{\lambda jk} \neq 0\}$

Theoretical Results

Theorem

Assume that $\bar{\alpha}_j - \beta_j = O_p(n^{-1/2})$ *and that* $\tilde{\Omega}$ *converges in probability to some positive definite matrix* Ω*,* $\sqrt{n}a_n \to 0$*, and* $\sqrt{n}b_n \to \infty$*. We have:*

1. $\bar{\beta_{\lambda j}} - \beta_j = O_p(n^{-1/2})$
2. $P(\bar{\beta}_{\lambda jk} = 0) \to 1$ *for every* $\beta_{jk} = 0$*.*

Theoretical Results

Theorem

Assume that $\bar{\alpha}_j - \beta_j = O_p(n^{-1/2})$ *and that* $\tilde{\Omega}$ *converges in probability to some positive definite matrix* Ω*. We have:*

$$P(\hat{s}_j^{BIC} = s_j) \to 1.$$

Simulation Example

We first created three hidden factors

$$V_1 \sim N(0, 290), \quad V_2 \sim N(0, 300)$$

$$V_3 = -0.3V_1 + 0.925V_2 + \varepsilon, \quad \varepsilon \sim N(0, 1)$$

V_1, V_2 and ε are independent.
Then 10 observed variables were generated as the follows
$X_i = V_1 + \varepsilon_i^1, \quad \varepsilon_i^1 \sim N(0, 1), \quad i = 1, 2, 3, 4,$
$X_i = V_2 + \varepsilon_i^2, \quad \varepsilon_i^2 \sim N(0, 1), \quad i = 5, 6, 7, 8,$
$X_i = V_3 + \varepsilon_i^3, \quad \varepsilon_i^3 \sim N(0, 1), \quad i = 9, 10,$

Simulation

Table: Comparision of performance of SPCA and GAS-SPCA

	SPCA			GAS-SPCA		
	PC1	PC2	PC3	PC1	PC2	PC3
1	0	0.499	0	0	0.500	0
2	0	0.500	0	0	0.500	0
3	0	0.500	0	0	0.500	0
4	0	0.501	0	0	0.500	0
5	0.499	0	0	0.500	0	0
6	0.500	0	0	0.500	0	0
7	0.500	0	0	0.500	0	0
8	0.500	0	0	0.500	0	0
9	0	0	0.707	0	0	0.707
10	0	0	0.707	0	0	0.707

Pitprops data

- n = 180 and p = 13.

Table: SPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.477	0	0	0	0	0
length	-0.476	0	0	0	0	0
moist	0	0.785	0	0	0	0
testsg	0	0.619	0	0	0	0
ovensg	0.177	0	0.641	0	0	0
ringtop	0	0	0.589	0	0	0
ringbut	-0.250	0	0.492	0	0	0
bowmax	-0.344	-0.021	0	0	0	0
bowdist	-0.416	0	0	0	0	0
whorls	-0.400	0	0	0	0	0
clear	0	0	0	-1	0	0
knots	0	0.013	0	0	-1	0
diaknot	0	0	-0.016	0	0	1

Pitprops data

Table: GAS-SPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	0	0	0	0	0	0
length	0	1	0	0	0	0
moist	0	0	0	0	0	0.240
testsg	0.043	0	0	0	0	0
ovensg	0	0	0	0	0	-0.971
ringtop	0.572	0	0	0	0	0
ringbut	0.461	0	0	0.124	0	0
bowmax	0	0	0	0	0	0
bowdist	0	0	0	0	0	0
whorls	0	0	0	0.438	0	0
clear	0.376	0	0	-0.891	0	0
knots	0	0	0	0	1	0
diaknot	-0.563	0	0	0	0	0

Teaching data

- This dataset is about the teaching evaluation scores of 251 courses taught in the Peking University.
- Each observation corresponds to one course taught during the period from 2002 to 2004, and records the average scores on the students' agreement with the nine statements.

Teaching data

Table: SPCA

Variable	PC1	PC2	PC3
Q 1	0.487	0	0.323
Q 2	0.346	0	0.338
Q 3	0.347	0	0.308
Q 4	0	0.619	0
Q 5	0	0.559	0
Q 6	0	0.552	0
Q 7	0.502	0	-0.636
Q 8	0.399	0	-0.430
Q 9	0.333	0	0.311

Teaching data

Table: GAS-SPCA

Variable	PC1	PC2	PC3
Q 1	0.483	0	0.320
Q 2	0.376	0	0.331
Q 3	0.328	0	0.224
Q 4	0.110	0.643	0
Q 5	0	0.515	0
Q 6	0	0.567	0
Q 7	0.458	0	-0.658
Q 8	0.394	0	-0.468
Q 9	0.375	0	0.291

Leng, C. and Wang, H. (2009). On general adaptive sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 18(1):201–215.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.