The R package sparsePCA for block approaches and group-sparse PCA.

Marie Chavent [1,2] Guy Chavent [3]

[1] Inria Bordeaux Sud-Ouest

[2] Université de Bordeaux

[3] Inria Paris

Numerical data matrix A of rank r :

- n observations
- p variables centered or standardized

The aim for Principal Component Analysis :

- find $m \leq r$ vectors v_j in $\mathbb{R}^p$, $j = 1, \ldots, m$ (loadings),
- to build non correlated variables $y_j = Av_j$ (components) explaining the maximum variance of A.

The aim of sparse and group-sparse PCA :

- build sparse and group-sparse loading vectors v_j.
- select variables or groups of variables important to build y_j.

A small simulated data example

True loadings

v1	v2
0.422	0.000
0.422	0.000
0.422	0.000
0.422	0.000
0.000	0.489
0.000	0.489
0.000	0.489
0.000	0.489
0.380	-0.147
0.380	0.147

PCA loadings

v1	v2
0.395	0.244
0.338	0.208
0.446	0.027
0.359	0.031
-0.107	0.513
-0.183	0.532
-0.143	0.336
-0.015	0.353
0.417	-0.212
0.402	0.260

Sparse PCA loadings

v1	v2
0.423	0.089
0.354	0.067
0.459	0.000
0.361	0.000
-0.001	0.524
-0.082	0.590
-0.058	0.390
0.000	0.381
0.397	-0.229
0.433	0.125

Group-sparse PCA loadings

v1	v2
0.446	0.000
0.386	0.000
0.464	0.000
0.389	0.000
0.000	0.493
0.000	0.596
0.000	0.453
0.000	0.429
0.360	-0.098
0.395	0.039

Many methods and R packages already exist for sparse PCA for instance :

- Deflation approaches :
  - Shen & Huang, 2008 [5]
  - D'Aspremont, Bach & Ghaoui, 2008 [2]
  - Mackey 2009 [4] (R package nsprcomp)
- Block approaches :
  - Zou, Hastie & Tibshirani, 2006 [6] (R package elasticnet)
  - Journée & al., 2010 [3]

Few methods for group-sparse PCA :

- Deflation approach : Anne Bernard PhD thesis (2014) [1]
- Block approach :
  - Paper : Chavent & Chavent (2017), *Group-sparse block PCA and explained variance*, arXiv :1705.00461
  - R package `sparsePCA` : https://github.com/chavent/sparsePCA

Outline

1. Deflation and block approaches for standard PCA
2. A block approach for group-sparse PCA
3. Numerical results
   - simulation study to compare deflation and block approaches,
   - small example where group-sparse PCA is used to perform sparse PCA of a mixture of numerical and categorical variables.

Deflation approach : the m loadings v_j are calculated recursively as solutions of a vector optimization problem :

$$\begin{aligned} \text{Set } A_0 &= A \ , \ v_0 = 0 \ , \ \text{and compute, for } j = 1 \dots m : \\ A_j &= A_{j-1}(I_p - v_{j-1} v_{j-1}^T) \\ v_j &= \arg\max_{\|v\|=1} \|A_j v\|^2 \end{aligned}$$

Block approach : the m loadings v_j are calculated simultaneously as solutions of a block optimization problem.

Three equivalent block optimization problems :

$$\max_{V \in \mathcal{S}_m^p} \sum_{j=1\dots m} \mu_j^2 \|Av_j\|^2 \quad \text{(loading formulation)} \tag{1}$$

$$\max_{U \in \mathcal{S}_m^n} \sum_{j=1\dots m} \mu_j^2 \|A^T u_j\|^2 \quad \text{(component formulation)} \tag{2}$$

$$\max_{U \in \mathcal{S}_m^n} \max_{V \in (\mathcal{B}^p)^m} \sum_{j=1\dots m} \mu_j^2 (u_j^T A v_j)^2 \quad \text{(component/loading formulation)} \tag{3}$$

with block unknown :

$$\begin{cases} V & = & [v_1 \dots v_m] \in \mathbf{R}^{p \times m} & \text{(tentative loadings),} \\ U & = & [u_1 \dots u_m] \in \mathbf{R}^{n \times m} & \text{(tentative normalized components) .} \end{cases}$$

and weights :

- If $\mu_1 = \cdots = \mu_m = 1$, the solution V is only a basis of the eigenspace of $A^T A$.
- If $\mu_1 > \mu_2 > \cdots > \mu_m > 0$, the solution V is the eigenvector basis itself.

The objective function in (3) is penalized to give the following block formulation of the group-sparse PCA problem :

$$\max_{U \in \mathcal{S}_m^n} \max_{V \in (\mathcal{B}^{|p|})^m} \sum_{j=1 \ldots m} \mu_j^2 \left[u_j^T A v_j - \gamma_j \|v_j\|_1 \right]_+^2 \tag{4}$$

where $[t]_+ = t$ if $t \geq 0$ and $[t]_+ = 0$ if $t < 0$.

The penality term promote the apparition of zeros in the loading vectors for some group of variables using the group ℓ_1-norm.

$$\|v_j\|_1 = \sum_{i=1}^{G} \|v_{i,j}\| \ ,$$

where $v_{i,j}$ is the vector of the loadings of the variables in group i :

$$v_j^T = \left[v_{1,j}^T \ldots {v_{i,j}}^T \ldots v_{G,j}^T \right]$$

Advantages and disavantages of this penalized component/loading formulation.

- The group-sparse loadings and the principal components produced by this formulation are not orthogonal.
+ The good side is that the numerical difficulties are split between U and V : the orthonormality constraint is for U and the non-differentiable ℓ_1-norm is for V.
+ The inner maximization loop on V can be solved analytically leadings to the maximization of a convex differentiable function :

$$\max_{U \in \mathcal{S}_m^n} F(U),$$

where

$$F(U) = \sum_{j=1 \ldots m} \mu_j^2 \sum_{i=1}^{G} \left[\|a_i{}^T u_j\| - \gamma_j \right]_+^2 ,$$

and

$$A = \left[a_1 \ldots a_i \ldots a_G \right] .$$

The solution (U^*, V^*) is obtained in two steps

1. Determine U^* which maximizes $F(U)$ using the following gradient algorithm :

input : $U_0 \in \mathcal{S}_m^n$
output : U_n (approximate solution)
begin
  $0 \longleftarrow k$
  repeat
    $T_k \longleftarrow S_\gamma(A^T U_k)$
    $G_k \longleftarrow \nabla_X F(U_k) = 2AT_k N^2$
    $U_{k+1} \longleftarrow \mathrm{polar}(G_k)$
    $k \longleftarrow k+1$
  until a stopping criterion is satisfied
end

2. Define V^* as the (normalized) matrix $S_\gamma(A^T U^*)$ where S_γ is a group-soft thresholing operator applying to the columns of $A^T U^*$ using the sparsity parameters $\gamma_1, \ldots, \gamma_m$..

With the R package `sparsePCA` :

```
groupsparsePCA(A, m, lambda, index = 1:ncol(A),
               block = 1, mu = 1/1:m,
               center = TRUE, scale = TRUE)
```

- `A` is the data matrix,
- `m` is the number of components,
- `lambda` is the vector of the m reduced sparsity parameters sparsité λ_j,
- `index` defines the groups of variables,
- `block` defines the approach (`block=0` for deflation and `block=1` for block).
- `mu` is the vector of weights used in the block approach.

The reduced sparsity parameters λ_j are chosen in $[0, 1]$:

$$\lambda_j = \gamma_j / \gamma_{j,max} \quad , \quad j = 1 \dots m \ .$$

with $\gamma_{j,max}$ the *nominal sparsity parameter* of each component :

$$\gamma_{j,max} = \frac{\sigma_j}{\sigma_1} \gamma_{max} \quad \text{where} \quad \gamma_{max} \stackrel{\text{def}}{=} \max_{i=1\dots p} \|a_i\|_2 \ ,$$

Other function of the package package :

- the function `simuPCA`,
- the function `sparsePCA`,
- the function `explainedVar` (implements 5 definitions of the variance explained by the sparse components),
- the function `pev` give the pourcentage of the variance of the data explained by each sparse component,

Data simulated with the eigenvalues $(200, 100, 50, 50, 6 \ldots 1)$ and the following loadings :

True loadings

v1	v2
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1
0.9	-0.3
0.9	0.3

In this model :

- $p = 10$ variables,
- $G = 3$ groups of variables,
- $m = 2$ components.

```
library(sparsePCA)
v1 <- c(1,1,1,1,0,0,0,0,0.9,0.9)
v2 <- c(0,0,0,0,1,1,1,1,-0.3,0.3)
valp <- c(200,100,50,50,6,5,4,3,2,1)
A <- simuPCA(n = 100, cbind(v1,v2), valp, seed = 1) # data matrix 100*10
A <- scale(A, center = T, scale = F)
index <- c(1,1,1,1,2,2,2,2,3,3) # 3 groups of variables
```

Deflation approach with $m = 2$ components.

The reduced sparsity parameters are $\lambda_1 = 0.3$ and $\lambda_2 = 0.3$.

```
Vtrue
```

```
##           v1     v2
##  [1,] 0.422  0.000
##  [2,] 0.422  0.000
##  [3,] 0.422  0.000
##  [4,] 0.422  0.000
##  [5,] 0.000  0.489
##  [6,] 0.000  0.489
##  [7,] 0.000  0.489
##  [8,] 0.000  0.489
##  [9,] 0.380 -0.147
## [10,] 0.380  0.147
```

```
groupsparsePCA(A, block = 0,lambda = c(0.3,0.3),
               m = 2, index)$Z
```

```
##         [,1]  [,2]
##  [1,] 0.441 0.000
##  [2,] 0.430 0.000
##  [3,] 0.473 0.000
##  [4,] 0.445 0.000
##  [5,] 0.000 0.468
##  [6,] 0.000 0.511
##  [7,] 0.000 0.532
##  [8,] 0.000 0.487
##  [9,] 0.309 0.000
## [10,] 0.321 0.000
```

The loadings of the third group of variables are wrongly set to 0 in dimension 2.

Block approach with $m = 2$ components.

The weights of the block approach are $\mu_1 = \mu_2 = 1$.

This approach is called "block, same mu".

```
Vtrue

##           v1     v2
##  [1,] 0.422  0.000
##  [2,] 0.422  0.000
##  [3,] 0.422  0.000
##  [4,] 0.422  0.000
##  [5,] 0.000  0.489
##  [6,] 0.000  0.489
##  [7,] 0.000  0.489
##  [8,] 0.000  0.489
##  [9,] 0.380 -0.147
## [10,] 0.380  0.147
```

```
groupsparsePCA(A, block = 1,
               lambda = c(0.3,0.3),
               mu = c(1,1),
               m = 2,index)$Z

##         [,1]  [,2]
##  [1,] 0.443 0.000
##  [2,] 0.432 0.000
##  [3,] 0.473 0.000
##  [4,] 0.446 0.000
##  [5,] 0.000 0.468
##  [6,] 0.000 0.508
##  [7,] 0.000 0.528
##  [8,] 0.000 0.495
##  [9,] 0.301 0.000
## [10,] 0.322 0.000
```

The weights of the block approach are $\mu_1 = 1$ and $\mu_2 = 1/2$.

This approach is called "block, different mu".

```
Vtrue

##            v1     v2
##  [1,] 0.422  0.000
##  [2,] 0.422  0.000
##  [3,] 0.422  0.000
##  [4,] 0.422  0.000
##  [5,] 0.000  0.489
##  [6,] 0.000  0.489
##  [7,] 0.000  0.489
##  [8,] 0.000  0.489
##  [9,] 0.380 -0.147
## [10,] 0.380  0.147
```

```
groupsparsePCA(A, block = 1,
               lambda = c(0.3,0.3),
               mu = c(1,1/2),
               m = 2, index)$Z

##          [,1]  [,2]
##  [1,] 0.441 0.000
##  [2,] 0.431 0.000
##  [3,] 0.473 0.000
##  [4,] 0.446 0.000
##  [5,] 0.000 0.467
##  [6,] 0.000 0.506
##  [7,] 0.000 0.526
##  [8,] 0.000 0.499
##  [9,] 0.306 0.000
## [10,] 0.321 0.000
```

Comparison of the deflation and block approaches.

Data simulated with the eigenvalues $(200, 180, 150, 130, 1 \ldots 1)$ and the following loadings :

True loadings

v1	v2	v3	v4
6	0	0	6
-6	0	0	6
6	0	0	6
-6	0	0	6
0	12	12	0
0	12	12	0
0	-12	12	0
0	-12	12	0
-5	8	0	5
-5	8	0	-5
5	8	0	5
5	8	0	-5
4	0	0	-10
4	0	0	-10
4	0	0	-10
4	0	0	-10
8	5	8	5
8	5	-8	5
8	-5	8	5
8	-5	-8	5

In this model :

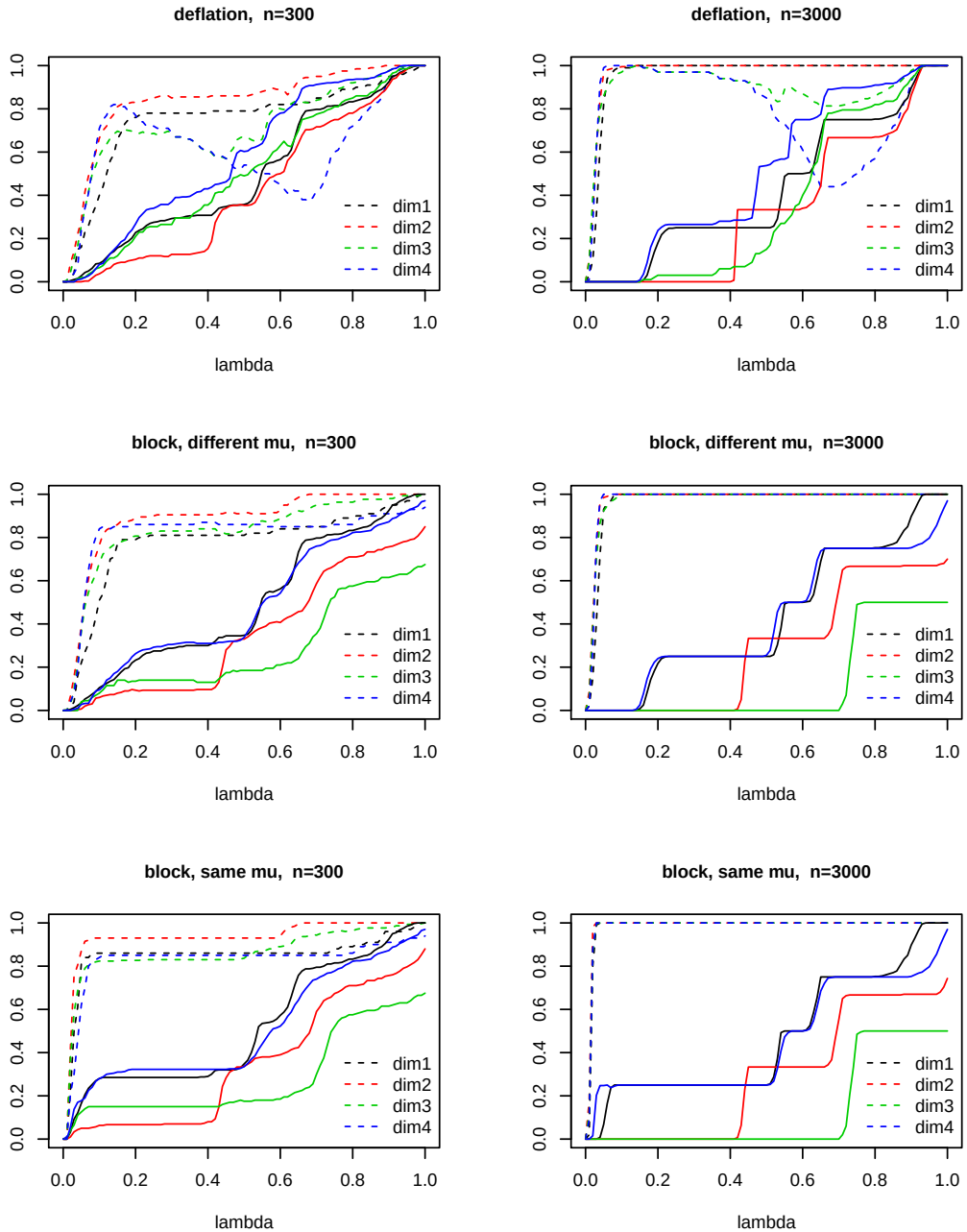
- $p = 20$ variables,
- $G = 5$ groups of variables of size 4,
- $m = 4$ components.

1. Simulation of :
   - 100 matrices A with $n = 300$ observations,
   - 100 matrices A with $n = 3000$ observations.
2. Perform for each matrix A the group-sparse loadings matrix V with the deflation approach, the block, same mu approach and the block different mu approach and a grid of reduced sparsity parameters :

$$\lambda = \lambda_1 = \cdots = \lambda_m \ ,$$

   with λ varying from 0 to 1 by steps of 0.01.
3. Compare the three approaches with :
   - true positive rate (tpr) : proportion of 0 in V_{true} set to 0 in V,
   - false positive rate (fpr) : proportion of non zéros in V_{true} set to 0 in V.

Mean true positive rates (dotted lines) and false positive rates (full lines).



Sparse PCA of a mixture of numerical and categorical variables.

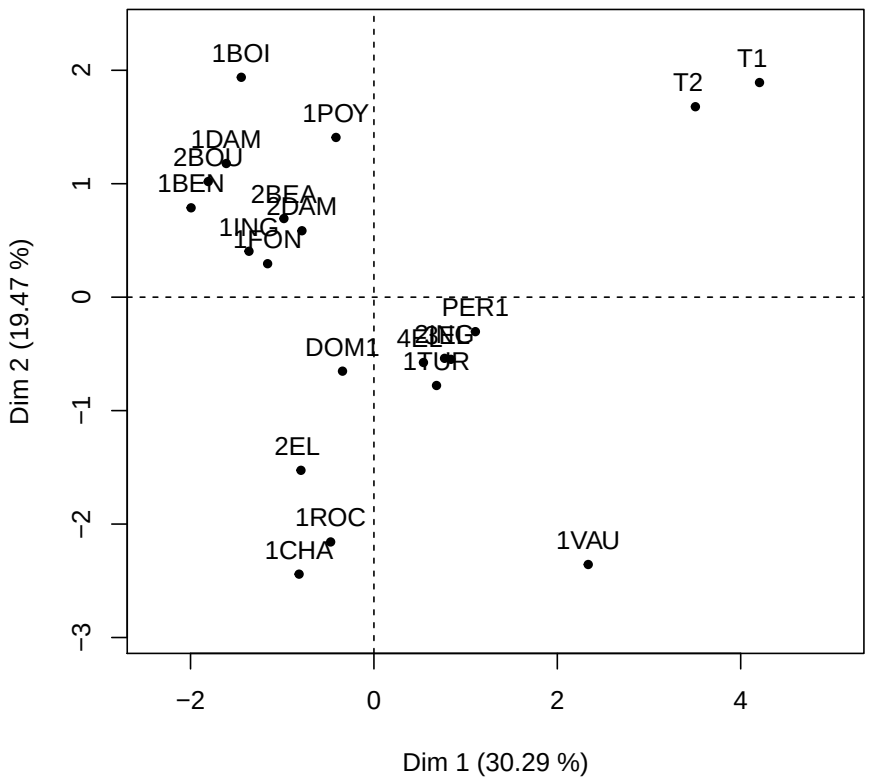
The wine data set of dimension 21×6 :

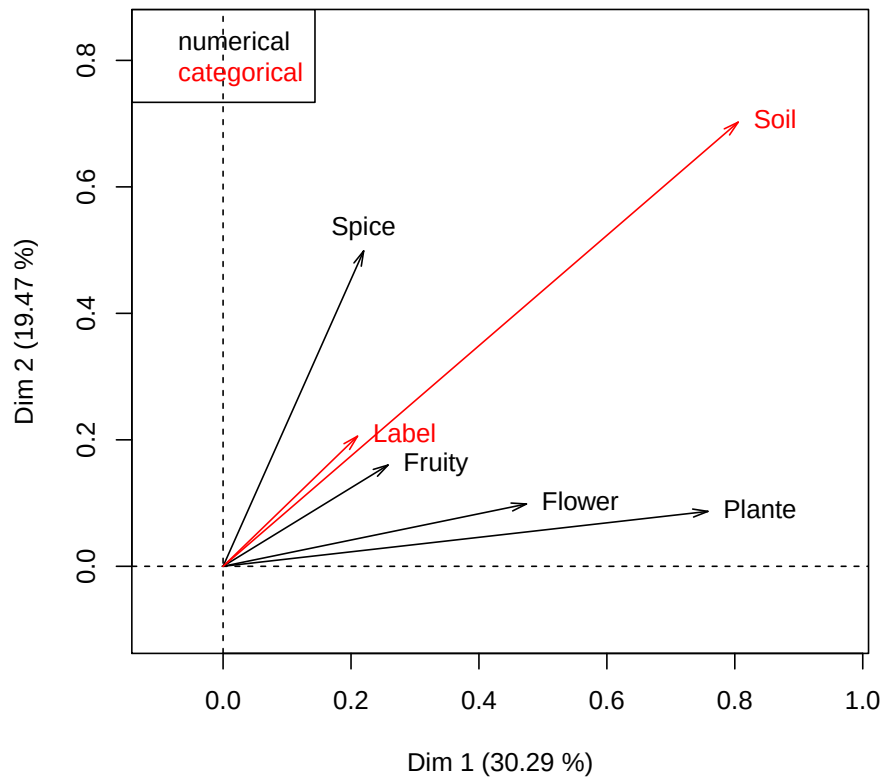
- ↪ 21 wines of Val de Loire
- ↪ 4 numerical sensory descriptors and 2 categorical variables :
  - Label of origin with 3 levels (Bourgueuil, Chinon, Saumur),
  - Soil with 4 levels (Env1, Env2, Env4, Reference).
- ↪ Two groups of indicator variables, one for each categorical variable.
- ↪ The function `sparsePCAmix` uses `groupsparsePCA`.

```
library(PCAmixdata)
data(wine)
head(wine[,c(1,2,13:16)])
```

```
##           Label      Soil Fruity Flower Spice Plante
## 2EL      Saumur      Env1   2.88   2.32  1.84   2.00
## 1CHA     Saumur      Env1   2.56   2.44  1.74   2.00
## 1FON Bourgueuil      Env1   2.77   2.19  2.25   1.75
## 1VAU     Chinon      Env2   2.39   2.08  2.17   2.30
## 1DAM     Saumur Reference   3.16   2.23  2.15   1.76
## 2BOU Bourgueuil Reference   2.80   2.24  2.15   1.75
```

```
X.quanti <- splitmix(wine[,c(1,2,13:16)])$X.quanti
X.quali <- splitmix(wine)$X.quali
pcamix <- PCAmix(X.quanti, X.quali, ndim = 2, graph=FALSE)
```





```
spcamix <- sparsePCAmix(X.quanti, X.quali,
                        m = 2, lambda = c(0.7,0.7),
                        block = 1, mu = c(1,1/2))
```

	v1	group-sparse v1	v2	group-sparse v2
Fruity	-0.308	0.000	0.302	0.000
Flower	-0.417	0.000	-0.237	0.000
Spice	0.284	0.000	0.533	0.592
Plante	0.527	0.710	-0.223	0.000
Bourgueuil	-0.122	0.000	0.106	0.000
Chinon	0.009	0.000	-0.122	0.000
Saumur	0.113	0.000	0.015	0.000
Env1	-0.050	-0.053	-0.136	-0.136
Env2	0.073	0.153	-0.168	-0.209
Env4	0.135	0.142	0.097	0.176
Reference	-0.158	-0.242	0.207	0.170

- Variance explained by non orthogonal components :
  - ↪ 5 definitions are proposed and compared in the paper and implemented in the R package.
- Choice of the sparsity parameters λ_j in abscence of information :
  - ↪ use the same reduced parameters λ for all loading and explore the influence of λ by letting it vary from 0 to 1 :
    - plot the proportion of explained variance by each dimension (implemented in the `pev` function),
    - perform the explained variance (using the `explainedVar` function) by cross validation (not implemented).
- Missing in the package :
  - a function to predict the (numerical) sparse components for further modelisations.
  - an option to taking the size of the groups into account.
- Block approaches in sparse LDA, sparse CCA....


Anne Bernard.

Development of statistical methods for genetic data analysis.

Theses, Conservatoire national des arts et metiers - CNAM, February 2014.


Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui.

Optimal solutions for sparse principal component analysis.

Journal of Machine Learning Research, 9(Jul) :1269–1294, 2008.


Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre.

Generalized power method for sparse principal component analysis.

Journal of Machine Learning Research, 11(Feb) :517–553, 2010.


Lester W Mackey.

Deflation methods for sparse pca.

In *Advances in neural information processing systems*, pages 1017–1024, 2009.


Haipeng Shen and Jianhua Z Huang.

Sparse principal component analysis via regularized low rank matrix approximation.

Journal of multivariate analysis, 99(6) :1015–1034, 2008.


Hui Zou, Trevor Hastie, and Robert Tibshirani.

Sparse principal component analysis.

Journal of computational and graphical statistics, 15(2) :265–286, 2006.

Merci de votre attention

Singular Value Decomposition of A :

$$A = U\Sigma V^T$$

- $u_1 \dots u_r$: left singular vectors (eigenvectors of AA^T)
- $v_1 \dots v_r$ of V : right singular vectors (eigenvectors of $A^T A$)
- $\sigma_1 \dots \sigma_r$: singular values (square roots of the eigenvalues of AA^T and AA^T)

Principal Component Analysis of A :

- The loadings v_j are the m first right singular vectors.
- The principal components y_j are proportionel to the m first left singular vectors :

$$y_j = \sigma_j u_j.$$

- The standard deviation of the principal components are the singular values :

$$Var(y_j) = \|Av_j\|^2 = \sigma_j^2.$$

Simulation of a matrix A of dimension $n \times p$ from a SVD model :

1. The $m \leq p$ eigenvectors $v_1, \ldots, v_m$ (right singular vectors) and the m eigenvalues $\sigma_1^2 \ldots \sigma_p^2$ of a covariance matrix are given in input and the covariance matrix C is built with :

$$C = V_{true} \Sigma_{true}^2 V_{true}^T$$

where :

- $\Sigma_{true}^{\mathbf{2}} = \mathrm{diag}(\sigma_{\mathbf{1}}^{\mathbf{2}} \ldots \sigma_p^{\mathbf{2}})$ is the diagonal matrix of the eigenvalues,
- V_{true} is the matrix of the eigenvalues, obtained using the following QR decomposition=

$$[v_{\mathbf{1}}, \ldots, v_m, U] = V_{true} R,$$

where U of dimension $p \times (p - m)$ is drawn randomly from a $U(0, 1)$ distribution.

2. n observations are randomly drawn from a $N(0_p, C)$ distribution to get the data matrix A.

The group-soft thersholing operator $S_\gamma(A^T U)$

- The matrix a_i of dimension $n \times p_i$ contains the data of group i and

$$A = \left[a_1 \dots a_i \dots a_p\right] .$$

- The vector $a_i^T u_j \in \boldsymbol{R}^{p_i}$ contains correlations between the p_i variables of group i end the jth normalized component u_j and

$$A^T U = \begin{pmatrix} a_1^T u_1 & \dots & a_1^T u_m \\ \vdots & & \vdots \\ a_G^T u_1 & \dots & a_G^T u_m \end{pmatrix}$$

- The vector $t_{ij} = S_{\gamma_j}(a_i^T u_j)$ is obtained by soft thresholding of the vector $a_i^T u_j$ i.e. the vector is set to 0 if its norm is smaller than γ_j and its length is reduced of γ_j otherwise :

$$t_{ij} = \frac{a_i^T x_j}{\|a_i^T x_j\|}[\|a_i^T x_j\| - \gamma_j]_+ \in \boldsymbol{R}^{p_i} .$$

- Finally

$$S_\gamma(A^T U) = \begin{pmatrix} t_{11} & \dots & t_{1m} \\ \vdots & & \vdots \\ t_{G1} & \dots & t_{G1} \end{pmatrix} = T$$

In order to take the size of the groups into account, the group-ℓ_1 norm is modified introducing a coefficient η_i for each group with usually $\eta_i = \sqrt{pi}$:

$$\|v_j\|_1 = \sum_{i=1}^{G} \eta_i \|v_{i,j}\|.$$

In that cas, the soft thresholding of the vector $a_i^T u_j$ is :

$$t_{ij} = \frac{a_i^T u_j}{\|a_i^T u_j\|} [\|a_i^T u_j\| - \eta_i \gamma_j]_+ \in \boldsymbol{R}^{p_i} \ .$$

Each sparsity parameter γ_j needs to be fitted to the norm of the vector $A^T u_j$ it is in charge of thresholding. This norm is simply estimated by its initial value

$$\|A^T u_j^0\| = \sigma_j,$$

where u_j^0 is the jth left singular vectors of A.

Moreover, it can be shown all the loading vector v_j is null if

$$\gamma_j \geq \gamma_{max}$$

with

$$\gamma_{max} \stackrel{\text{def}}{=} \max_{i=1\ldots p} \|a_i\|_2$$

and $\|a_i\|_2$ the first singular value of a_i.

A nominal maximum sparsity parameters $\gamma_{j,max}$ is then defined for each dimension :

$$\gamma_{j,max} = \frac{\sigma_j}{\sigma_1} \gamma_{max},$$

and finaly the reduced sparsity parameters λ_j are chosen in $[0, 1]$ with :

$$\lambda_j = \gamma_j / \gamma_{j,max} \quad , \quad j = 1 \dots m \ .$$

When no a priori information on the sparsity of the underlying loadings is known, we can use the same reduced parameters λ for all loadings :

$$\lambda = \lambda_1 = \cdots = \lambda_m \ ,$$

and explore the influence of λ by letting it vary from 0 to 1 by steps of $0,01$ for instance.

We recall the polar decomposition of a $k \times \ell$ matrix G :

$$G = UP \ , \tag{5}$$

where U is a $k \times \ell$ unitary matrix ($U^t U = I_\ell$) - not to be confused with the matrix U in the SVD of A, and P is a positive $\ell \times \ell$ semidefinite matrix ($P \geq 0$). The matrix U is called the polar matrix of G :

$$U = \mathrm{polar}(G) \ . \tag{6}$$

When G happens to be a vector, U is simply the unit vector pointing in the direction of G (or any unit vector if $G = 0$), and P is the norm of G.