le cnam

CE.R.I.E.S.

A Generalization of Sparse PCA to Multiple Correspondence Analysis

G. Saporta[1], A. Bernard[1,2], C. Guinot[2,3]

[1] CNAM, Paris, France
[2] CE.R.I.E.S., Neuilly sur Seine, France
[3] Université François Rabelais, Tours, France

5th International Conference of the ERCIM, 1-3 December 2012, Oviedo, Spain

Background

In case of high dimensional data, PCA or MCA components are nearly impossible to interpret.
Two ways for obtaining simple structures:

1. **Factor rotation** (varimax, quartimax etc.) well known in factor analysis for obtaining simple structure. Generalized to CA and MCA by Van de Velden & al (2005), Chavent & al (2012)
   But components are still combinations of **all** original variables

2. **Sparse methods** providing components which are combinations of **few** original variables, like **sparse PCA**.
   Extension for categorical variables:
   → Development of a new sparse method: **sparse MCA**

Outline

1. From Sparse PCA to Group Sparse PCA
    - 1.1 Lasso & elastic net
    - 1.2 Sparse PCA
    - 1.3 Group Lasso
    - 1.4 Group Sparse PCA

2. Sparse MCA
    - 2.1 Definitions
    - 2.2 Algorithm
    - 2.3 properties
    - 2.4 Toy example: “dogs” data set

3. Application to genetic data

Conclusion

1. From Sparse PCA to Group Sparse PCA

1.1 Lasso & elastic net

Lasso: shrinkage and selection method for linear regression (Tibshirani, 1996)

- Imposes the L1 norm on the linear regression coefficients

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Lasso continuously shrinks the coefficients towards zero
- Produces a sparse model but the number of variables selected is bounded by the number of units

Elastic net: combine ridge penalty and lasso penalty to select more predictors than the number of observations (Zou & Hastie, 2005)

$$\hat{\boldsymbol{\beta}}_{en} = \arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|_2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right) \quad \text{with} \quad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$$

1.2 Sparse PCA

In PCA each PC is a linear combination of **all** the original variables
→ Difficult to interpret the results

Challenge of SPCA: to obtain components easily interpretable (lot of zero loadings in principal factors)

Principle of SPCA: to modify PCA imposing lasso/elastic-net constraint to construct modified PCs with sparse loadings

Warning: Sparse PCA does not provide a global selection of variables but a selection **dimension by dimension** : different from the regression context (Lasso, Elastic Net, …)

Several attempts:

Simple PCA

by Vines (2000) : integer loadings

Rousson, V. and Gasser, T. (2004) : loadings (+ , 0, -)

SCoTLASS by Jolliffe & al. (2003) : extra L_1 constraints

Our technique is based on H. Zou, T. Hastie, R. Tibshirani (2006)

Let the SVD of X be $\mathbf{X} = \mathbf{UDV}^{\mathbf{T}}$ with $\mathbf{Z} = \mathbf{UD}$ the principal components

Ridge regression:

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\beta} \left\| \mathbf{Z} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\|^2$$

$$\mathbf{X}^{\mathbf{T}}\mathbf{X} = \mathbf{V}\mathbf{D}^{\mathbf{2}}\mathbf{V}^{\mathbf{T}} \; with \; \mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I}$$

$$\hat{\boldsymbol{\beta}}_{\mathbf{i,ridge}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \boldsymbol{\lambda}\,\mathbf{I}\right)^{-1}\mathbf{X}^{\mathbf{T}}\left(\mathbf{X}\mathbf{V}_{\mathbf{i}}\right) = \mathbf{V}_{\mathbf{i}}\frac{\mathbf{D}_{ii}^{2}}{\mathbf{D}_{ii}^{2} + \lambda} \quad \Longrightarrow \quad \tilde{v} = V_i$$

Loadings can be recovered by regressing (ridge regression) PCs on the p variables

→PCA can be written as a **regression-type optimization problem**

1.2 Sparse PCA

Sparse PCA add a new penalty to produce sparse loadings:

Lasso penalty

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

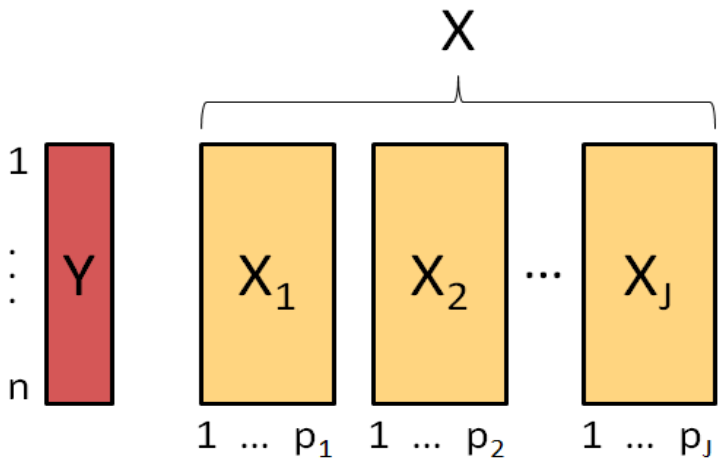
$\hat{\mathbf{V}}_i = \frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|}$ is an approximation to $\mathbf{V}_i$, and $\mathbf{X}\hat{\mathbf{V}}_i$ the ith approximated component

→ Produces sparse loadings with zero coefficients to facilitate interpretation

Alternated algorithm between elastic net and SVD

1.3 Group Lasso

X matrix divided into J sub-matrices X_j of p_j variables



Group Lasso: extension of Lasso for selecting groups of variables

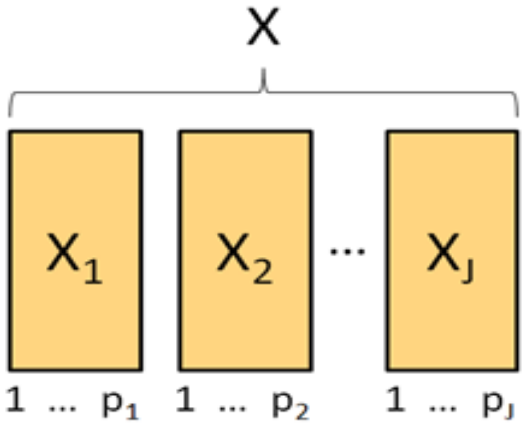
The group Lasso estimate is defined as a solution to (Yuan & Lin, 2007):

$$\hat{\boldsymbol{\beta}}_{GL} = \arg\min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} \left| \boldsymbol{\beta}_j \right|$$

If p_j=1 for all j, group Lasso = Lasso

1.3 Group Sparse PCA

Data matrix X still divided into J groups X_j of p_j variables, but no Y



Group Sparse PCA: compromise between SPCA and group Lasso

Goal: select groups of continuous variables (zero coefficients to entire blocks of variables)

Principle: replace the penalty function in the SPCA algorithm

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\| \mathbf{Z} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\|^2 + \lambda_1 \left\| \boldsymbol{\beta} \right\|_1$$

by that defined in the group Lasso

$$\hat{\boldsymbol{\beta}}_{GL} = \arg\min_{\beta} \left\| \mathbf{Z} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} \left| \boldsymbol{\beta}_j \right|$$

2. Sparse MCA

2.1 Definition

Original table

X_J
1
p_J
.
.
.
3

In MCA:

Selection of ***1 column*** in the original table (categorical variable X_j)

=

Selection of ***a block of p_j indicator variables*** in the complete disjunctive table

Complete disjunctive table

X_{J1}	…	X_{Jpj}
1		0
0		1
.		.
.		.
.		.
0		0

Challenge of Sparse MCA : select categorical variables, not categories

Principle: a straightforward extension of Group Sparse PCA for groups of indicator variables, with the chi-square metric

2.1 Correspondence analysis: notations

Let F be the n x q disjunctive table divided by the number of units

$$\mathbf{r} = \mathbf{F}\mathbf{1}_q \qquad \mathbf{c} = \mathbf{F}^T\mathbf{1}_n \qquad \mathbf{D}_r = \mathbf{diag}(\mathbf{r}) \qquad \mathbf{D}_c = \mathbf{diag}(\mathbf{c})$$

Let $\tilde{\mathbf{F}}$ be the matrix of standardised residuals:

$$\tilde{\mathbf{F}} = \mathbf{D}_r^{-\frac{1}{2}}\left(\mathbf{F} - \mathbf{r}\mathbf{c}^T\right)\mathbf{D}_c^{-\frac{1}{2}}$$

Singular Value Decomposition $\tilde{\mathbf{F}} = \mathbf{U}\Lambda\mathbf{V}^T$

2.2 Algorithm

1. Let α start at V[,1:K], the loadings of the first K PCs

2. Given a fixed $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}^{\mathbf{1}}, \ldots, \boldsymbol{\alpha}^{\mathbf{K}}\right]$; solve the group lasso problem for k=1, …, K (number of factors, K ≤ J) and j=1,…,J

$$\hat{\boldsymbol{\beta}}^{\mathbf{k}} = \arg\min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^{J} \tilde{F}_j \beta_j^k \right\|^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} \left| \beta_j^k \right|$$

$\mathbf{y} = \tilde{\mathbf{F}}\boldsymbol{\alpha}^{\mathbf{k}}$ and λ the tuning parameter

3. For a fixed $\boldsymbol{\beta} = \left[\boldsymbol{\beta}^{\mathbf{1}}, \ldots, \boldsymbol{\beta}^{\mathbf{K}}\right]$, compute the SVD of $\tilde{\mathbf{F}}^{\mathbf{T}}\tilde{\mathbf{F}}\boldsymbol{\beta} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}$ and update $\boldsymbol{\alpha} = \mathbf{U}\mathbf{V}^{\mathbf{T}}$

4. Repeat steps 2-3 until convergence

5. Normalization: $\tilde{\mathbf{V}}^{\mathbf{k}} = \boldsymbol{\beta}^{\mathbf{k}} / \left\| \boldsymbol{\beta}^{\mathbf{k}} \right\|$

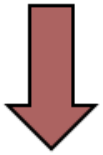
Properties	MCA	Sparse MCA
Uncorrelated Components	TRUE	FALSE
Orthogonal loadings	TRUE	FALSE
Barycentric property	TRUE	TRUE
% of inertia	$\lambda_j / tot \times 100$	$\left\| \tilde{\mathbf{z}}_{\mathbf{j.1,\ldots,j\text{-}1}} \right\|^2$
Total inertia	$\frac{1}{p}\sum_{j=1}^{p} p_j - 1$	$\sum_{j=1}^{k} \left\| \tilde{\mathbf{z}}_{\mathbf{j.1,\ldots,j\text{-}1}} \right\|^2$

$\tilde{\mathbf{z}}_{\mathbf{j.1,\ldots,j\text{-}1}}$ are the residuals after adjusting $\tilde{\mathbf{z}}_{\mathbf{j}}$ for $\tilde{\mathbf{z}}_{\mathbf{1,\ldots,j\text{-}1}}$ (regression projection)

2. Sparse MCA

Toy example: Dogs

X_1 Size	...	X_6 Aggressiveness
large (L)		agressive (A)
medium (M)		agressive (A)
.	...	.
.		.
.		.
small (S)		nonagressive (N)



K_1 Size			...	K_6 Aggressiveness	
S.	M.	L.		A	N
0	0	1		1	0
0	1	0		1	0
.	.	.	...	.	.
.	.	.		.	.
.	.	.		.	.
1	0	0		0	1

Data:

n=27 breeds of dogs
p=6 variables
q=16 (total number of columns)

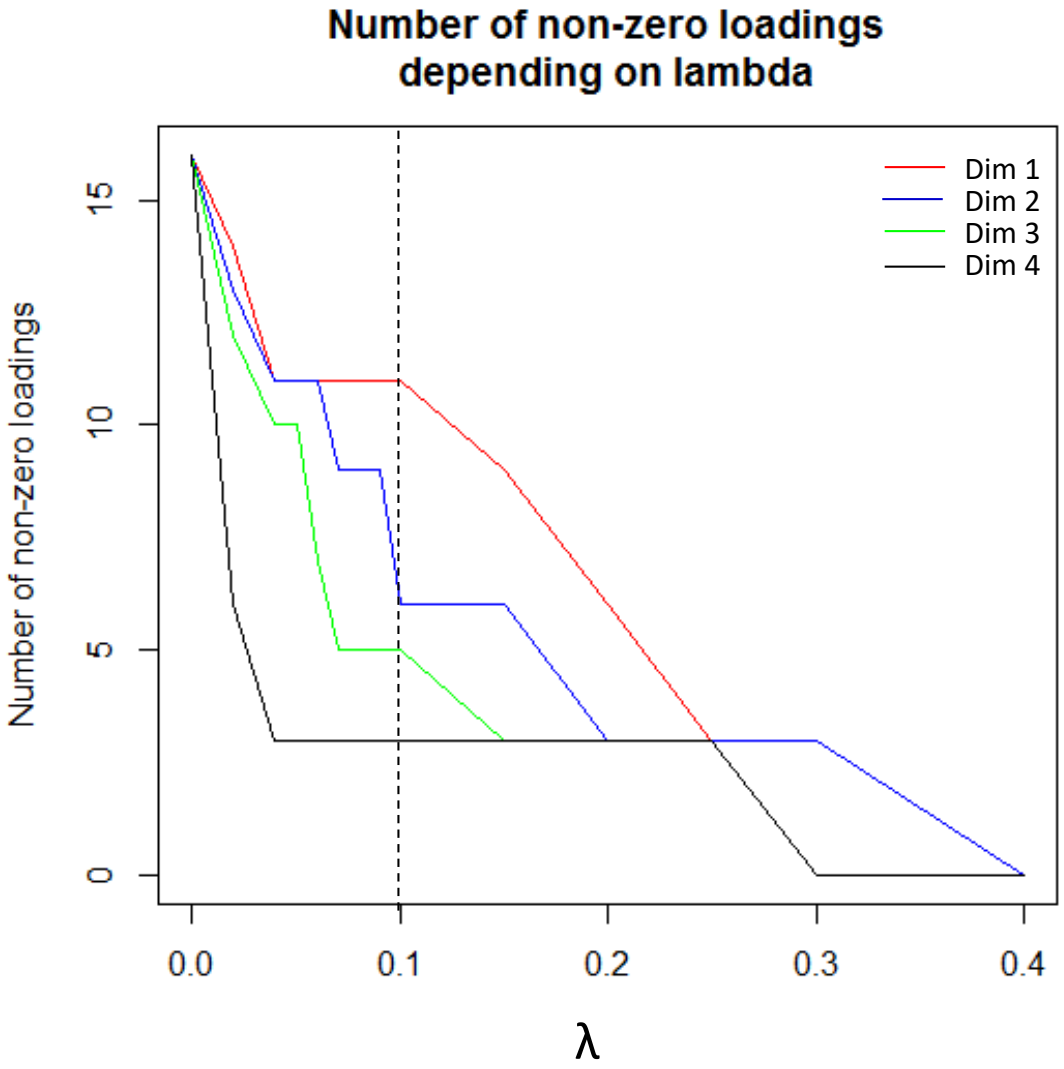
X : 27 x 6 matrix of categorical variables

K : 27 x 16 complete disjunctive table → $K=(K_1, \ldots, K_6)$

1 bloc
=
1 SNP = 1 K_j matrix

2. Sparse MCA

2.4 Toy example: Dogs



For λ=0.10:

- 11 non-zero loadings on the 1st axis
- 6 non-zero loadings on the 2nd axis
- 5 non-zero loadings on the 3rd axis
- 3 non-zero loadings on the 4th axis

2. Sparse MCA

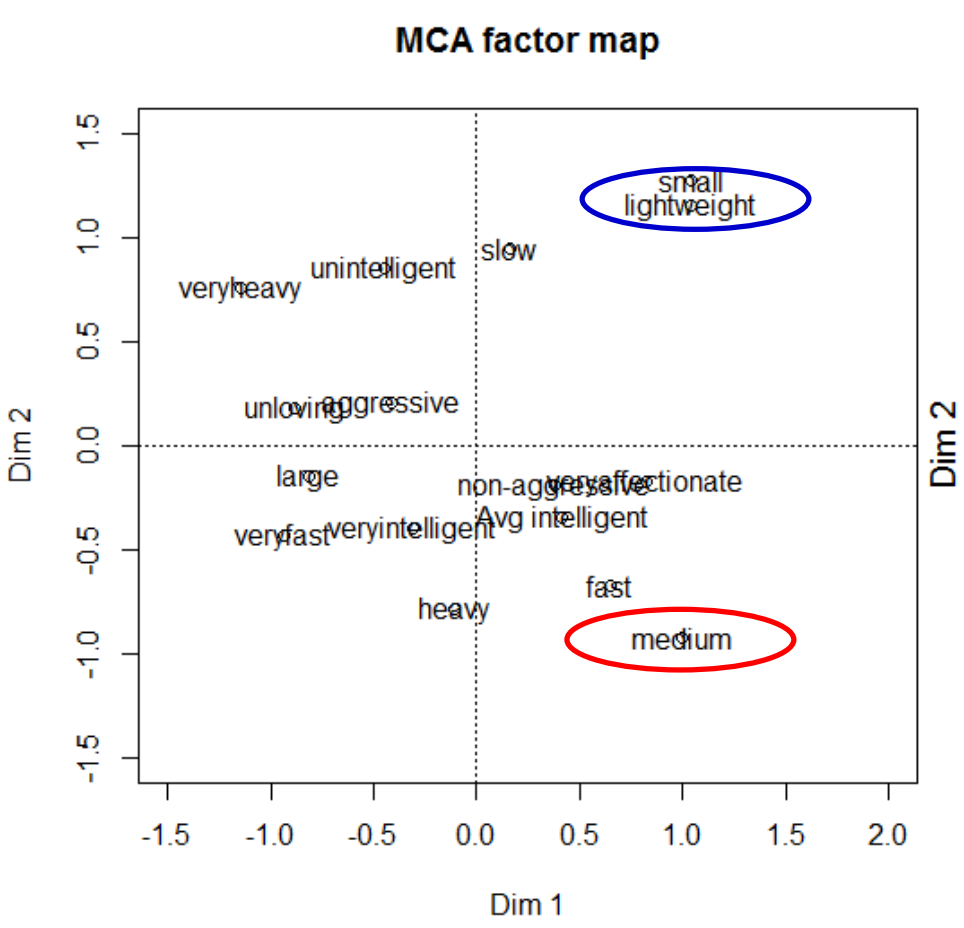
2.4 Toy example: Comparison of the loadings

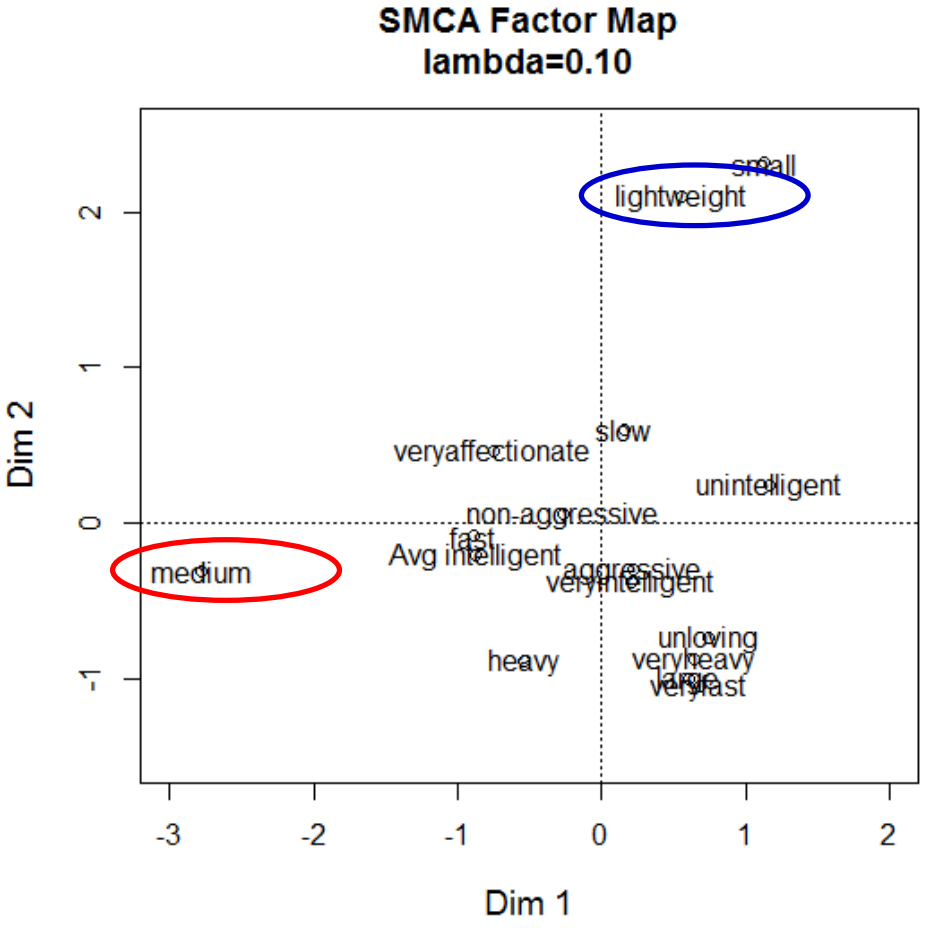
SNPs	MCA				Sparse MCA			
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 1	Dim 2	Dim 3	Dim 4
large	**-0.270**	**0.017**	-0.072	0.060	**-0.399**	**-0.517**	**0.000**	**0.000**
medium	**0.222**	**-0.444**	0.384	-0.065	**0.808**	**0.008**	**0.000**	**0.000**
small	**0.453**	**0.402**	-0.205	-0.085	**-0.331**	**0.610**	**0.000**	**0.000**
lightweight	0.437	**0.332**	**-0.098**	-0.091	**0.000**	**0.471**	**0.278**	**0.000**
heavy	-0.061	**-0.265**	**-0.118**	0.154	**0.000**	**-0.369**	**0.426**	**0.000**
veryheavy	-0.428	**0.332**	**0.493**	-0.334	**0.000**	**-0.059**	**-0.860**	**0.000**
slow	0.070	0.297	0.285	-0.144	-0.002	**0.000**	**0.000**	**0.000**
fast	0.177	-0.269	0.065	-0.019	0.013	**0.000**	**0.000**	**0.000**
veryfast	-0.286	-0.068	-0.429	0.201	-0.011	**0.000**	**0.000**	**0.000**
unintelligent	-0.052	0.328	-0.087	**0.417**	-0.184	**0.000**	**0.000**	**-0.248**
avg intelligent	0.087	-0.140	0.255	**0.096**	0.197	**0.000**	**0.000**	**-0.488**
veryintelligent	-0.118	-0.134	-0.437	**-0.764**	-0.035	**0.000**	**0.000**	**0.836**
unloving	-0.264	0.123	-0.028	0.076	-0.040	**0.000**	-0.007	**0.000**
veryaffectionate	0.245	-0.115	0.026	-0.070	0.040	**0.000**	0.007	**0.000**
aggressive	-0.113	0.079	0.053	-0.034	**0.000**	**0.000**	**0.000**	**0.000**
non-agressive	0.105	-0.074	-0.049	0.032	**0.000**	**0.000**	**0.000**	**0.000**
#non-zero loadings	16	16	16	16	11	6	5	3
% inertia	28.19	22.79	13.45	9.55	21.37	20.81	12.04	5.88

2. Sparse MCA

2.4 Toy example : comparison of displays

Comparison between MCA and Sparse MCA
on the first plan

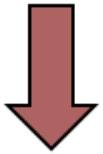




3. Application on genetic data

Single Nucleotide Polymorphisms

SNP 1=X_1	...	SNP 100=X_{100}
AA		AB
AB		BB
.	. . .	.
.		.
.		.
AA		AA
BB		AA



SNP 1=K_1			...	SNP 100=K_{100}		
AA	AB	BB		AA	AB	BB
1	0	0		0	1	0
0	1	0		0	0	1
.	.	.	. . .	.	.	.
.	.	.		.	.	.
.	.	.		.	.	.
1	0	0		1	0	0
0	0	1		1	0	0

Data:

n=502 individuals

p=100 SNPs (among more than 300 000 of the original data base)

q=281 (total number of columns)

X : 502 x 100 matrix of qualitative variables

K : 502 x 281 complete disjunctive table → $K=(K_1, \ldots, K_{100})$

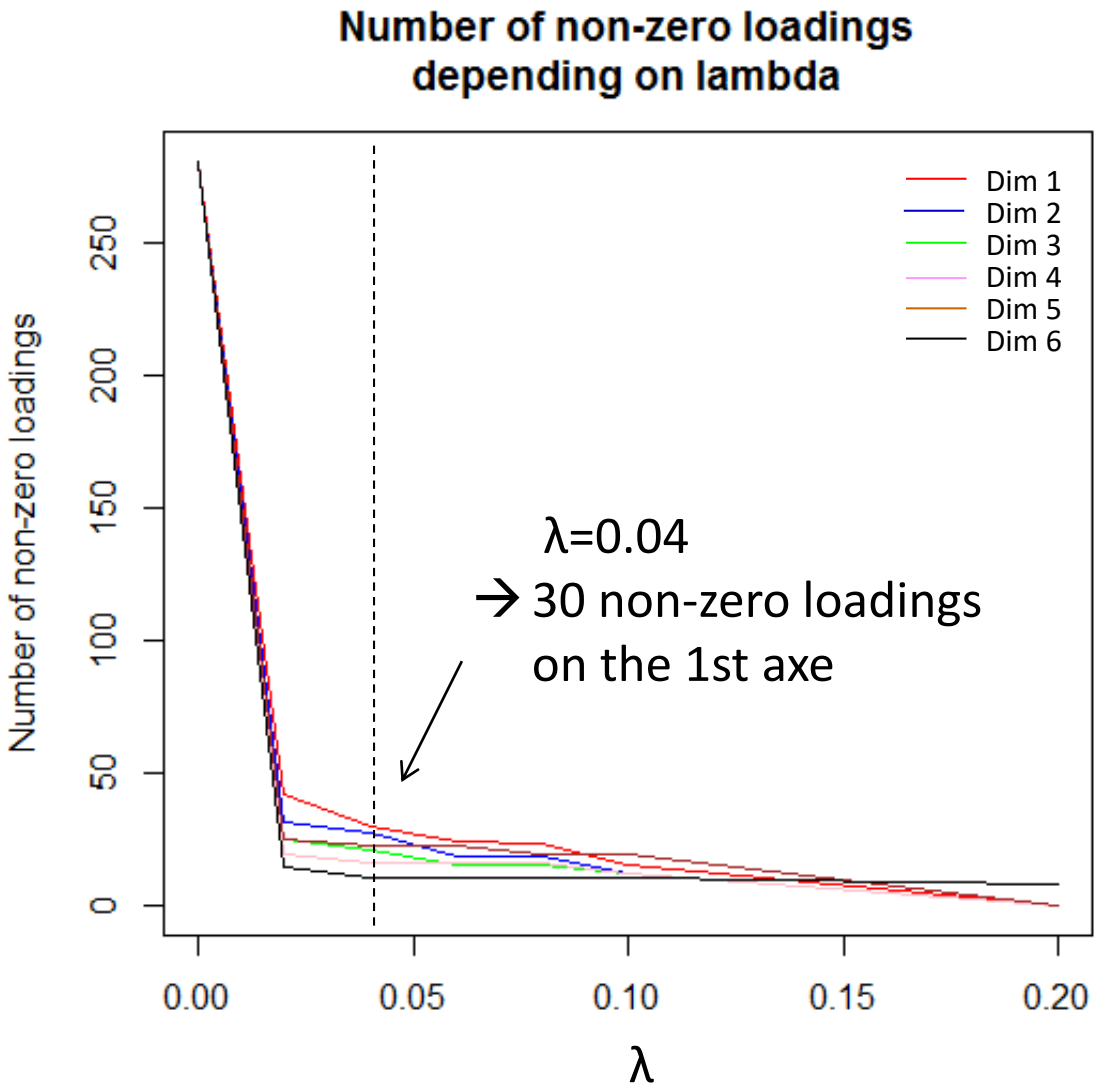
1 block

=

1 SNP = 1 K_j matrix

3. Application on genetic data

Single Nucleotide Polymorphisms



Application on genetic data

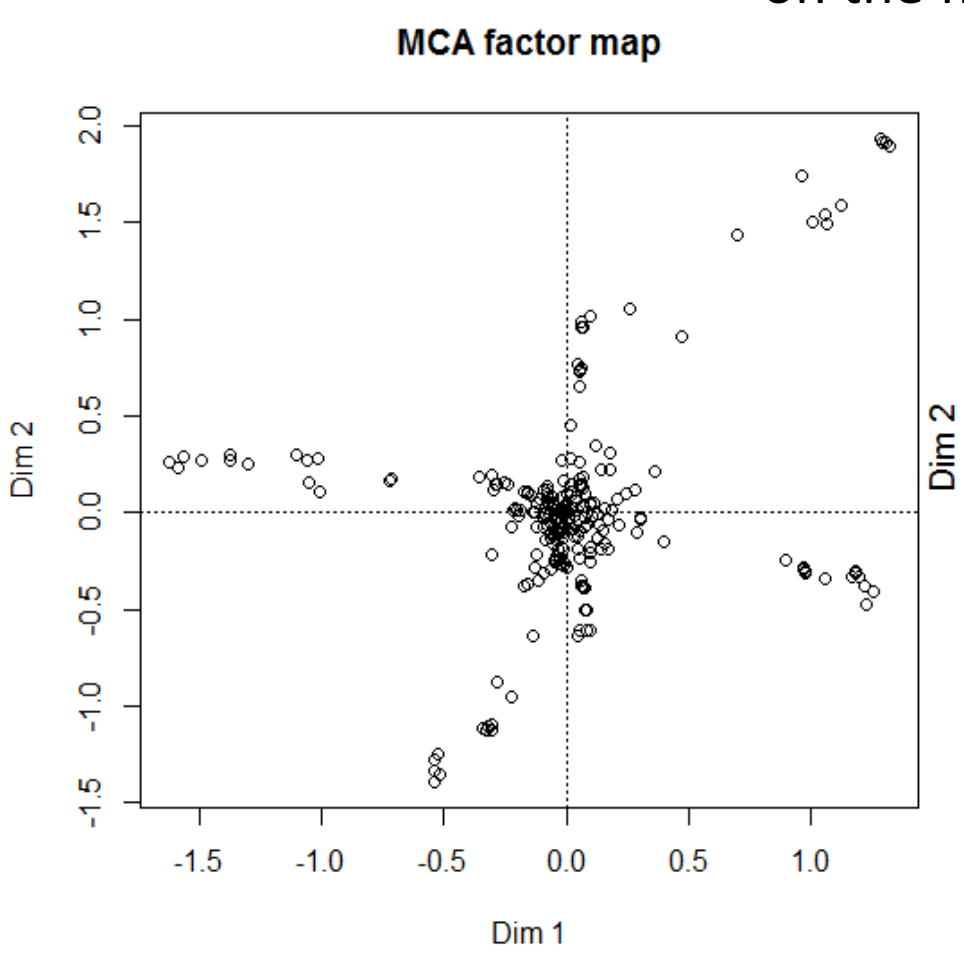
Comparison of the loadings

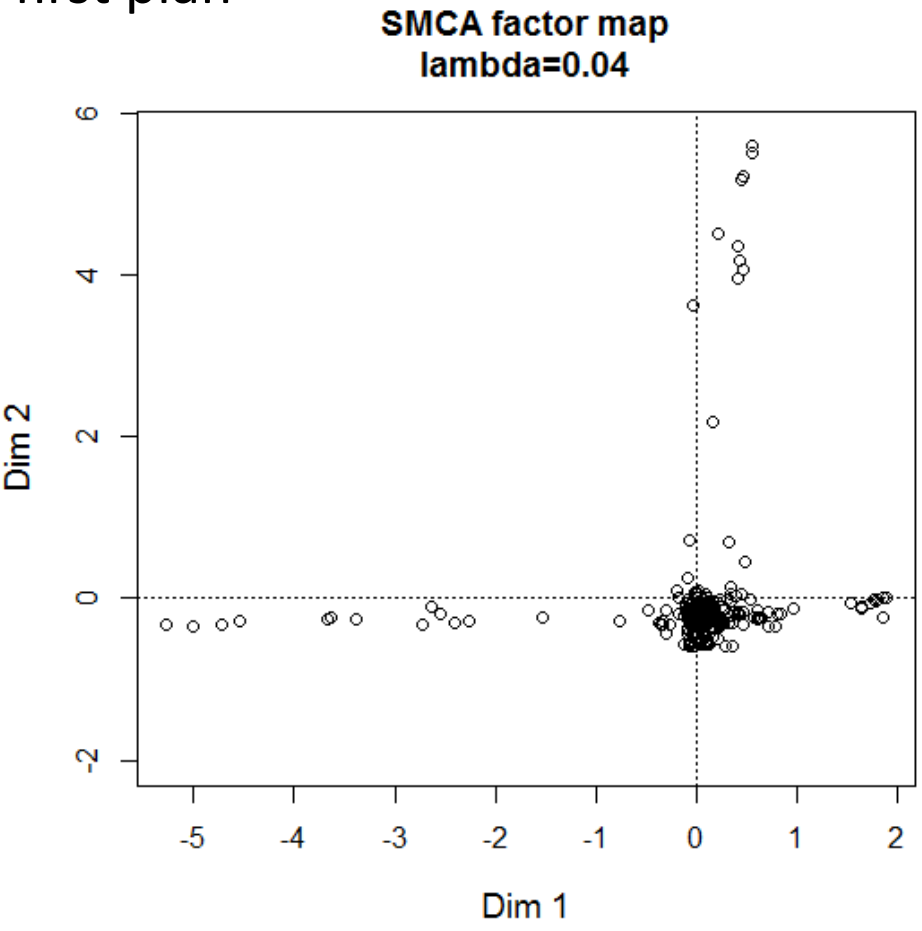
SNPs	MCA		Sparse MCA	
	Dim 1	Dim 2	Dim 1	Dim 2
rs4253711.AA	**-0.323**	-0.043	**-0.309**	**0.000**
rs4253711.AG	**0.009**	0.016	**0.057**	**0.000**
rs4253711.GG	**0.024**	-0.006	**0.086**	**0.000**
rs4253724.AA	**-0.264**	-0.025	**-0.424**	**0.000**
rs4253724.AT	**0.018**	0.014	**0.115**	**0.000**
rs4253724.TT	**0.027**	-0.008	**0.116**	**0.000**
rs26722.AG	0.054	**-0.421**	**0.000**	**-0.574**
rs26722.GG	-0.003	**0.024**	**0.000**	**0.574**
rs35406.AA	-0.002	**0.024**	**0.000**	**0.241**
rs35406.AG	0.038	**-0.388**	**0.000**	**-0.241**
⋮	⋮	⋮	⋮	⋮
#non-zero loadings	281	281	30	24
% inertia	6.86	6.73	5.03	4.95

3. Application on genetic data

Single Nucleotide Polymorphisms

Comparison between MCA and Sparse MCA on the first plan

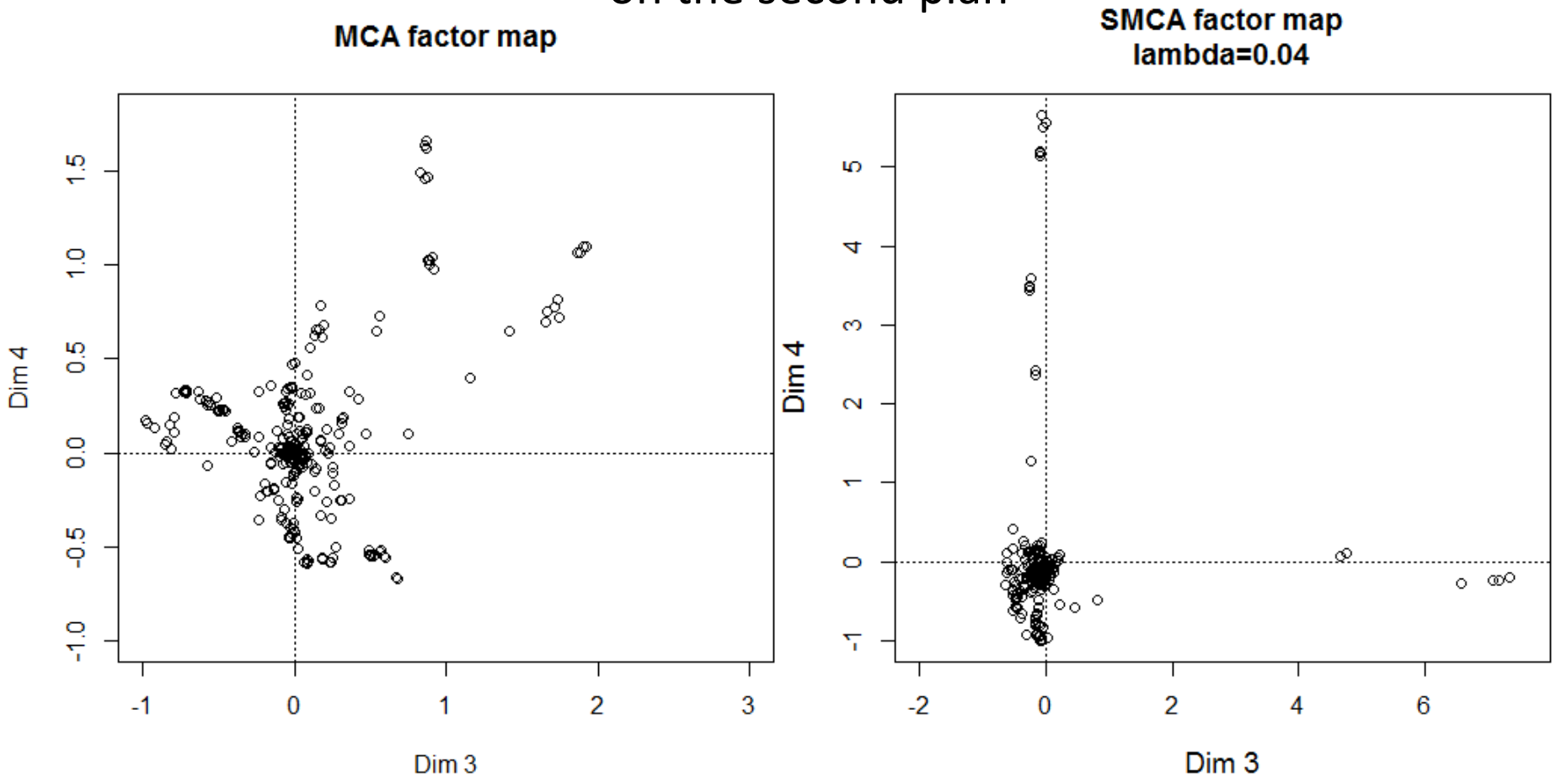




3. Application on genetic data

Single Nucleotide Polymorphisms

Comparison between MCA and Sparse MCA on the second plan



Application on genetic data

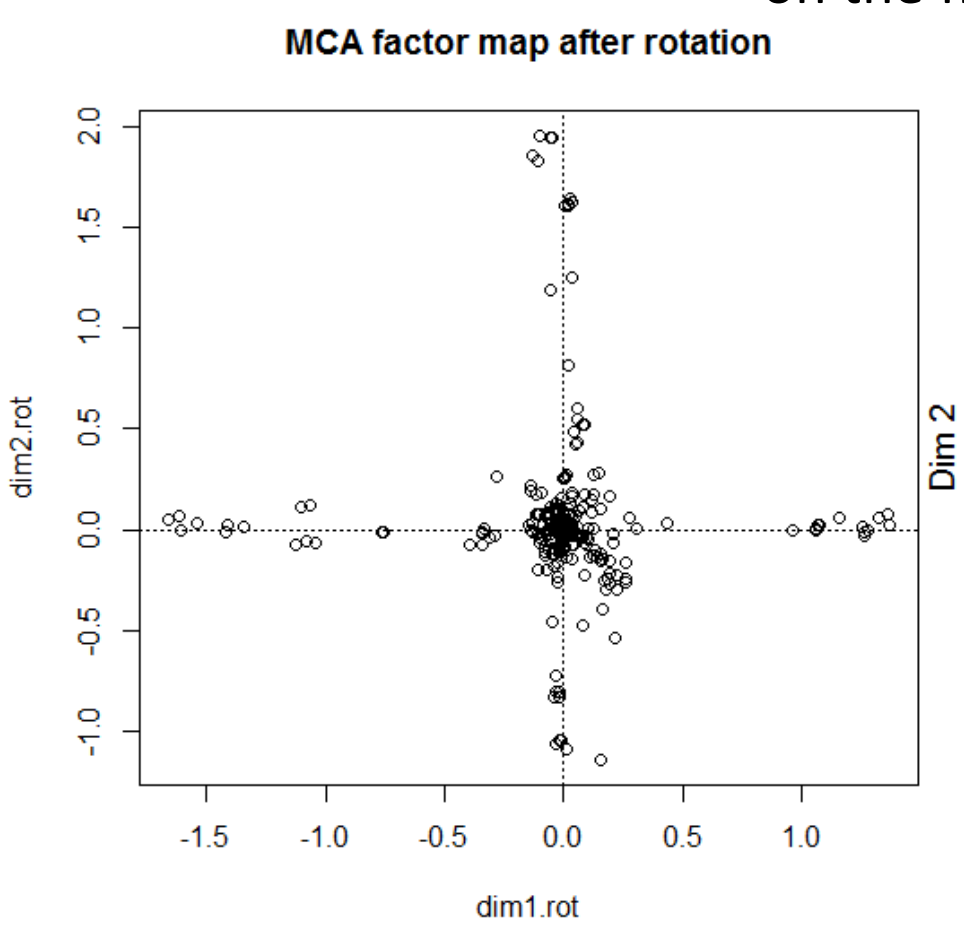
Comparison of the squared loadings

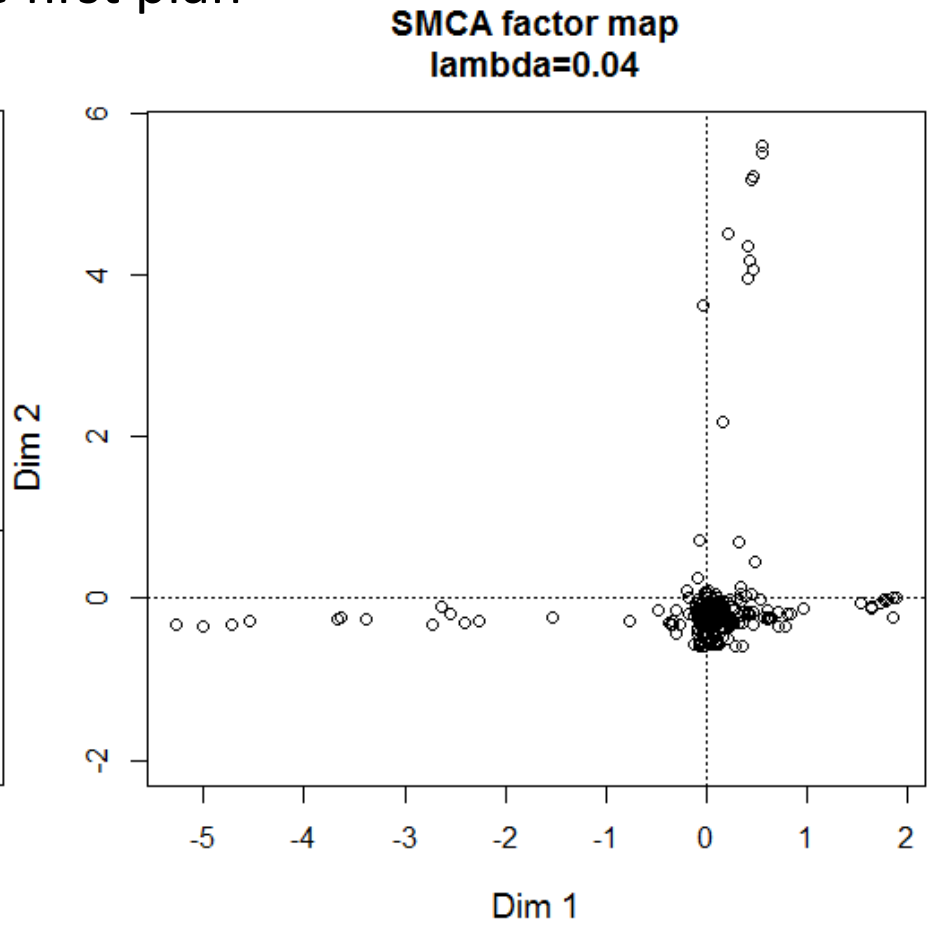
SNPs	MCA		MCA with rotation		Sparse MCA	
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
rs4253711	**0.104**	0.232	0.003	0.003	**0.106**	**0.000**
rs4253724	**0.119**	0.238	0.003	0.002	**0.206**	**0.000**
rs26722	0.001	0.003	0.003	0.003	**0.000**	**0.659**
rs35406	0.001	0.000	0.003	0.005	**0.000**	**0.115**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
#of non-zero loadings	281	281	281	281	30	24
% inertia	6.86	6.73	6.73	6.46	5.03	4.95

3. Application on genetic data

Single Nucleotide Polymorphisms

Comparison between MCA with rotation and sparse MCA on the first plan

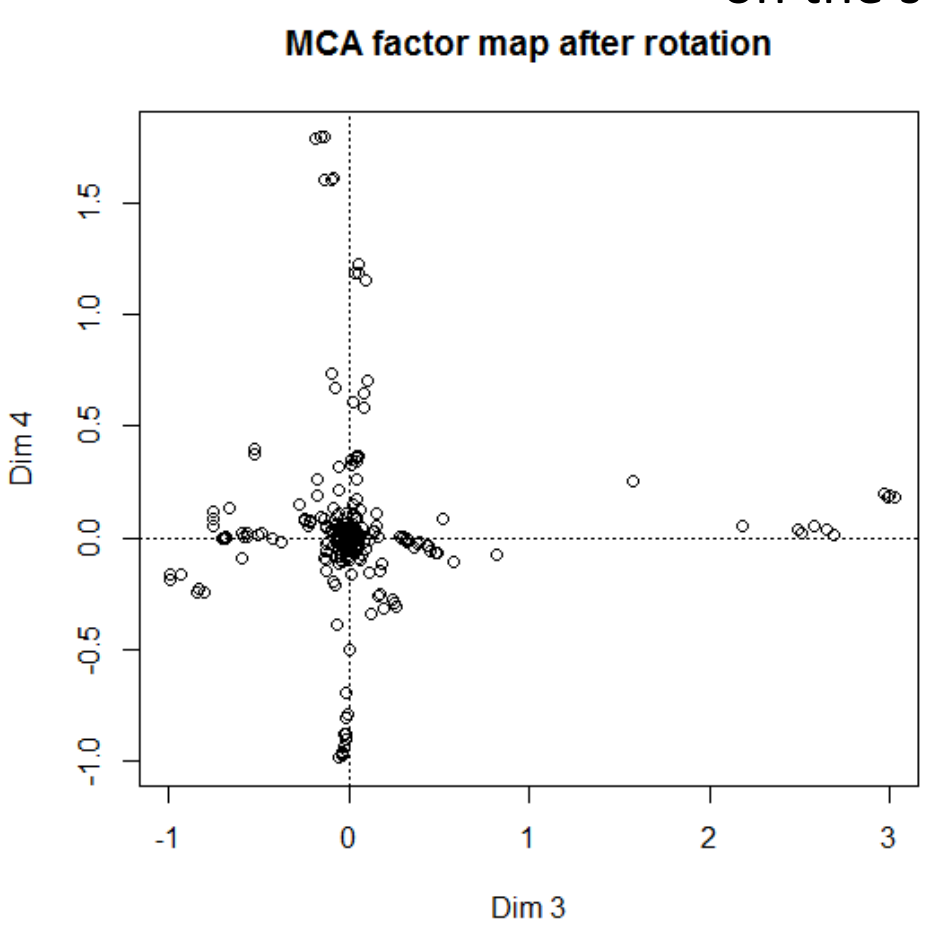


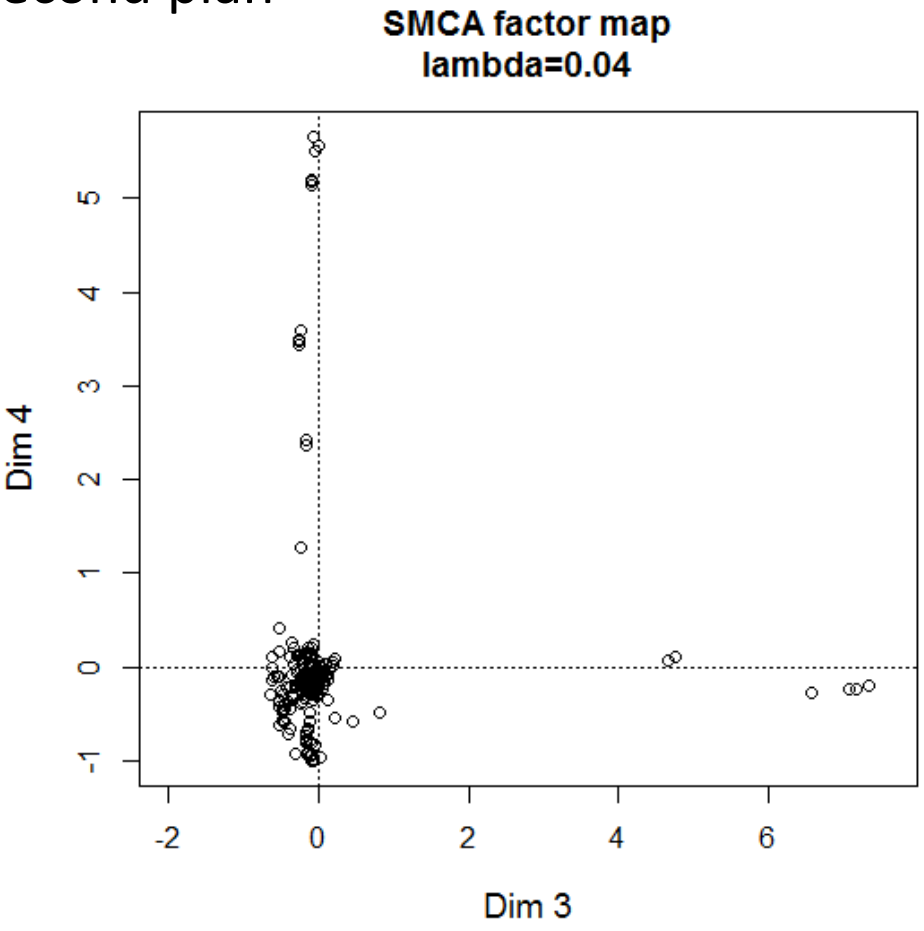


3. Application on genetic data

Single Nucleotide Polymorphisms

Comparison between MCA with rotation and sparse MCA on the second plan





Conclusions and perspectives

- We proposed 2 new methods in a unsupervised multiblock data context: **Group Sparse PCA** for continuous variables, and **Sparse MCA** for categorical variables
- Both methods produce sparse loadings structures that makes easier the interpretation and the comprehension of the results

However these methods do not yield sparsity within groups

Research in progress:

- Criteria for choosing the tuning parameter λ
- Extension of Sparse MCA
  - To select groups and predictors within a group, in order to produce sparsity at both levels
  - A compromise between the Sparse MCA and the sparse group lasso developed by Simon et al. (2002)

References

Chavent, M., Kuentz-Simonet, V., and Saracco, J. (2012). Orthogonal rotation in PCAMIX. *Advances in Data Analysis and Classification*, **6** (2), 131-146.

Jolliffe, I.T. , Trendafilov, N.T. and Uddin, M. (2003) A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics,* **12,** 531–547,

Rousson, V. , Gasser, T. (2004), Simple component analysis. *Journal of the Royal Statistical Society*: *Series C (Applied Statistics),* **53**,539-555

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2012) A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics,*

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288,

References

Van de Velden, M., Kiers, H. (2005) Rotation in Correspondence Analysis, *Journal of Classification*,22, 2, 251-271

Vines, S.K., (2000) Simple principal components, *Journal of the Royal Statistical Society*: *Series C (Applied Statistics)*, **49**, 441-451

Yuan, M., Lin, Y. (2007) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49-67,

Zou, H., Hastie , T. (2005*)* Regularization and variable selection via the elastic net. *Journal of Computational and Graphical Statistics*, **67**, 301-320,

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286.