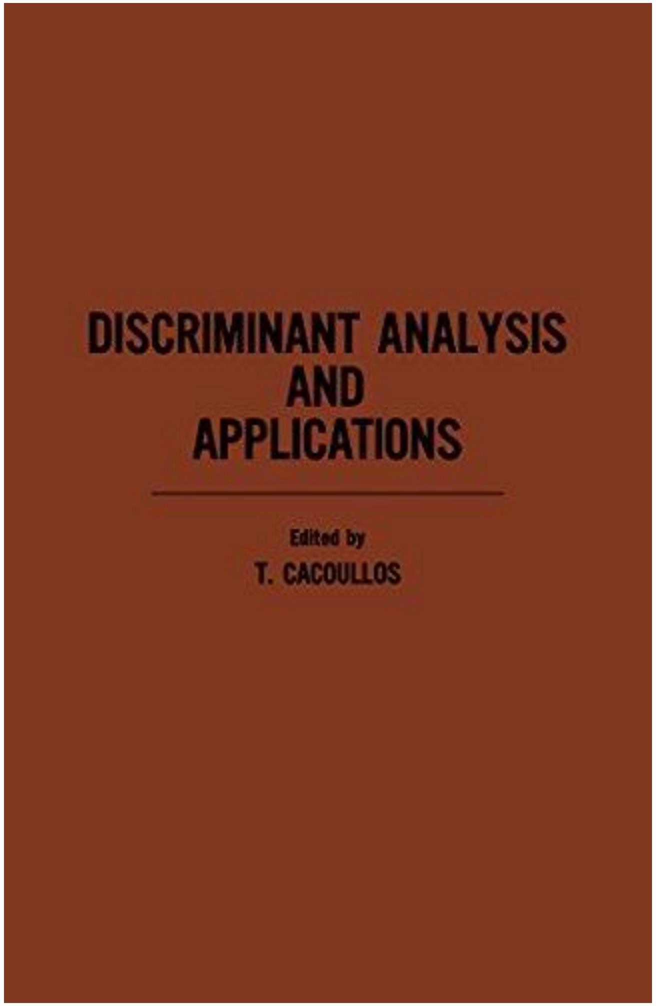
Sparse Correspondence Analysis for Contingency Tables

Gilbert Saporta
CEDRIC,
Conservatoire National des Arts et Métiers, Paris
gilbert.saporta@cnam.fr



Proceedings of the NATO Advanced Study Institute on Discriminant Analysis and Applications held in Kifissia, Athens, Greece in June 1972.

Academic Press New York and London 1973
A Subsidiary of Harcourt Brace Jovanovich, Publishers

- Summer school at Mamaia, Black Sea, (Romania) July 1993

Outline

1. Introduction
2. Reminders on sparse PCA
3. Sparse CA
4. Conclusion and perspectives

A joint work with:

Liu Ruiping (Beihang University, Beijing)

Ndeye Niang (CNAM, Paris)

Wang Huiwen (Beihang University, Beijing)

1.Introduction

- Correspondence Analysis of contingency tables (CA) is both:
  - a double PCA
  - a generalized SVD

  with weights and the chi squared metric
- Doubly sparse CA: an application of sparse SVD
- Sparse PCA of row profiles only leads to column sparse CA. Useful for contingency tables with many columns like documents-terms matrix

2.Reminders on sparse PCA

- In **PCA**, each PC is a linear combination of **all** the original variables : difficult to interpret the results for large p.
- **Objective of sPCA:** obtain pseudo components easily interpretable as combinations of only a few variables. Most coefficients (weights) should be equal to zero.

2.1 First attempts:

- **Simple PCA**
  - Hausman (1982) weights -1,0,1
  - Vines (2000) : integer weights
  - Generalized by Rousson, V. and Gasser, T. (2004) : blocks of weights (+ , 0, -)

Hausman, Robert E., Jr. (1982) Constrained multivariate analysis. In S.H. Zanakis, Jagdish S. Rustagi eds, *Optimization in statistics: With a view towards applications in management science and operations research,* TIMS Stud. Management Sci., 19, 137–151, *North-Holland, Amsterdam,* 1982

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

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

2.2 SCoTLASS (Simplified Component Technique – Lasso) by Jolliffe & al. (2003) : extra L_1 constraints

$$\max \mathbf{v}'\mathbf{V}\mathbf{v} \;\text{ with }\; \|\mathbf{v}\|^2 = 1 \text{ and } \|\mathbf{v}\|_1 = \sum_{j=1}^{p} |v_j| \le \tau$$

$$1 < \tau < \sqrt{p}$$

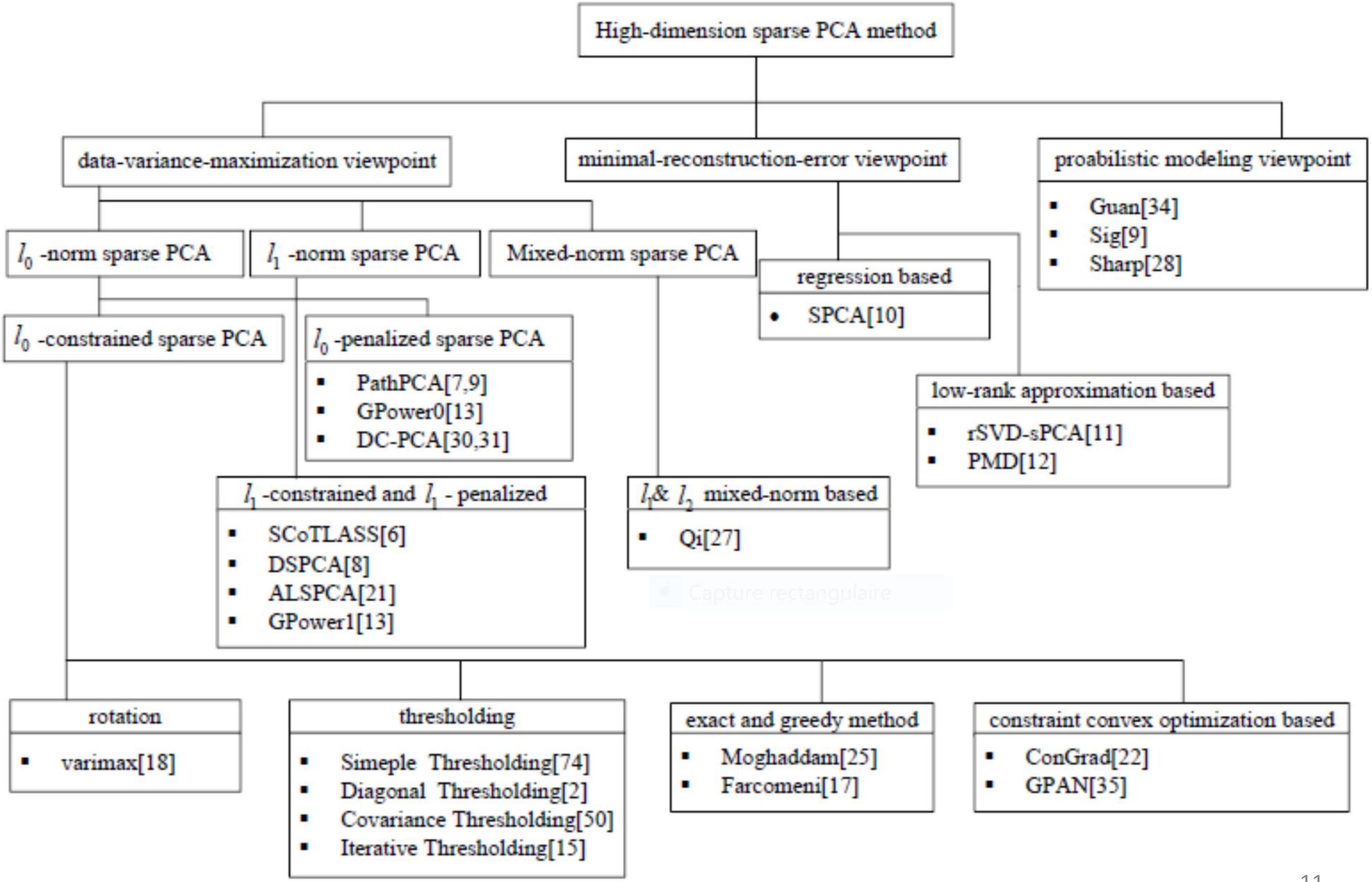
$$\tau \ge \sqrt{p} \;\text{ usual PCA}$$

$$\tau < 1 \text{ no solution}$$

$$\tau = 1 \text{ only one nonzero coefficient}$$

2.3 More than 20 variants

Shen & Li, 2015



2.4 Sparse SVD

- A rank 1 sparse SVD or Penalized Matrix Decomposition (Witten et al, 2009):

$$\min \|\mathbf{X} - d\mathbf{u}\mathbf{v}'\|_F^2 \text{ subject to } \|\mathbf{u}\|^2 = \|\mathbf{v}\|^2 = 1,$$

$$\text{and } \sum_{i=1}^{I} |u_i| \le \alpha, \sum_{j=1}^{J} |v_j| \le \beta, \quad d \ge 0$$

- Equivalent formulation:

$$\max \mathbf{u}'\mathbf{X}\mathbf{v} \text{ subject to } \|\mathbf{u}\|^2 \le 1, \|\mathbf{v}\|^2 \le 1,$$

$$\sum_{i=1}^{I} |u_i| \le \alpha, \sum_{j=1}^{J} |v_j| \le \beta$$

2.5 Lost properties and issues

- Sparse PCA does not provide a global selection of variables but a selection **dimension by dimension** : different from the regression context (Lasso, Elastic Net, …)
- SCoTLASS: orthogonal factors but correlated components
- Usually: neither factors, nor components are orthogonal
  - Necessity of adjusting the % of explained variance
- No clear criterium like R^2 or MSE to choose the tuning parameters *ie* the degree of sparsity.

- Deflation in SVD
  - Usual solution: repeat the penalized decomposition for $\mathbf{X}\text{-}d\mathbf{uv}'$ (Hotelling's deflation) but the solution is not orthogonal to the rank one matrix $\mathbf{uv}'$.
  - **Projected PMD** provides an almost orthogonal solution:

replace $\mathbf{X}$ by $(\mathbf{I}\text{-}\mathbf{uu}')\mathbf{X}(\mathbf{I}\text{-}\mathbf{vv}')$

3.Sparse Correspondence Analysis

3.1 Standard correspondence analysis

- For a contingency table $\mathbf{N}$, CA is
  - a double PCA
  - A weighted SVD of centered $\mathbf{P}=\mathbf{N}/n$

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{rc}'\right)\mathbf{D}_c^{-1/2}$$

$$\frac{p_{ij} - p_i p_j}{\sqrt{p_i p_j}}$$

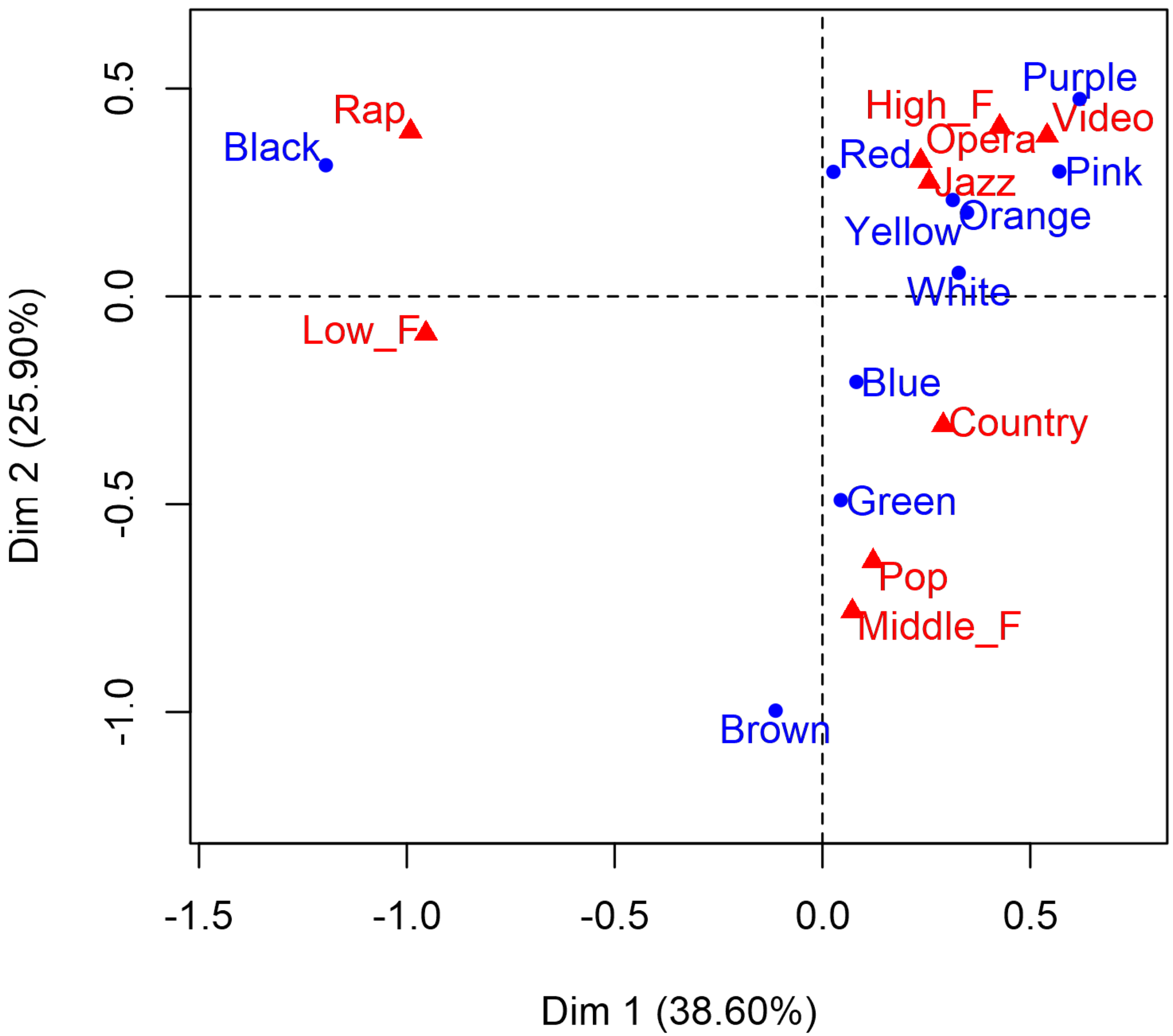
$\mathbf{r}$, $\mathbf{c}$ vectors of marginal distributions

3.2 A toy example: colours of sound

Color	Video	Jazz	Country	Rap	Pop	Opera	Low F	High F	Middle F	x_{i+}	**r**
Red	4	2	4	4	1	2	2	4	1	24	0.121
Orange	3	4	2	2	1	1	0	3	2	18	0.091
Yellow	6	4	5	2	3	1	1	3	0	25	0.126
Green	2	0	5	1	3	3	3	1	5	23	0.116
Blue	2	5	0	1	4	1	2	1	3	19	0.096
Purple	3	3	1	0	0	3	0	2	1	13	0.066
White	0	0	0	0	1	4	1	5	3	14	0.071
Black	0	2	0	11	1	3	10	1	1	29	0.146
Pink	2	1	1	0	2	4	0	2	0	12	0.061
Brown	0	1	4	1	6	0	3	0	6	21	0.106
x_{+j}	22	22	22	22	22	22	22	22	22	$N = 198$	1.000
$\mathbf{c}^T$	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11		

Abdi H., Béra M. (2017) Correspondence Analysis
In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining,
Springer, New York,

CA factor map



CA rows and columns coordinates and contributions

	a1	a2	ctr1	ctr2		b1	b2	ctr1	ctr2
Red	0.026	0.299	0	56	Video	0.541	0.386	113	86
Orange	0.314	0.232	31	25	Jazz	0.257	0.275	25	44
Yellow	0.348	0.202	53	27	Country	0.291	-0.309	33	55
Green	0.044	-0.490	1	144	Rap	-0.991	0.397	379	91
Blue	0.082	-0.206	2	21	Pop	0.122	-0.637	6	234
Purple	0.619	0.475	87	77	Opera	0.236	0.326	22	61
White	0.328	0.057	26	1	LowF	-0.954	-0.089	351	5
Black	-1.195	0.315	726	75	HighF	0.427	0.408	70	96
Pink	0.570	0.300	68	28	MiddleF	0.072	-0.757	2	330
Brown	-0.113	-0.997	5	545					
					total			1000	1000
total			1000	1000					

Both sides sparse CA through sparse SVD

- sumabsu = $\sum_{i=1}^{I} |u_i|$
- sumabsv = $\sum_{j=1}^{J} |v_j|$

- The smaller they are, the sparser $\mathbf{u}$ or $\mathbf{v}$ will be. Need for a compromise between sparseness and fit

Criteria:

- $$BIC(\tau) = \frac{\left\| \mathbf{X} - \hat{\mathbf{X}} \right\|^2}{np\hat{\sigma}^2} + \frac{\ln(np)}{np} df(\tau)$$

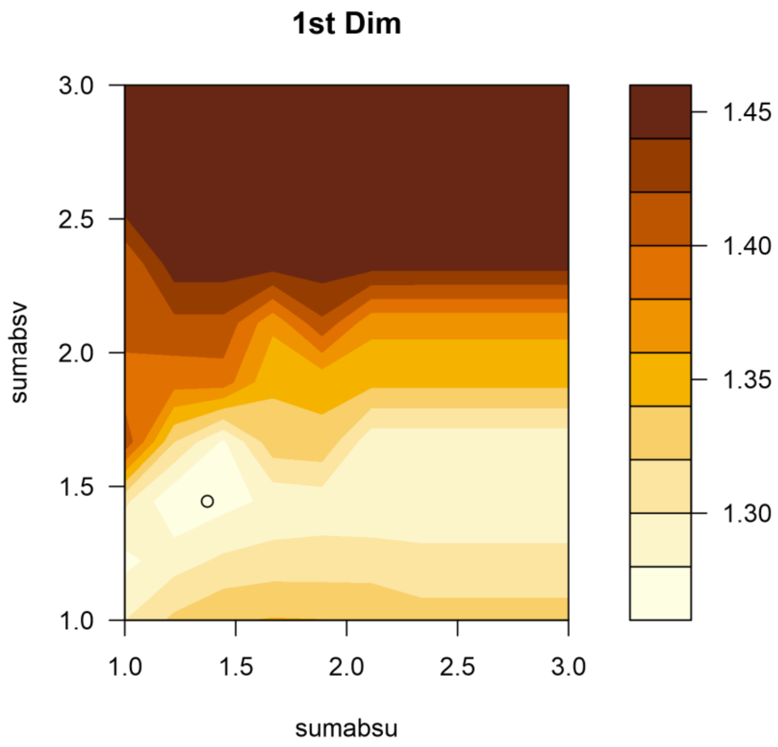
  Zou et al. (2007), Shen et al. (2013)

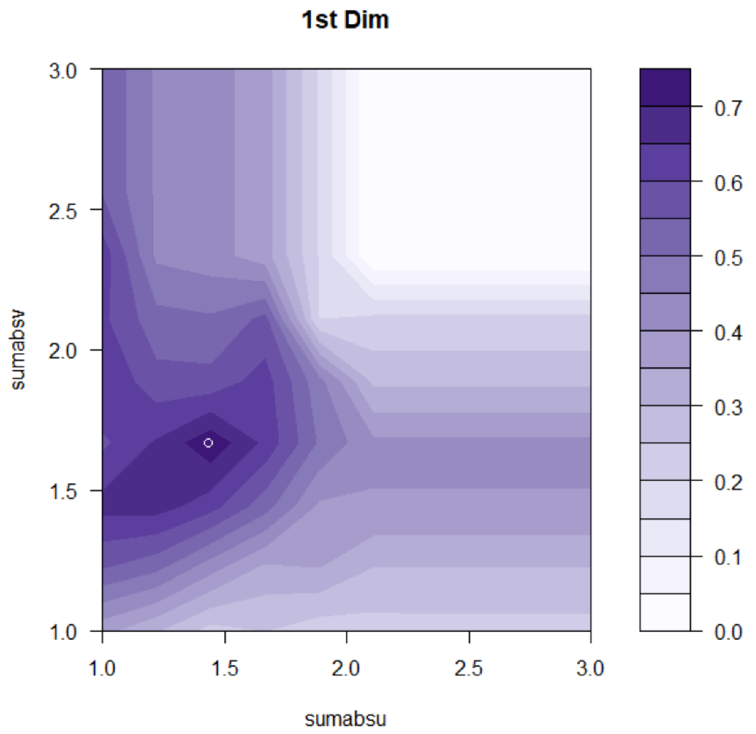
- Index of sparseness derived from Trendafilov (2014)

$$IS = \frac{V_a V_s}{V_0^2} \frac{\#0}{pr}$$

V_a, V_s and V_o are the adjusted, unadjusted and ordinary total variances for the problem, and #0 is the number of zero loadings with r components

Simultaneous optimization: first dimension



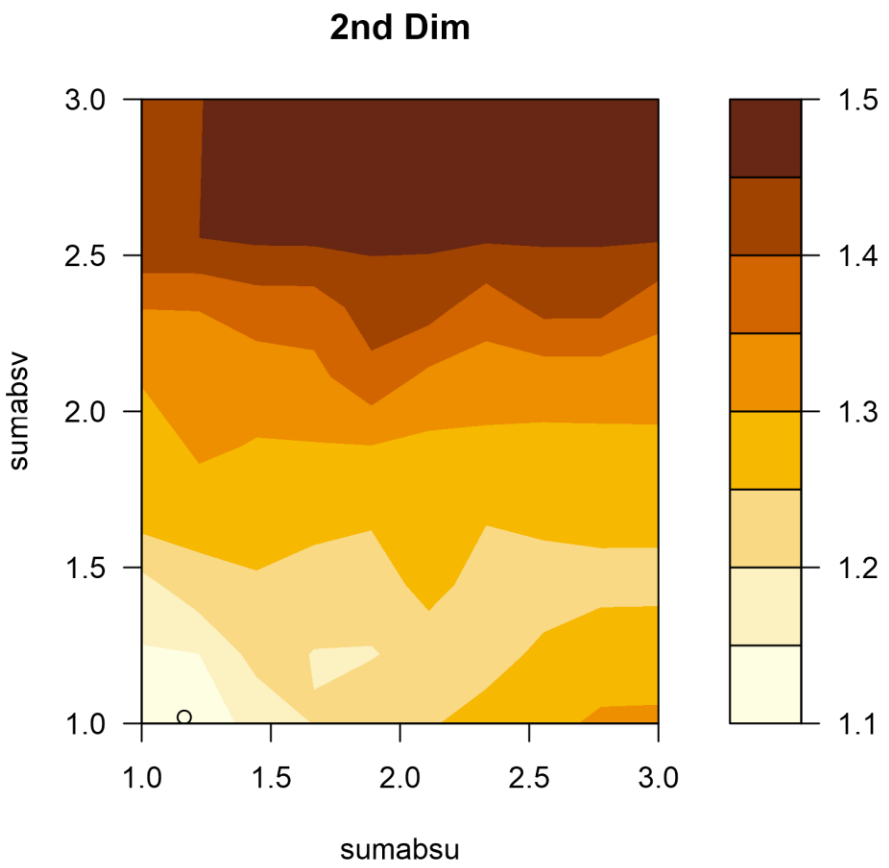


BIC

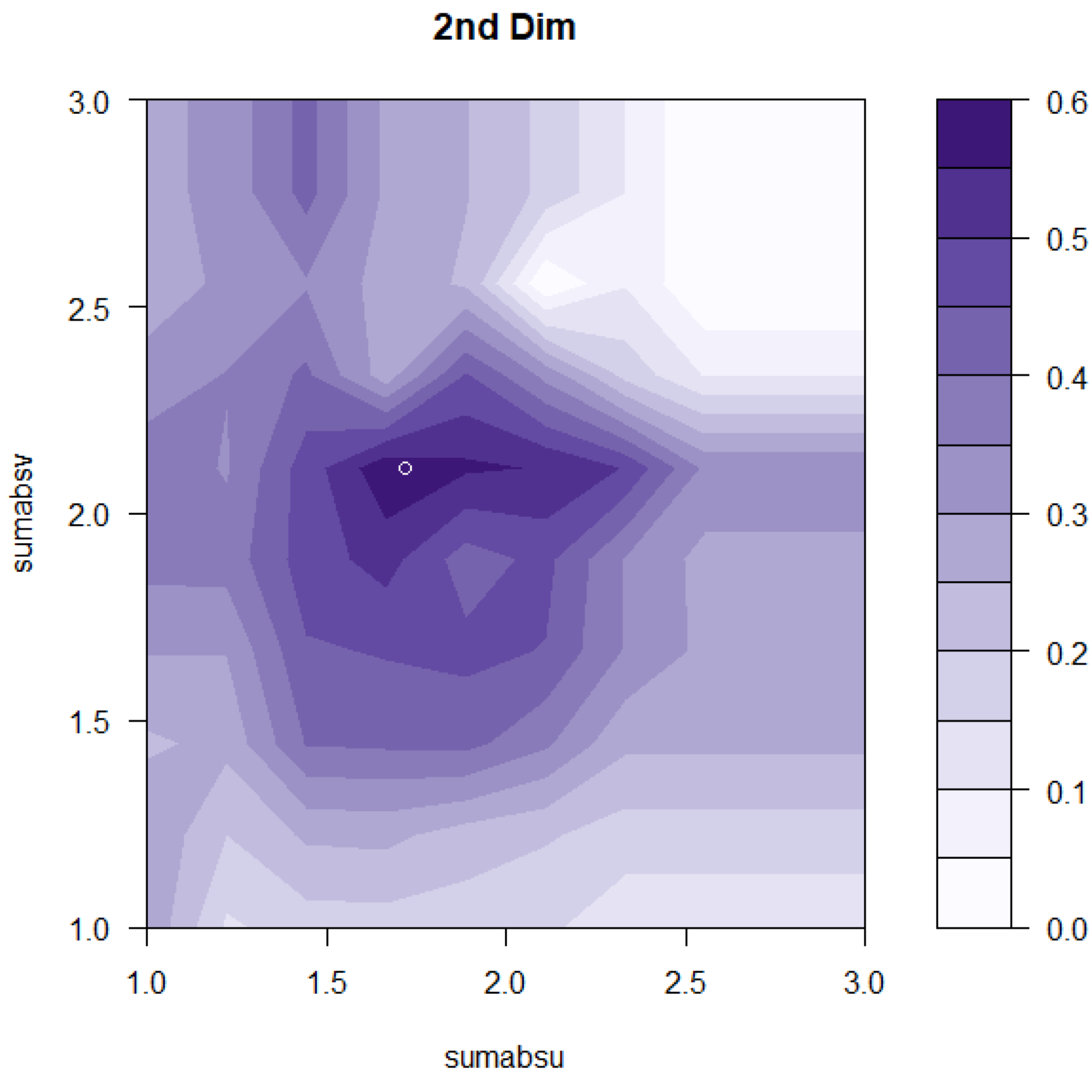
IS

Second dimension

- BIC fails to give an acceptable solution

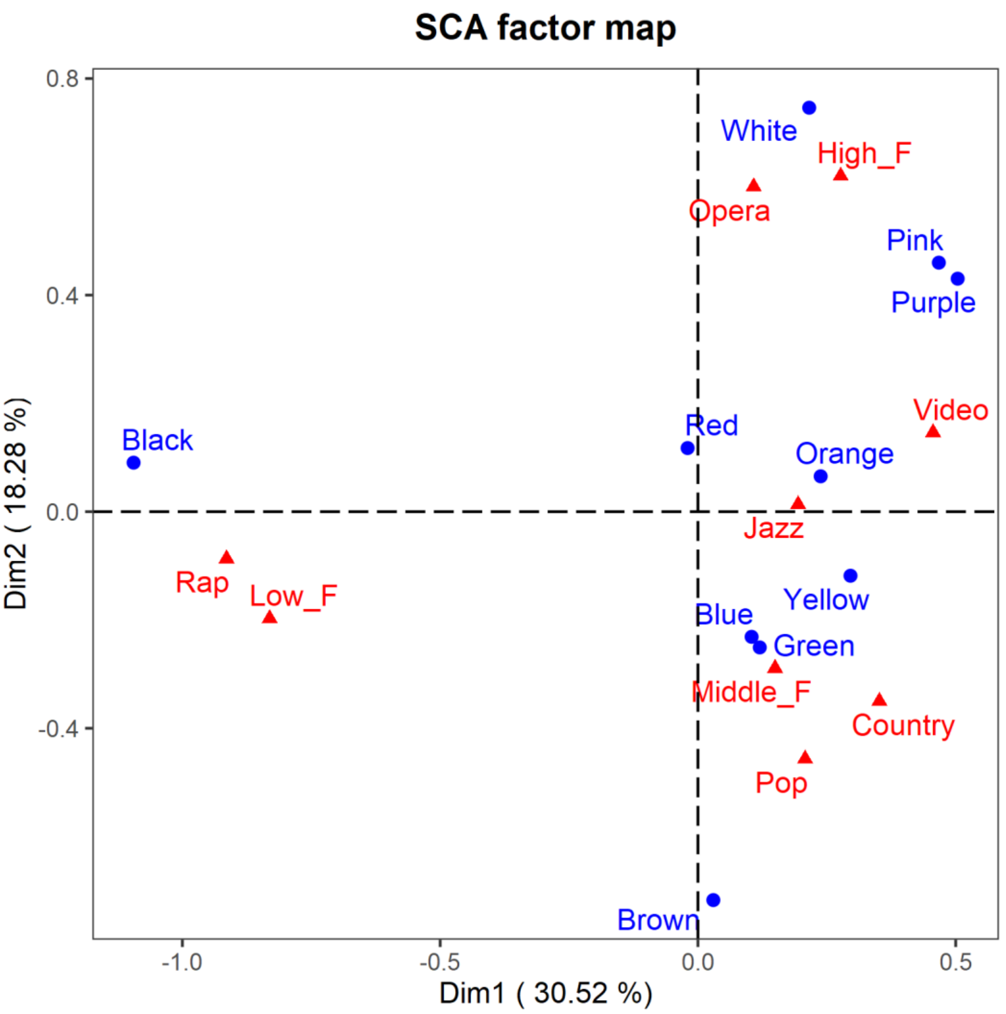


While IS does



Sparse CA

	u1	u2	ctr1	ctr2	a1	a2
Red	0	0	0	0	-0.020	0.117
Orange	0.047	0	1	0	0.238	0.065
Yellow	0.161	0	14	0	0.296	-0.118
Green	0	0	0	0	0.120	-0.251
Blue	0	0	0	0	0.104	-0.231
Purple	0.343	0.201	34	19	0.504	0.430
White	0	0.832	0	358	0.216	0.746
Black	-1.202	0	929	0	-1.095	0.091
Pink	0.284	0.233	21	24	0.467	0.460
Brown	0	-0.877	0	598	0.030	-0.717
	v1	v2	ctr1	ctr2	b1	b2
Video	0.295	0	42	0	0.456	0.146
Jazz	0	0	0	0	0.194	0.013
Country	0.122	-0.345	7	97	0.352	-0.350
Rap	-1.054	0	542	0	-0.914	-0.087
Pop	0	-0.466	0	177	0.208	-0.457
Opera	0	0.639	0	333	0.108	0.600
LowF	-0.915	0	409	0	-0.831	-0.197
HighF	0	0.654	0	349	0.277	0.620
MiddleF	0	-0.235	0	45	0.150	-0.289



- Percentages of explained variance are a little smaller than in the standard CA
- Graphical displays look very similar
- Low contributions have been set to zero, while high contributions are enlighted

  - Weight vectors nearly orthogonal :

    $<u1; u2 > = 0.0085$ and $< v1; v2 > = 0.0047$

  - Coordinates vectors nearly orthogonal:

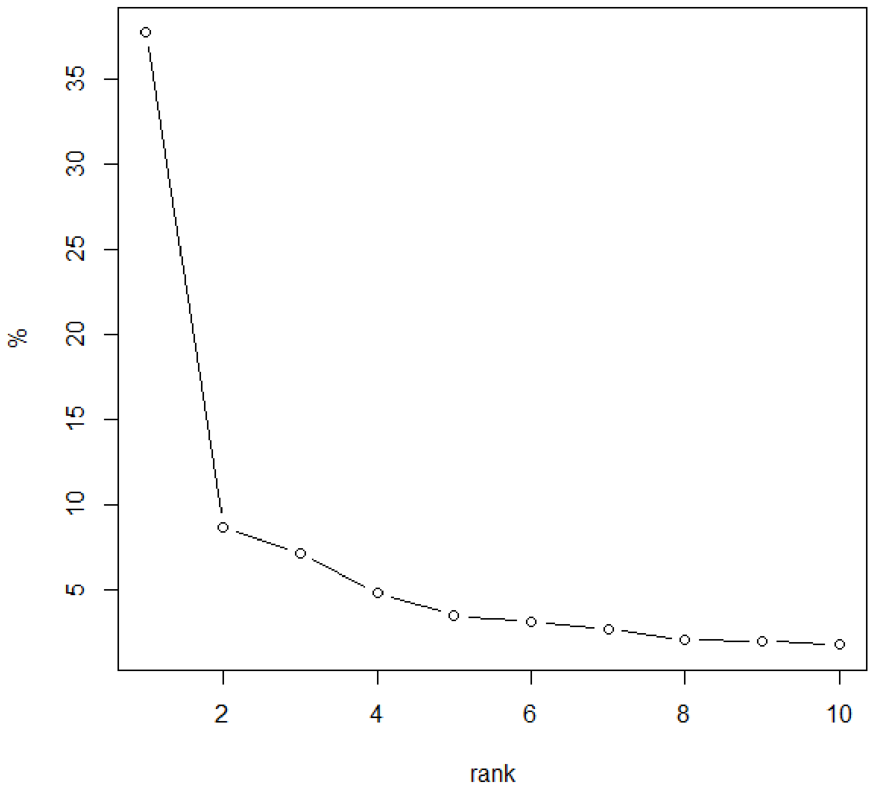
    $<a1; a2 >= 0.0128$ and $< b1; b2 > = 0.0320$

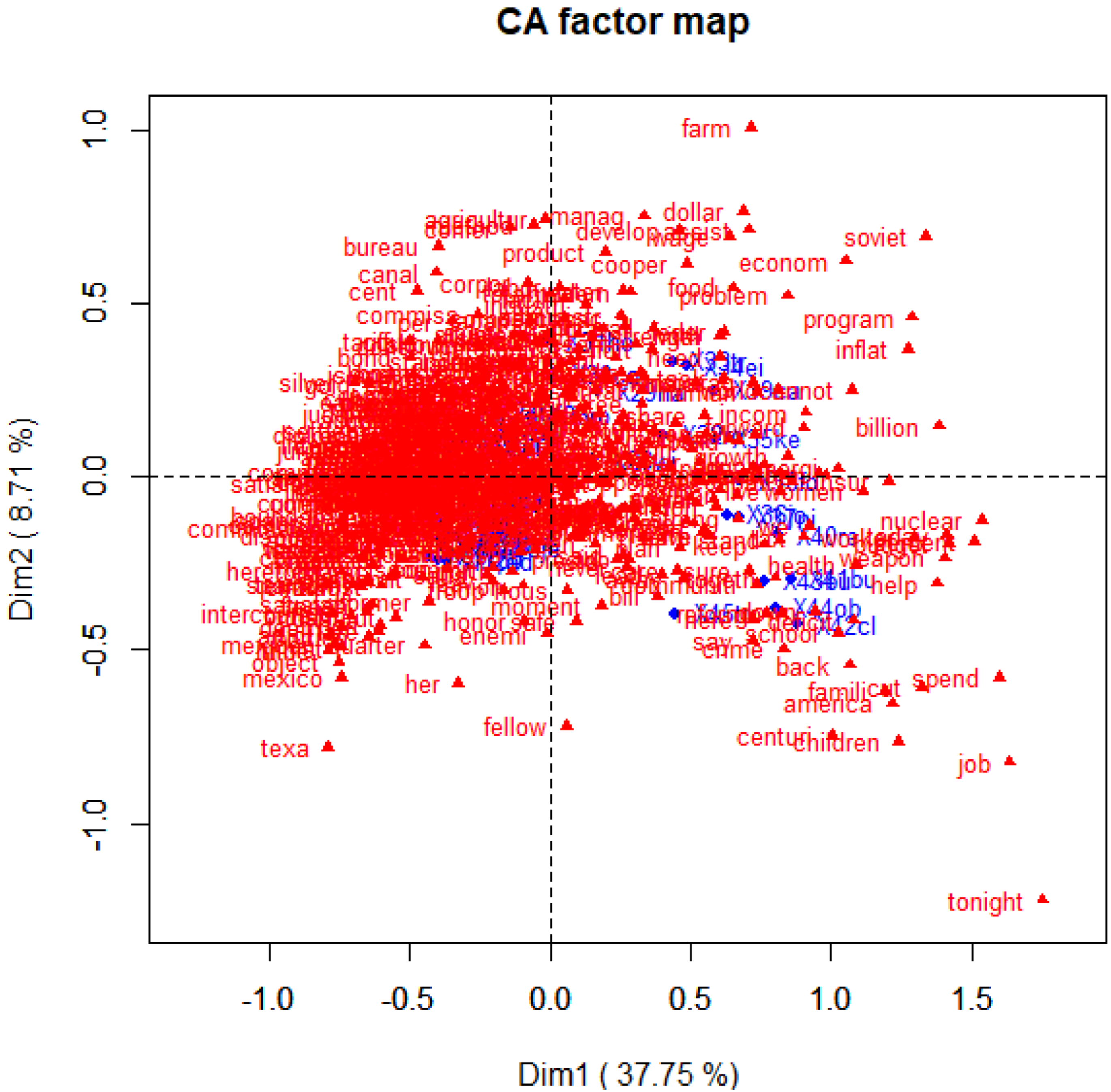
3.4 Textual data

- State of the Union Addresses
  - speeches of 43* presidents of the United States (from G.Washington to D.Trump). The data set contains 934 high-frequency words that appear more than 220 times in the speeches.
  - Preprocessing reduces the number of words to 572

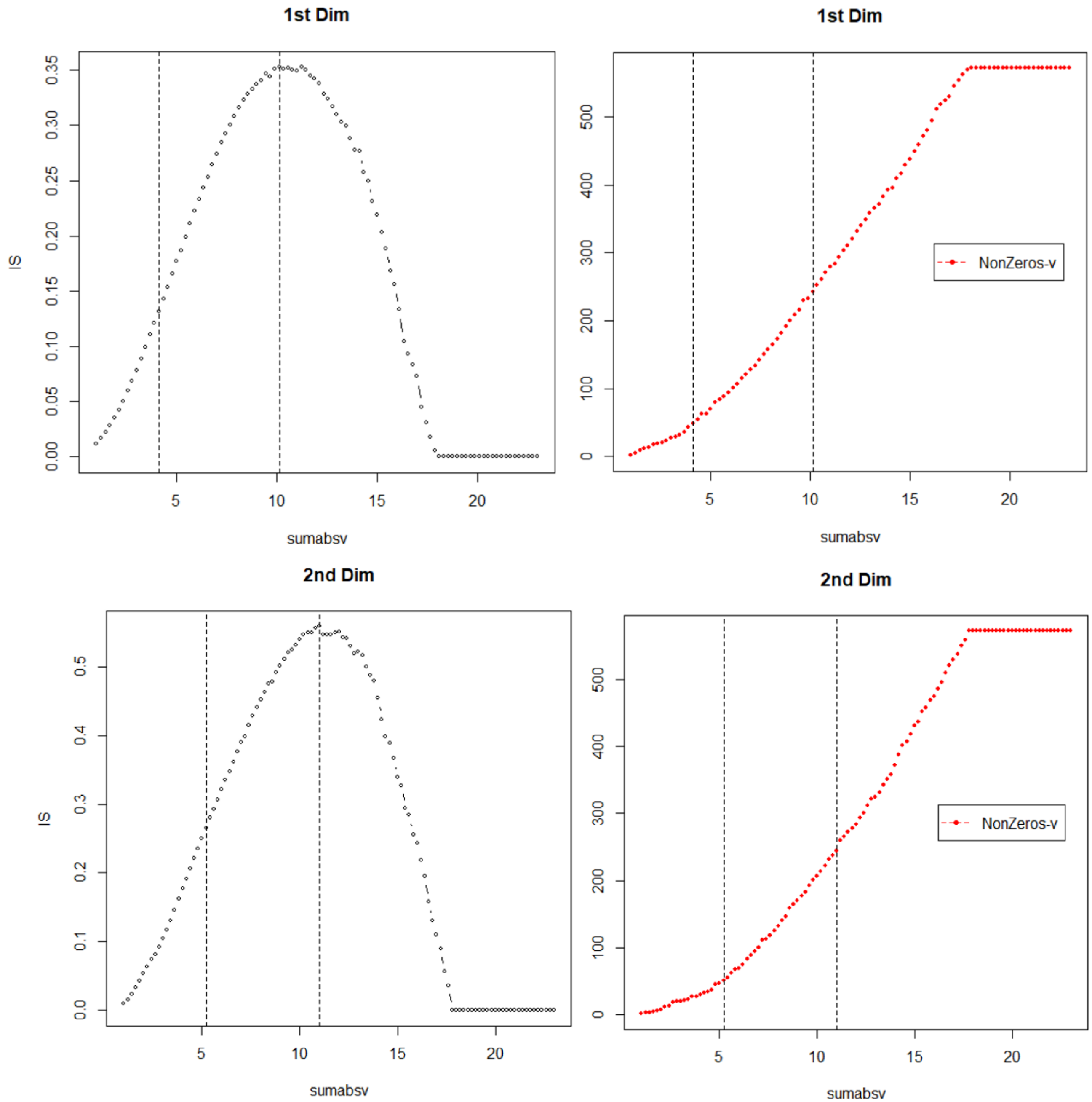
* Some speeches are missing

Scree plot of eigenvalues




CA factor map
Dim1 (37.75 %)
Dim2 (8.71 %)
-1.0
-0.5
0.0
0.5
1.0
1.5
farm
dollar
soviet
bureau
product
cooper
econom
canal
cent
program
inflat
billion
nuclear
weapon
health
help
spend
back
america
centuri
children
job
tonight
texa
fellow
her
mexico
object
honor
enemi
moment
school
crime
incom
growth
women
keep

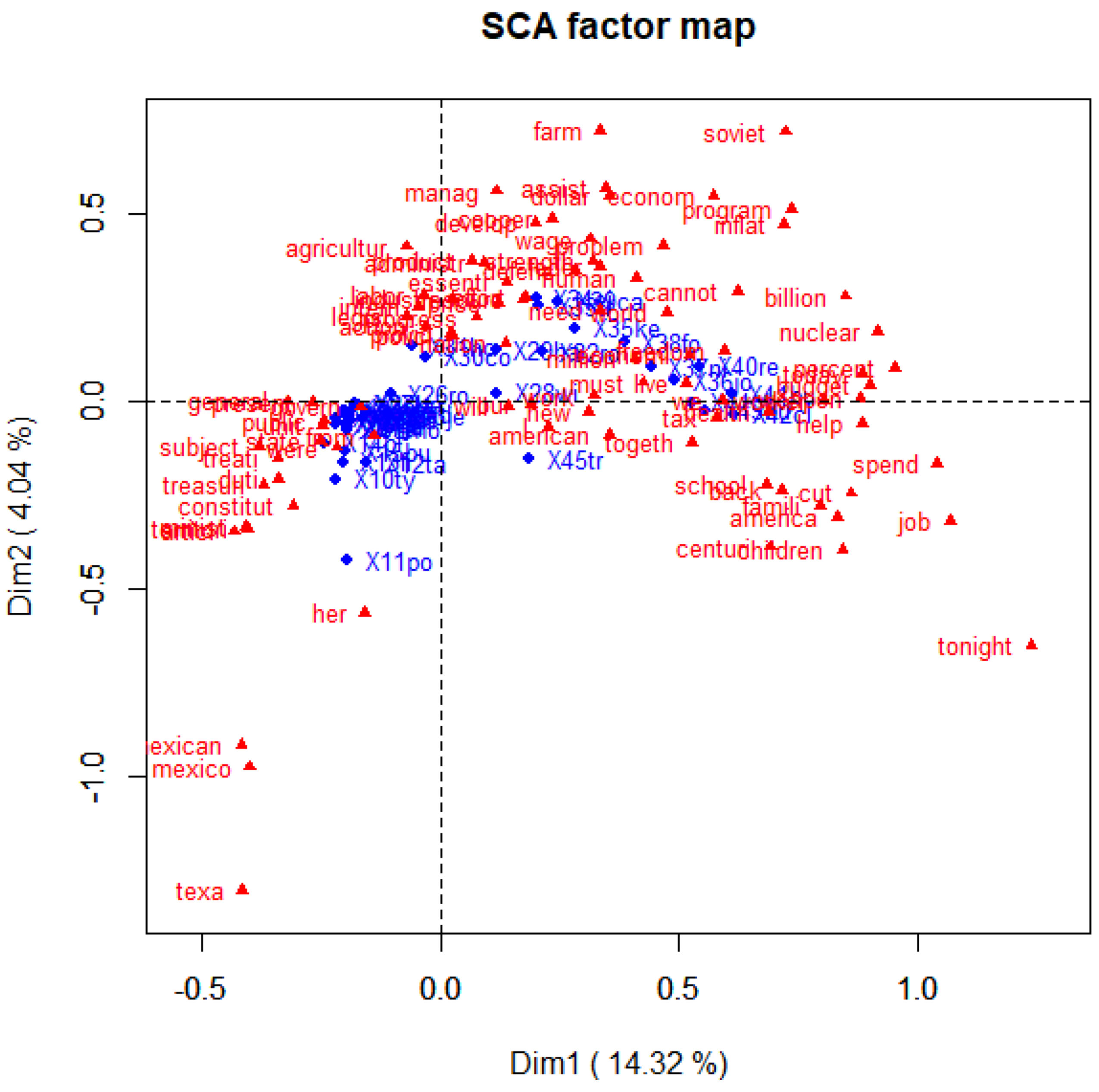
- One side sparse CA
  - Sparsify columns (words) not rows (presidents)
  - No constraints on $\sum_{i=1}^{I} |u_i|$
  - Grid search for IS as a function of $\sum_{j=1}^{J} |v_j|$



Optimal values of sumabsv gives too many non-zero weights.

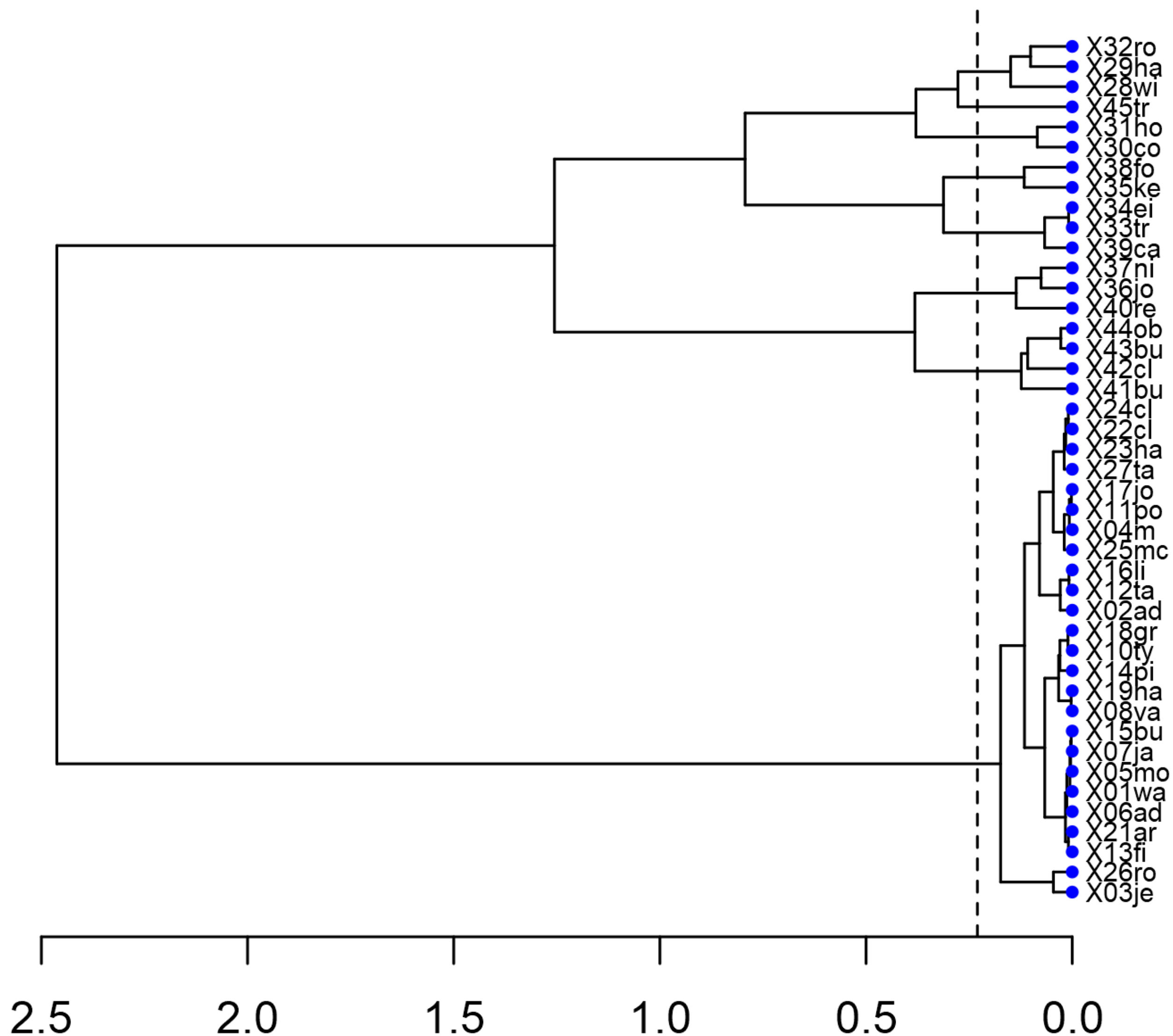
Our choice:
50 non zero weights

Sparse CA-Oneside

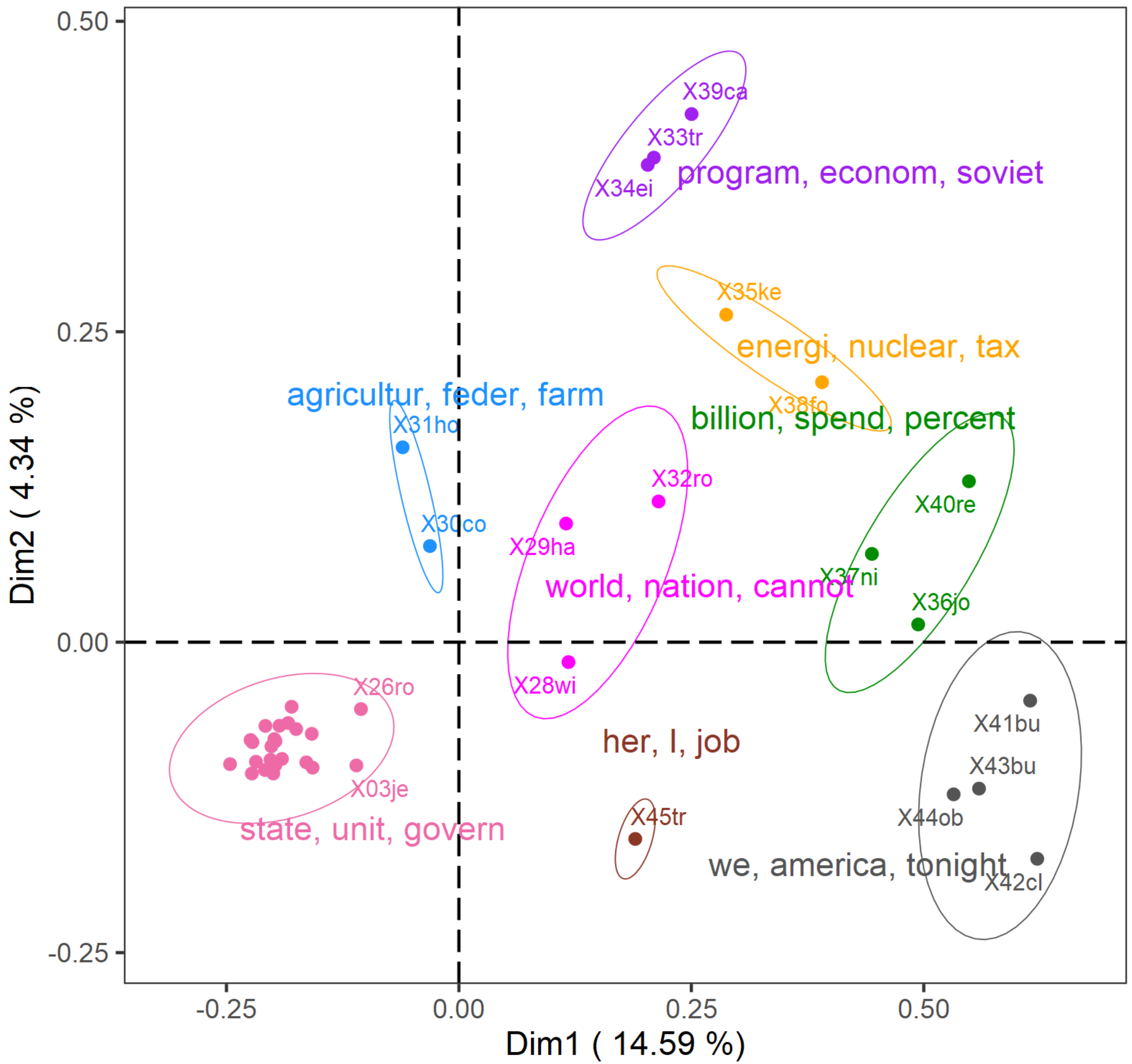


Cluster dendrogram

8 clusters



SCA factor map



4. Conclusions and perspectives

- Sparse methods meet the challenge of high dimensional data and makes interpretation easier.
- Sparse correspondence analysis useful for large contingency tables
- Future works
  - Packaging sparse CA in R
  - Sparsify non symmetric correspondence analysis

Preprint

Cornell University

arXiv.org > stat > arXiv:2012.04271

Statistics > Methodology

[Submitted on 8 Dec 2020]

Sparse Correspondence Analysis for Contingency Tables

Ruiping Liu, Ndeye Niang, Gilbert Saporta, Huiwen Wang

References

- Adachi, K., Trendafilov, N.T. (2015) : Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics* **31**, 1–25
- Guillemot V, Beaton D, Gloaguen A, Löfstedt T, Levine B, Raymond N, et al. (2019) A constrained singular value decomposition method that integrates sparsity and orthogonality. *PLoS ONE* 14(3). https://doi.org/10.1371/journal.pone.0211463
- Jolliffe, I.T., Trendafilov, N.T. , Uddin, M. (2003) A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**, 531–547
- Shen, N., Li, J. (2015) A Literature Survey on High-Dimensional Sparse Principal Component Analysis *International Journal of Database Theory and Application,* **8**, 6, 57-74
- Trendafilov, N.T. (2014) . From simple structure to sparse components: a review. *Computational Statistics*, 2**9,** 431–454.
- Witten, D.M., Tibshirani, R., Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation. *Biostatistics* **10,** 515–534
- Zou, H., Hastie, T. , Tibshirani, R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286.