Analyse des correspondances Taxi

Librairie « TaxicabCA »

Vartan Choulakian
Jacques Allard

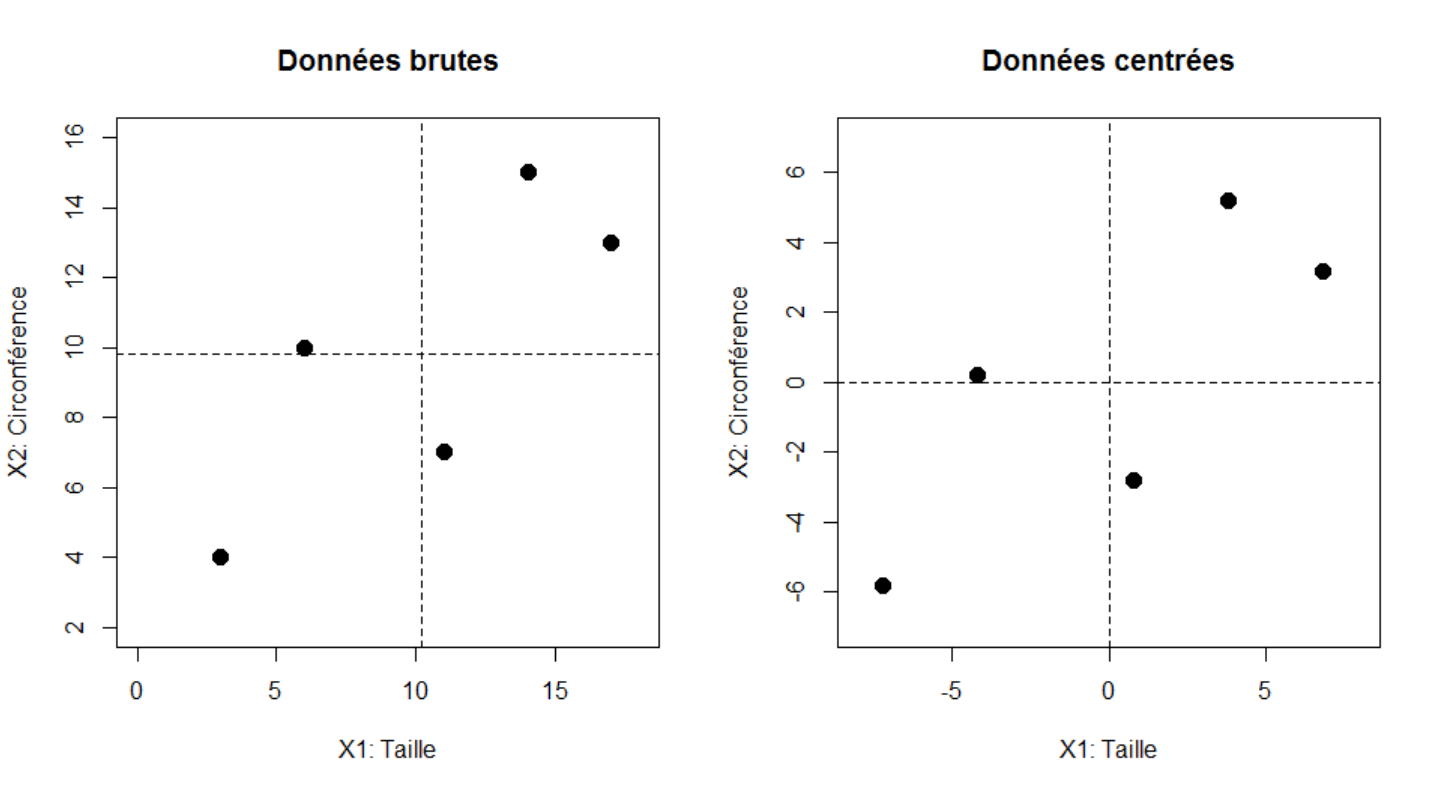
Contenu

Analyse en composantes principales classique

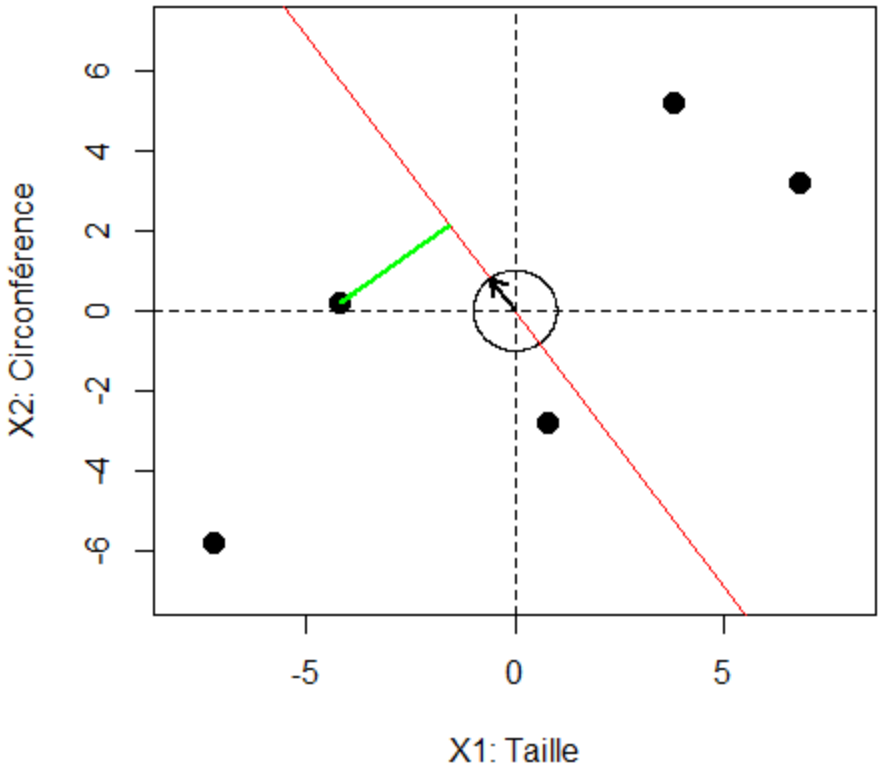
Principe

Tableau : I rangées, J colonnes

x_1	x_2
3	4
6	10
11	7
14	15
17	13



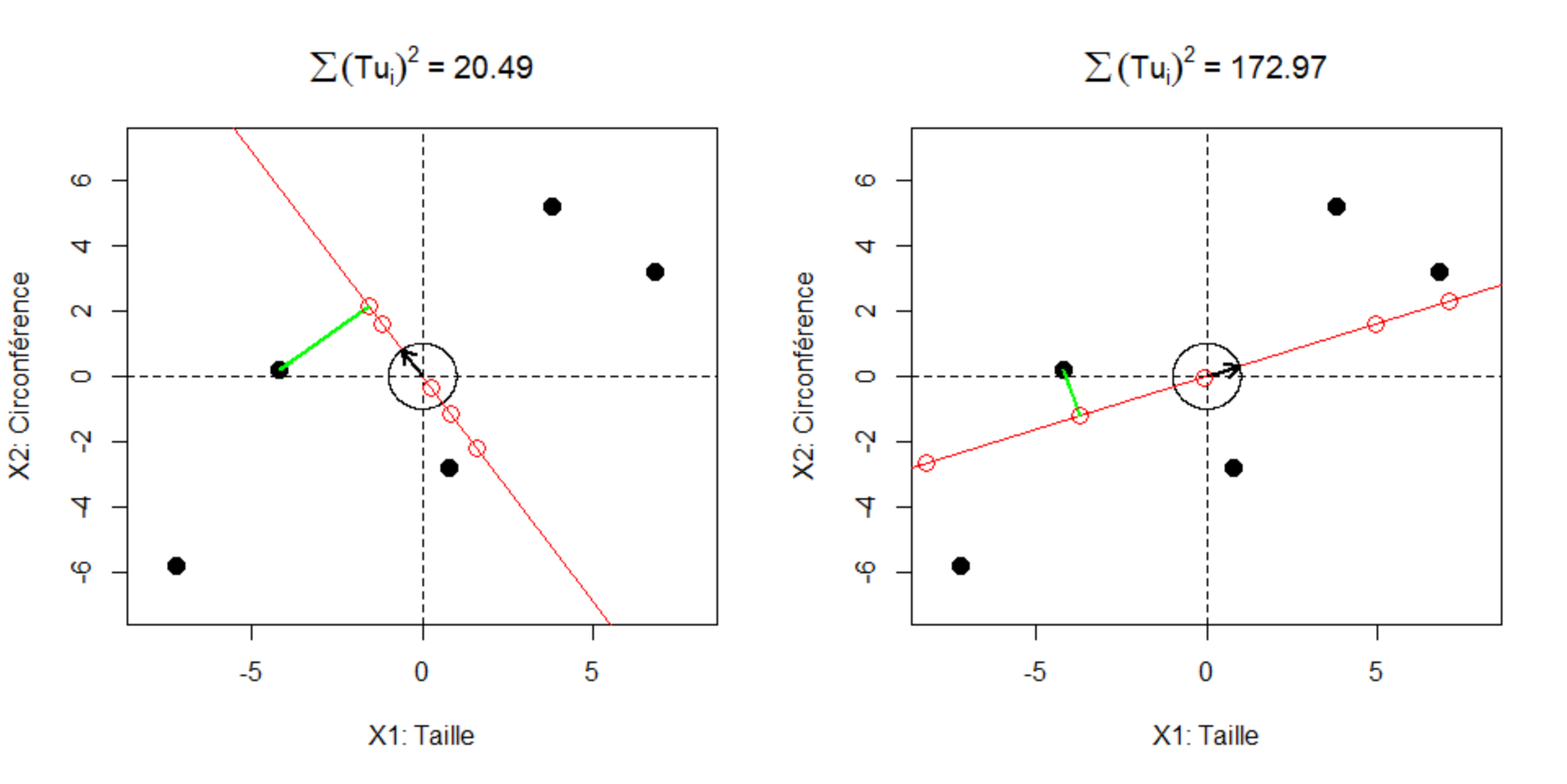
Centrer les données : soustraire la moyenne.



Considérer toutes les directions

i.e. les vecteurs u dans R^J unitaires sous la norme euclidienne « $\boldsymbol{L_2}$ »

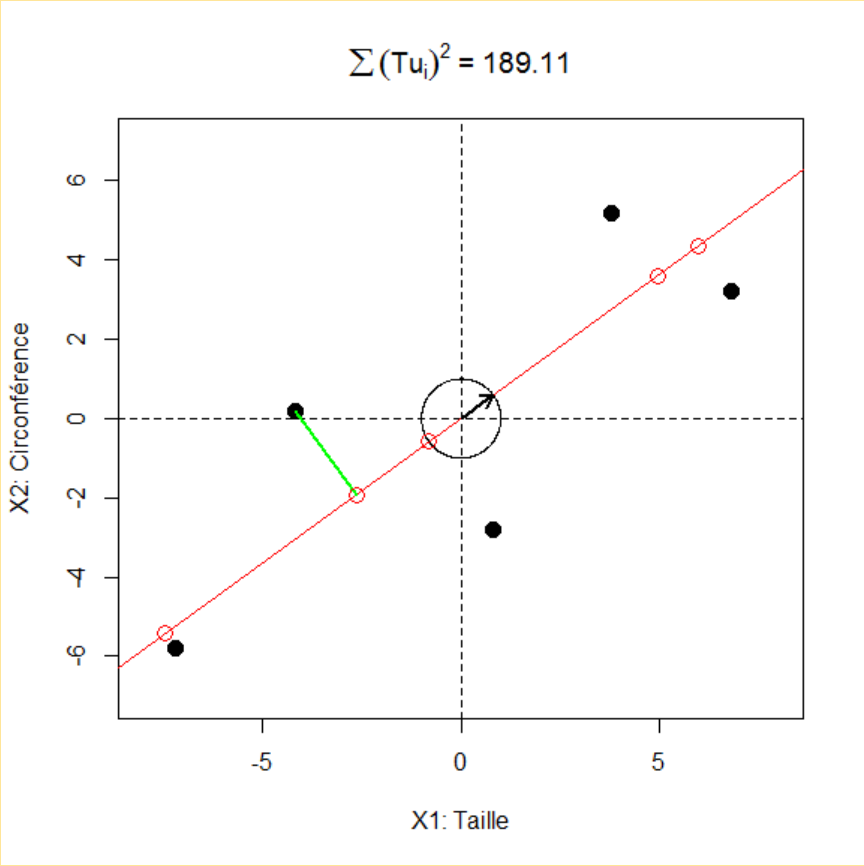
= la sphère « $\boldsymbol{L_2}$ » de rayon 1



Considérer la projection Tu des points sur chaque droite.

C'est un vecteur dans R^I.

Maximiser : $\|\boldsymbol{Tu}\|_2$ sous $\|\boldsymbol{u}\|_2 = \mathbf{1}$



Solution par algèbre linéaire : Méthode de décomposition par les valeurs singulières

Résultat : La première composante principale (définie au signe près)

Itération

Projeter les points dans l'espace perpendiculaire à la composante principale

Reprendre le calcul pour trouver la deuxième composante principale, etc…

==========

$\max(\|Tu\|_2)^2$ = la variance des projections

= importance relative des composantes principales

Elle diminuera progressivement d'une composante à l'autre

Diagramme en « éboulis » (« scree plot »)

Interprétation…

Axe 1 : « Bigness »

(Joueurs de football américain et de basket vs jockeys et gymnastes)

Axe 2 : « % de gras »

Normes

$$L_2$$

$$\|(x_1, x_2, \dots, x_I)\|_2 = \sqrt{\sum {x_i}^2}$$

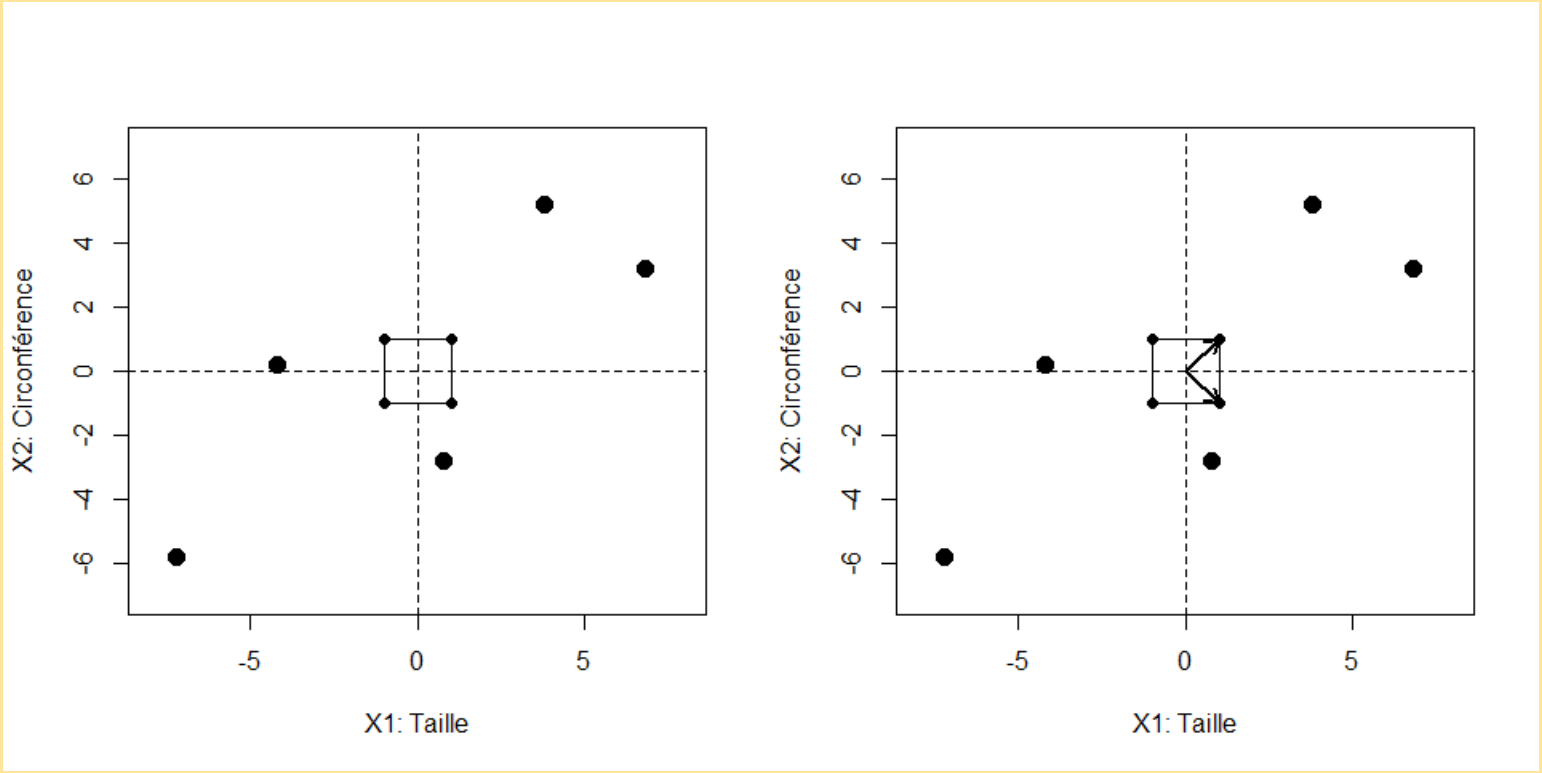
« Taxi » ou « Manhattan »

$$L_1$$

$$\|(x_1, x_2, \dots, x_I)\|_1 = \sum |x_i|$$

Emprunté de : https://afterstudying.wordpress.com/2012/08/13/Taxi-geometry/

Analyse en composantes principales Taxi

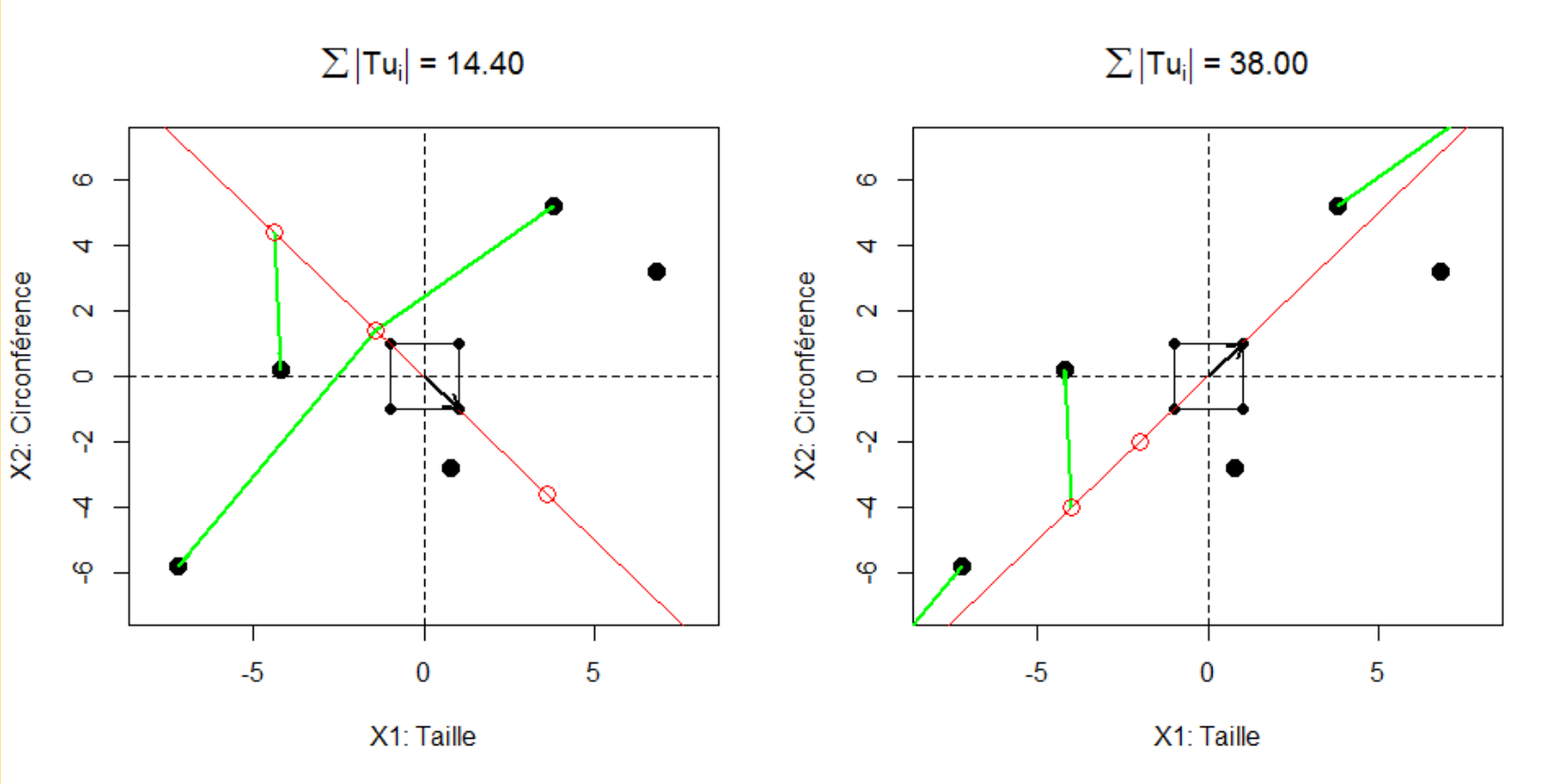


Considérer toutes les directions, i.e.

les vecteurs dans R^J unitaires sous la norme « L_∞ » : $\|(u_j)\|_\infty = \max(u_j) = 1$

= la sphère « L_∞ » de rayon 1

Il suffit de considérer les sommets de l'hypercube, i.e. les 2^J vecteurs ($\pm 1, \pm 1, \ldots, \pm 1$)



Considérer la projection Tu des points sur chaque droite, c'est un vecteur dans R^I

Maximiser : $\|\boldsymbol{Tu}\|_1$ sous $\|\boldsymbol{u}\|_\infty = \mathbf{1}$

(ACP Classique : Maximiser : $\|\boldsymbol{Tu}\|_2$ sous $\|\boldsymbol{u}\|_2 = \mathbf{1}$)

Il suffit de considérer les 2^{J-1}projections sur les vecteurs unitaires $u = (1, \pm 1, \ldots, \pm 1)$

C'est un **PROBLÈME COMBINATOIRE (FINI)** !

Itérations

Projeter les points dans l'espace perpendiculaire à la composante principale

Reprendre le calcul pour trouver la deuxième composante principale Taxi, etc…

==========

$\max(\|Tu\|_1)$ = la dispersion Taxi

= importance relative des composantes principales Taxi.

Algorithmes

Exhaustif :

Calculer Tu pour tous les $u = (1, \pm 1, \dots, \pm 1)$

Produit matriciel : $T_{I \times J} U_{J \times 2^{J-1}}$

OK pour $J \leq 22$; par bloc en 2^{J-1} si I est grand

Directif + Criss-cross

Examiner les I vecteurs unitaires $u_i = signe(T_{ij})$

Appliquer un algorithme de type « criss-cross » (2 itérations) pour chaque u_i

Génétique

Librairie GA

R CRAN: Ne pas inclure de « Depends » (« Suggests » OK!)

Performance de R en calcul numérique

https://modelingguru.nasa.gov/docs/DOC-2625

Language	Option	n=2000	Facteur
Python	intrinsic	95	3
Python + Numba (loops)		1340	37
Julia	intrinsic	73	2
R	intrinsic	2863	80
IDL	intrinsic	36	1
Matlab	intrinsic	99	3
Fortran	gfortran (loop)	8366	232
	gfortran -O3 (loop)	1213	34
	gfortran (matmul)	434	12
	gfortran -O3 (matmul)	368	10
	ifort (loop)	448	12
	ifort -O3 (loop)	124	3
	ifort (matmul)	448	12
	ifort -O3 (matmul)	125	3
	ifort (DGEMM)	33	1
C	gcc (loop)	3177	88
	gcc -Ofast (loop)	430	12
	icc (loop)	399	11
	icc -Ofast (loop)	262	7
Scala	Simple loop	3230	90
	with la4j	959	27
	with JAMA	938	26

Table 2.1: Elapsed times (in seconds) obtained by multiplying two randomly generated matrices.

Performance des algorithmes

Exemple: $I = 26, J = 15$

```
Exhaustif : 1.3 s
Directif : 0.14 s
Génétique : 17.5 s
```

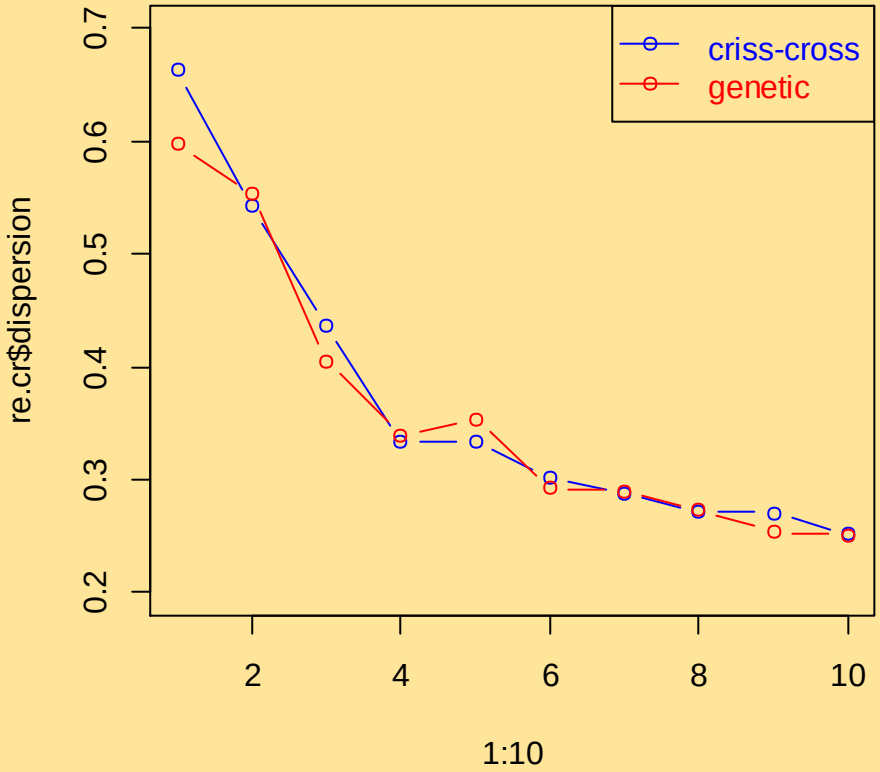
Exemple

$I = 126, J = 129$

```
Directif :  2.1 s
Génétique :  46.9 s
```

Utiliser l'algorithme génétique en particulier si $I \approx J$

(À venir : Directif + Génétique)



2^{J-1}= « Malédiction de la dimensionnalité »

Étude des tableaux de contingence

STAT1001 : Petit tableau – « Tourisme » Données fictives

		Mode de transport					
		bicyclette	**tr.commun**	**voiture**	**bateau**	**bus.groupe**	**Total**
Âge	**18-24**	20	25	2	1	0	**48**
	25-39	10	10	8	3	0	**31**
	40-59	6	3	18	5	6	**38**
	60-74	4	19	35	32	22	**112**
	75-99	0	5	2	20	15	**42**
	Total	**40**	**62**	**65**	**61**	**43**	**271**

$$\chi^2 = \sum \frac{(O - A)^2}{A}$$

	bicyclette	tr.commun	voiture	bateau	bus.groupe	Total
18-24	12.9	14	-9.5	-9.8	-7.6	0
25-39	5.4	2.9	0.6	-4	-4.9	0
40-59	0.4	-5.7	8.9	-3.6	0	0
60-74	-12.5	-6.6	8.1	6.8	4.2	0
75-99	-6.2	-4.6	-8.1	10.5	8.3	0
Total	0	0	0	0	0	0

Tableau… éparse (sous une définition quelconque !!!)

	M1	M2	M3	M4	M5	M6	M8	M9	M10	M16	M18	M20	M4A	M11	M12	M13	M14	M15	M19	Total
29cookT	2	5	3		1		2	4	2	1	1	2	1	3			1		1	**29**
5cookV								1		1		3								**5**
2cookS		1					1													**2**
1cookT									1											**1**
10cookH			2	1				1		1	1		3		1					**10**
21dinS		1	4	2	2	1	1			3	2	1		2				1	1	**21**
29dinV		1	2	1	1		1	1	4	3	2	2		6		3	1	1		**29**
5dinV									1	1	1			2						**5**
33dinB	1	3	1	1	1	2	2	5	4	2	1	3		2		2		2	1	**33**
10dinB								2	1		1	2	1				2		1	**10**
9dinV	1	1	2		1			1	2			1								**9**
47dinJ	1	3	4	3	2	2	2	4	4	2	5	5	1	7				1	1	**47**
8dinB	1			2	1			2					1					1		**8**
19dinB	1	3	2	1			2	1	2	1		1	1	1	1	1			1	**19**
5dinB	1	3										1								**5**
3pouJ							1		1			1								**3**
29proMP		2	2	2			1	**12**	3	1		2	1	2				1		**29**
16proMH		4	1				1	3	1			6								**16**
10spinW								2	5							2			1	**10**
21storV		2	2	1		1	2	2	2	4		1	1	1				1	1	**21**
8storV		1	1					1	1	1	1			1					1	**8**
27storV	2	2	1	2	1	1	3	4	2	1	2	3		1		1			1	**27**
3storV	1								2											**3**
11storV	2	2	1	1		1		1	1			1							1	**11**
20storV	2	3	1	1	1		2	2	1	1	1		1		1	1		1	1	**20**
13workST		2	1		1		2	3	4											**13**
53workL		4	7		1	6	1	7	8	5	5	6				3				**53**
1covL			1																	**1**
5workL			1	2	1			1												**5**
18workV	1			3		2			1	1		3		4		1	1		1	**18**
18workP						**12**	1							5						**18**
Total	**16**	**43**	**39**	**23**	**14**	**28**	**25**	**60**	**53**	**29**	**23**	**44**	**11**	**37**	**3**	**14**	**5**	**9**	**13**	**489**

Historique

Analyse des correspondances (Benzécri ~ 1965)

Concept : « ACP sur les rangées du tableau $O - A$ avec pondération par la marginale »

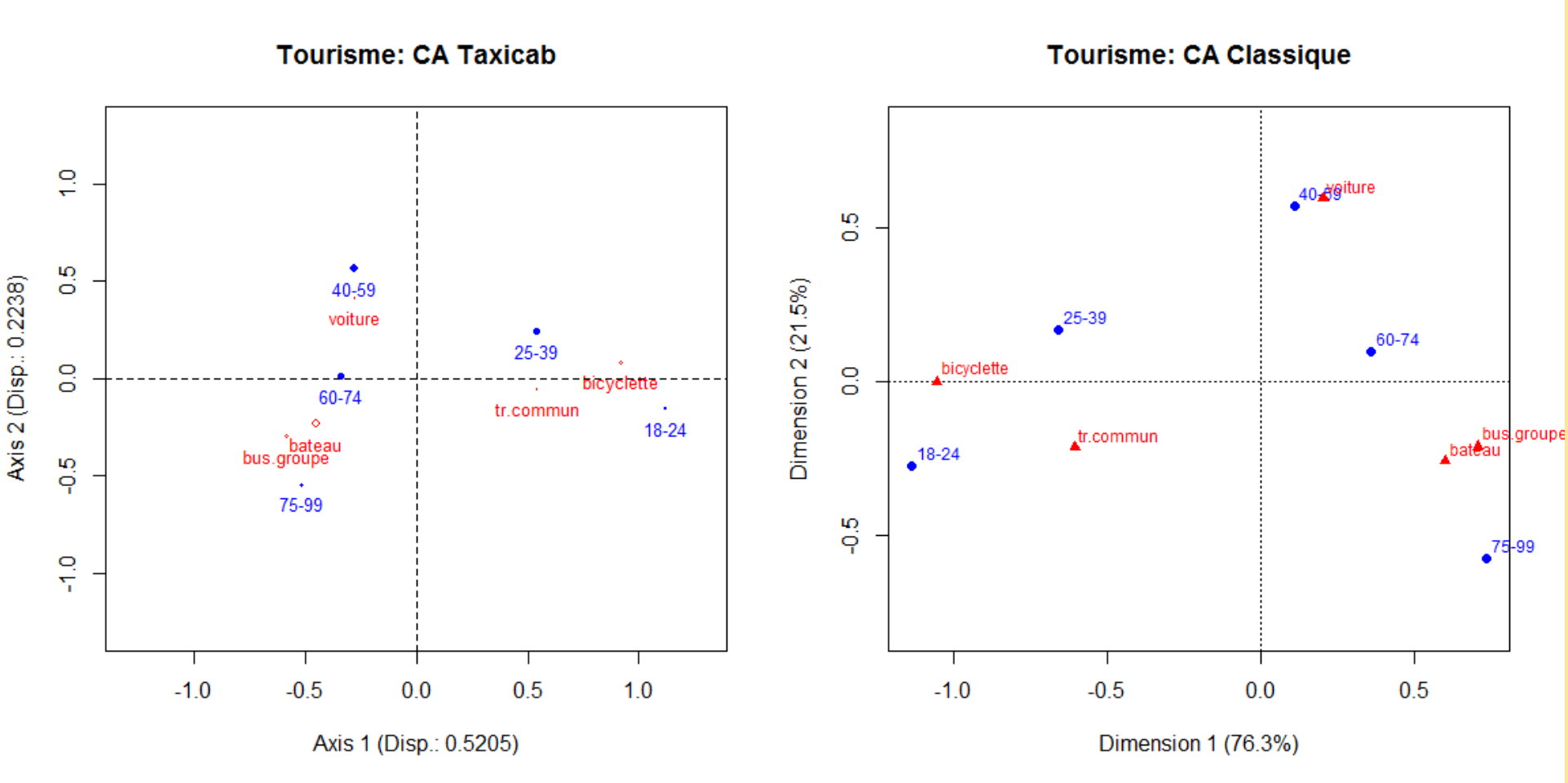
Métriques L_2 et L_2

Analyse des correspondances Taxi (Choulakian ~ 2006)

Métriques L_1 et L_{Inf}

Note : Dualité rangées colonnes

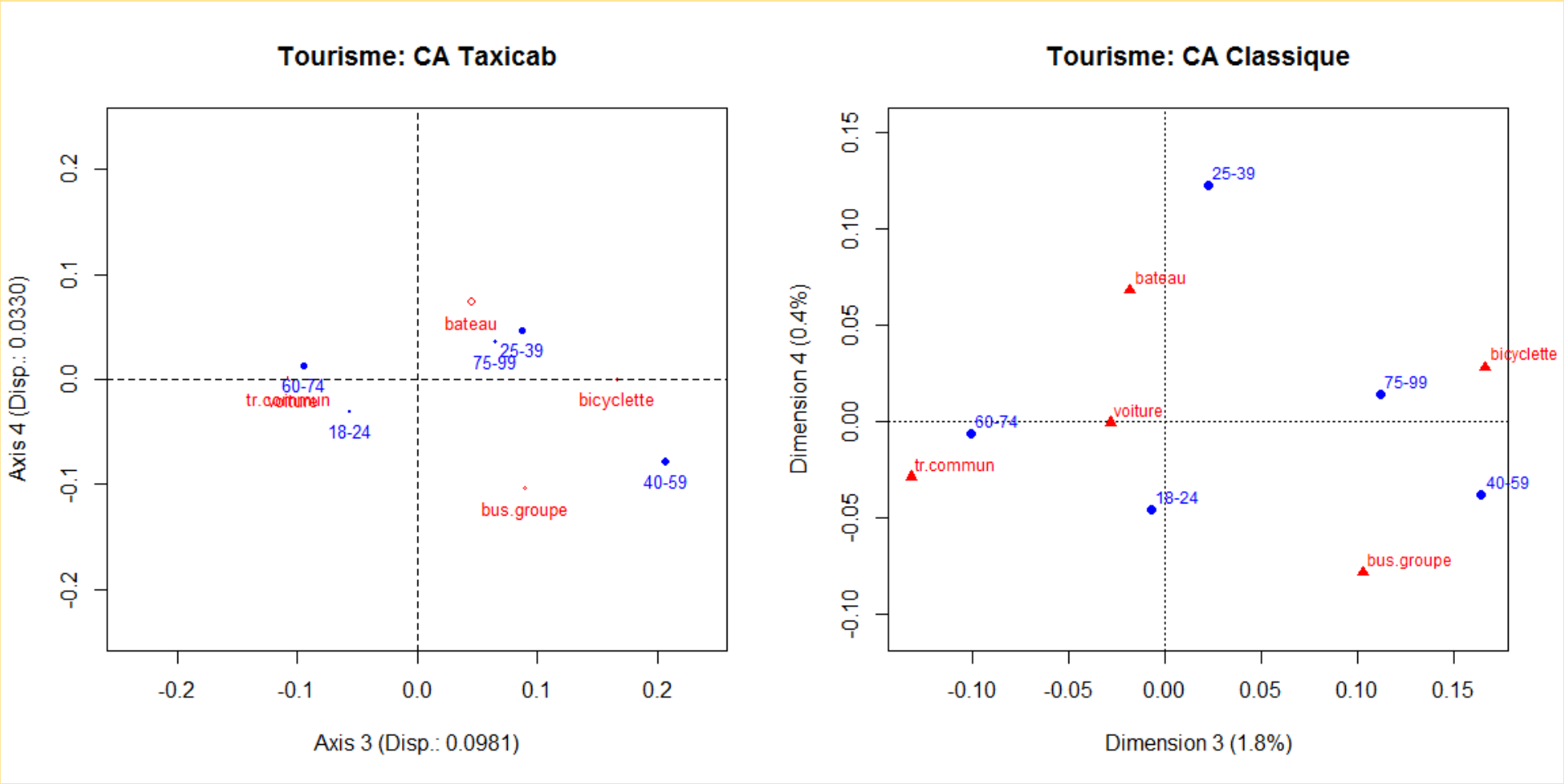
Analyse de « Tourisme » (5x5)



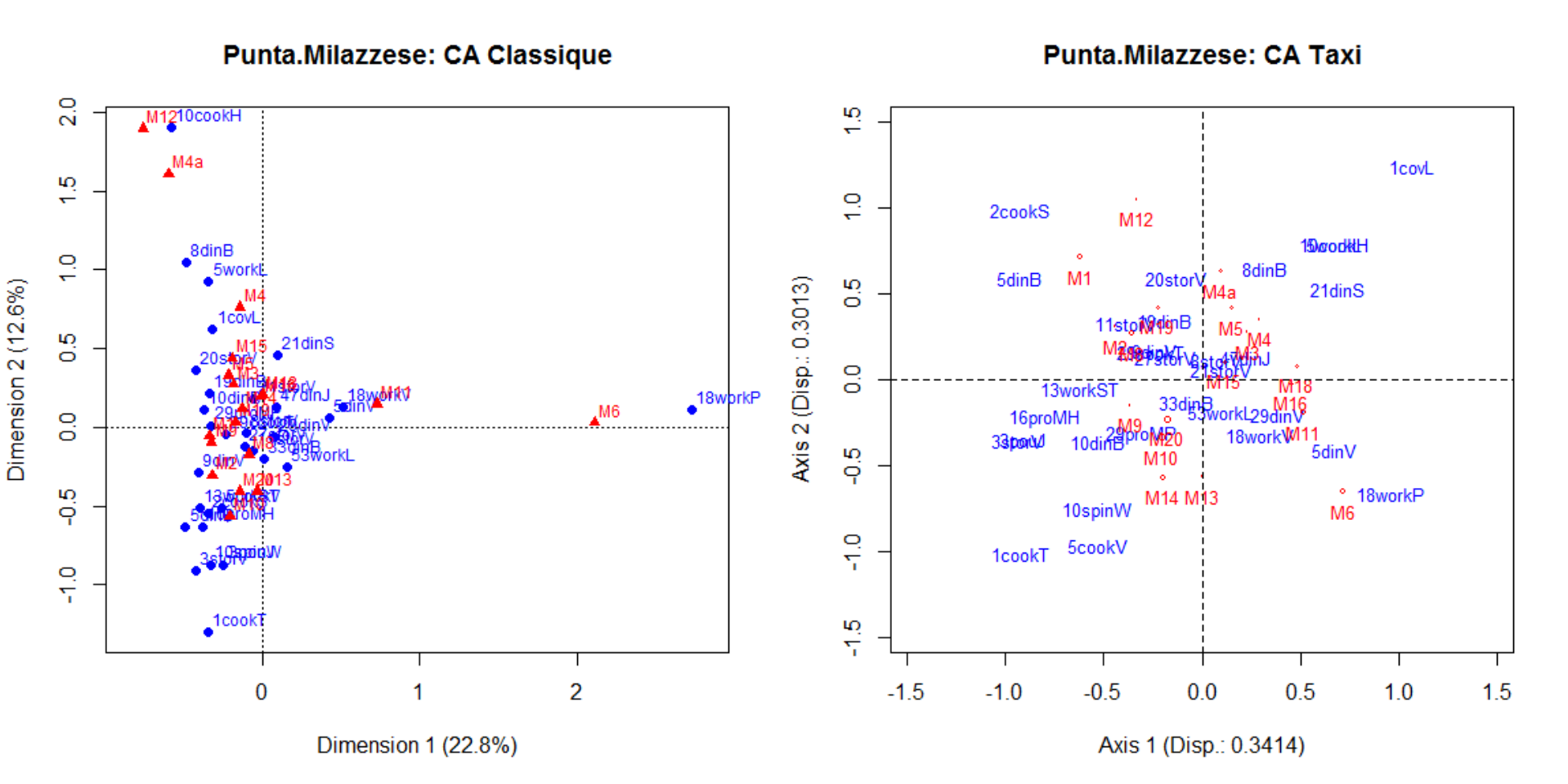
Axe 1 : Jeune vs Âgé

Axe 2 : Autonome vs Organisé

CA ET TCA DONNENT DES RÉSULTATS SIMILAIRES

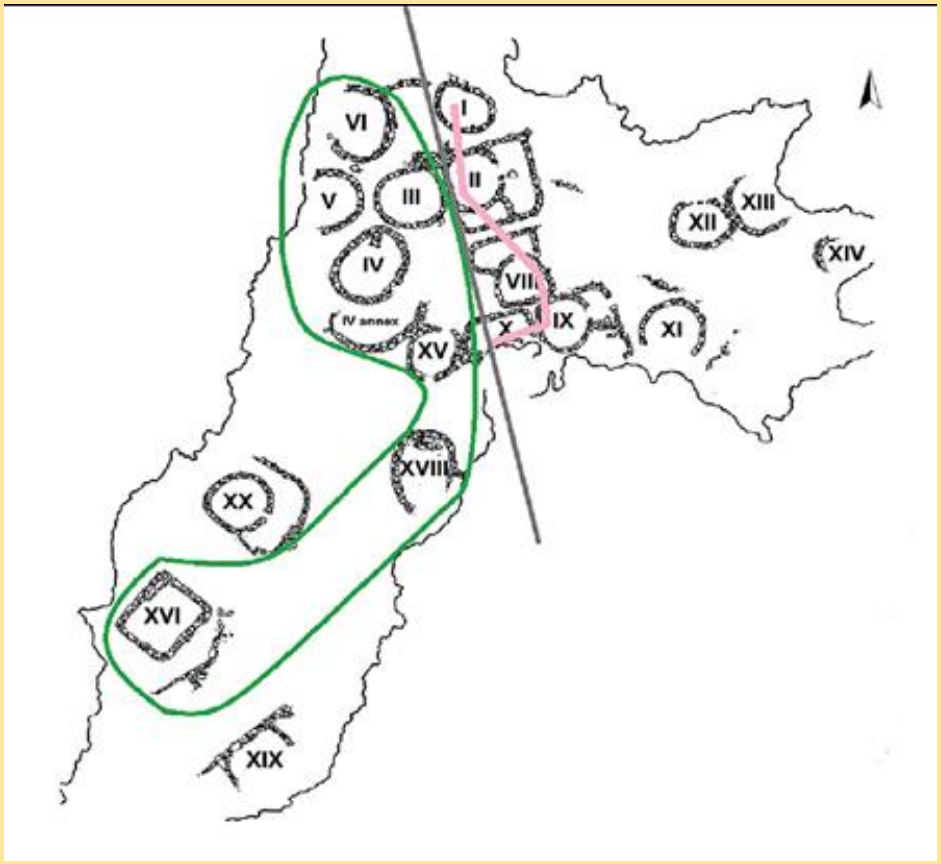

Tourisme: CA Taxicab
Axis 4 (Disp.: 0.0330)
Axis 3 (Disp.: 0.0981)
bateau
25-39
75-99
60-74
tr.commun
voiture
18-24
bicyclette
40-59
bus.groupe
Tourisme: CA Classique
Dimension 4 (0.4%)
Dimension 3 (1.8%)
25-39
bateau
bicyclette
75-99
60-74
voiture
tr.commun
18-24
40-59
bus.groupe

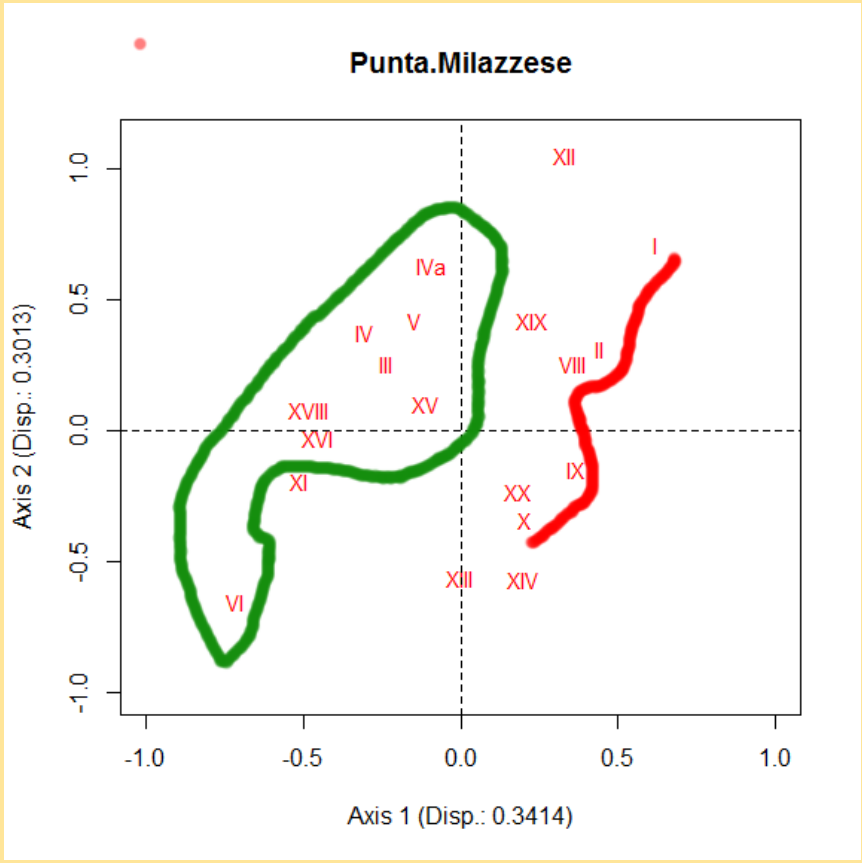
Analyse des données archéologiques (31x19)



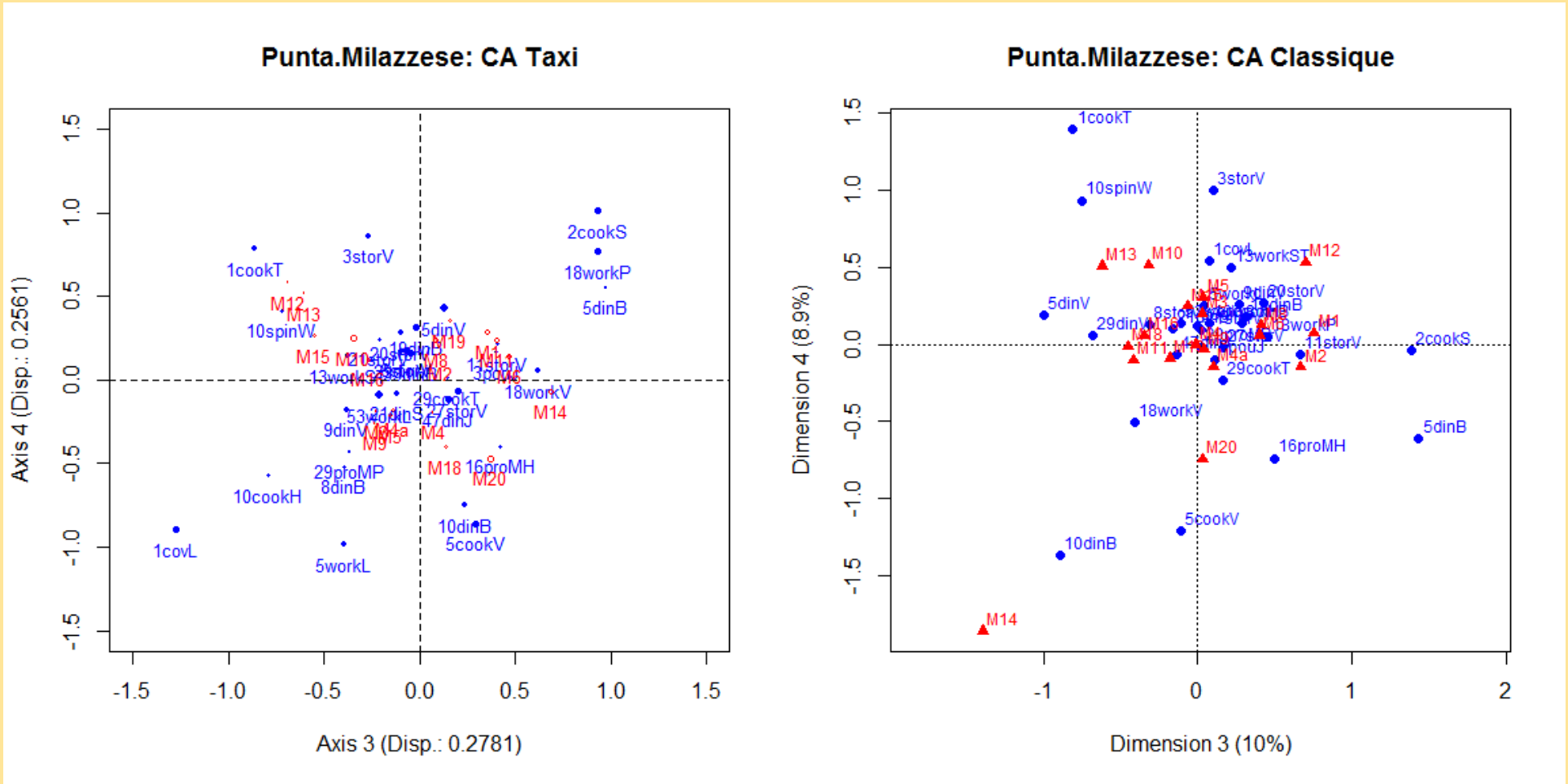
CA ET TCA NE DONNENT PAS DES RÉSULTATS SIMILAIRES

Coïncidence de la topologie du site d'excavation et de la projection CA Taxi…





INTERPRÉTATION ?


Punta.Milazzese: CA Taxi
Axis 4 (Disp.: 0.2561)
Axis 3 (Disp.: 0.2781)
1cookT
3storV
2cookS
18workP
5dinB
M12
M13
10spinW
5dinV
M15
M19
M1
M8
13storV
18workV
M14
29cookT
27storV
9dinV
M4
M18
16proMH
M20
29proMP
8dinB
10cookH
10dinB
5cookV
1covL
5workL
Punta.Milazzese: CA Classique
Dimension 4 (8.9%)
Dimension 3 (10%)
1cookT
3storV
10spinW
M13
M10
1covL
13workST
M12
5dinV
20storV
29dinV
18workP
11storV
2cookS
M2
29cookT
18workV
5dinB
M20
16proMH
5cookV
10dinB
M14

Conclusion statistique

« Le concept développé dans “Du Cubisme” d'observer un sujet à partir de différents points dans l'espace en même temps… »

Source:
https://fr.wikipedia.org/wiki/Cubisme

“In Les Demoiselles d'Avignon, he depicts one of the demoiselles simultaneously full face and in profile, two perspectives at once, a projection from the fourth dimension. He had gone beyond Poincaré.”

Source :
Arthur I Miller, “Henri Poincaré: the unlikely link between Einstein and Picasso”, www.theguardian.com

print.tca

Présentation traditionnelle des contributions des rangées et des colonnes

```
Data name: Tourisme: CA Taxi

Algorithm used: Exhaustive

Dispersion

               Axis1    Axis2    Axis3    Axis4
Dispersion 0.520527 0.223838 0.098125 0.032997


Column contributions x 1000

      bicyclette tr.commun voiture bateau bus.groupe
Axis1        917       546    -275   -452       -583
Axis2         83       -54     415   -231       -300
Axis3        166      -107    -102     46         90
Axis4          0         0       0     73       -104
MASS         148       229     240    225        159


Row contributions (x 1000) and Row masses (x 10^3)

      Axis1 Axis2 Axis3 Axis4 MASS
18-24  1122  -157   -56   -31  177
25-39   538   243    87    48  114
40-59  -279   563   208   -78  140
60-74  -342    13   -95    13  413
75-99  -515  -543    64    35  155
```

plot.tca

```
## S3 method for class 'tca'
plot(
  tcaObject,
  axes = c(1, 2),
  labels.rc = c(0, 1),
  col.rc = c("blue", "red"),
  pch.rc = c(16, 21),
  mass.rc = c(T,T),
  cex.rc = c(NA, NA),
  jitter = c(T, F),
  saveToFile = F,
  path = ".",
  folder = NULL,
  type = "pdf"
)
```

summary.tca

```
Data name: Tourisme: CA Taxi

Algorithm used: Exhaustive

Dispersion

               Axis1    Axis2    Axis3    Axis4
Dispersion 0.520527 0.223838 0.098125 0.032997
```

saveTCA

```
saveTCA(
  tcaObject,
  path = ".",
  folder = NULL,
  what = c("report", "csv",
           "plot", "dataMatrix", "tcaObject"),
  plotAxes = matrix((1:2),
                    nr = 1,
                    nc = 2,
                    byrow = T
  ),
  type = "pdf",
  csvFormat = c("csv", "csv2")
)
```

Dossier « Tourisme »

```
colContribs.csv
colMass.csv
dataMatrix.csv
dataMatrixTotal.csv
dispersion.csv
Report.txt
rowContribs.csv
rowMass.csv
TCA Plot 1x2.pdf
TCA Plot 3x4.pdf
```

Report.txt

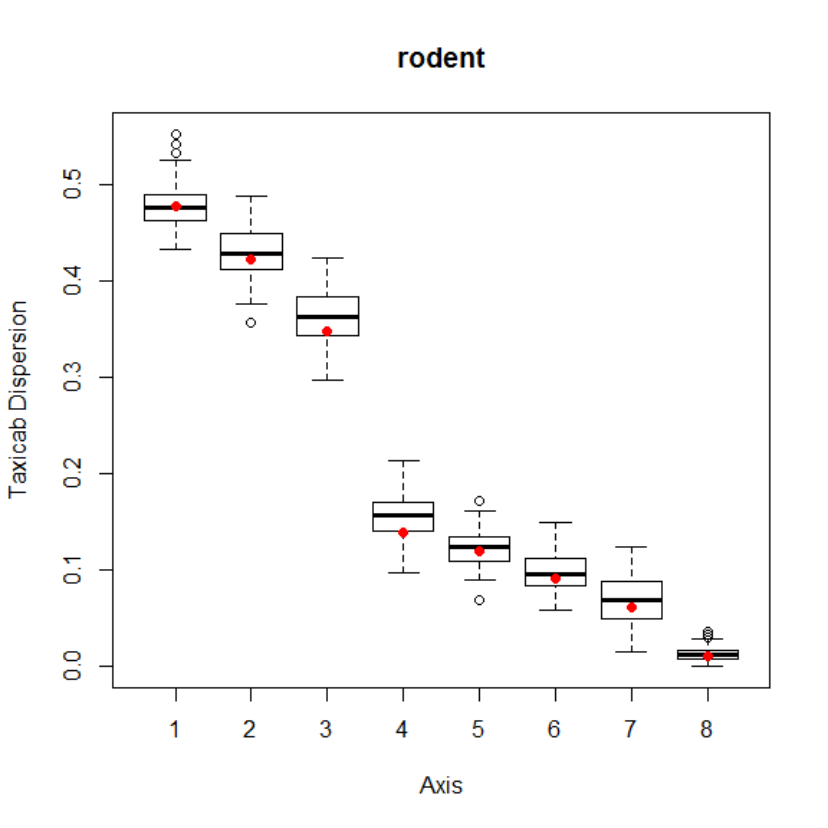
```
Data name: Tourisme

Algorithm used: Exhaustive

Dispersion

              Axis1    Axis2    Axis3    Axis4
Dispersion 0.520527 0.223838 0.098125 0.032997


Column contributions x 1000

      bicyclette tr.commun voiture bateau bus.groupe
Axis1        917       546    -275   -452       -583
Axis2         83       -54     415   -231       -300
Axis3        166      -107    -102     46         90
Axis4          0         0       0     73       -104
MASS         148       229     240    225        159


Row contributions (x 1000) and Row masses (x 10^3)

      Axis1 Axis2 Axis3 Axis4 MASS
18-24  1122  -157   -56   -31  177
25-39   538   243    87    48  114
40-59  -279   563   208   -78  140
60-74  -342    13   -95    13  413
75-99  -515  -543    64    35  155
```

À venir

Amélioration de la librairie

- Option pour un algorithme d'optimisation défini par l'usager

Bootstrap



Sélection du nombre de composantes par validation croisée

« Composantes à retenir = Composantes *interprétables* ! »

vs

« Composantes à retenir = Composantes *stable sous la validation croisée* ! »

Art B. Owen, Jingshu Wang (2016): Bi-cross-validation for factor analysis

Questions ?