Analyses factorielles avec R

Principes généraux

- Méthodes multivariées :
  - permettent d'analyser les relations entre un grand nombre de variables (par opposition aux statistiques univariées et bivariées)
- Résumer un ensemble de variables par des variables synthétiques
- Représentations géométriques qui transforment en distance euclidienne des ressemblances statistiques entre profils

Source :Escofier, Pagès, 1998

Trois techniques classiques

- ACP : tableau croisant des individus et des variables numériques
- ACF : tableaux de fréquence
- ACM : tableaux croisant des individus et des variables qualitatives
- Un même principe : on construit 2 nuages de point, l'un représentant les lignes (les individus), l'autre représentant les colonnes (les variables)
  - Il ya bien sur une association très forte entre ces deux nuages

Source :Escofier, Pagès, 1998

Analyses en Composantes Principales (ACP)

- S'applique à des tableaux à 2 dimensions croisant individus et variables
  - Individus en ligne, variable en colonnes
- A propos de 2 individus, on essaie d'évaluer leur ressemblance : deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables
- A propos de 2 variables, on évalue leur liaison à partir du coefficient de corrélation linéaire

Source :Escofier, Pagès, 1998

Objectifs

- Bilan des relations entre individus
  - Quels sont les individus qui se ressemblent ?
  - Peut-on mettre en évidence une typologie des individus ?
- Bilan des relations entre variables
  - Quelles variables sont corrélées entre elles ? Peut on mettre en évidence une typologie des variables ?
- Résumer l'ensemble des variables par des variables synthétiques appelées composantes principales

Source :Escofier, Pagès, 1998

Nuage des individus (1)

- Impossible de représenter le nuage (nb de dimensions bien supérieure à 3)
- On cherche à fournir des images planes
  - On cherche des axes factoriels qui pris 2 à 2 vont former des axes factoriels
  - Chaque direction est orthogonale aux axes précédents
  - On parle aussi des principaux facteurs de variabilité, dans la mesure où ils rendent compte le plus possible de la diversité des individus

Source :Escofier, Pagès, 1998

Nuage des individus (2)

- Les axes rendent minimum l'écart entre le nuage des individus et sa projection
- La projection ne pouvant que réduire la distance entre points, les axes factoriels apparaissent comme les directions telles que les distances entre les points projetés ressemblent le plus possibles aux distances entre les points homologues de N_I

Source :Escofier, Pagès, 1998

Nuage des variables (1)

- Ce sont les angles entre les vecteurs représentant les variables qui sont peu déformés par les projections et non pas les distances entre les points

- On met en évidence une suite de variables synthétiques, les composantes principales, non corrélées entre elles, qui résument au mieux l'ensemble de variables initiales

Source :Escofier, Pagès, 1998

Nuage des variables (2)

- Deux individus situés à une même extrémité d'un axe sont proches car ils ont tous deux généralement de fortes valeurs pour les variables situées du même côté qu'eux et de faibles variables situées à l'opposé

Source :Escofier, Pagès, 1998

Applications sur R

- Utilisation du plugin FactomineR de Rcmdr
  - Développé par le département de mathématiques de l'agrocampus de Rennes

- Installation

  source("http://factominer.free.fr/install-facto-fr.r")

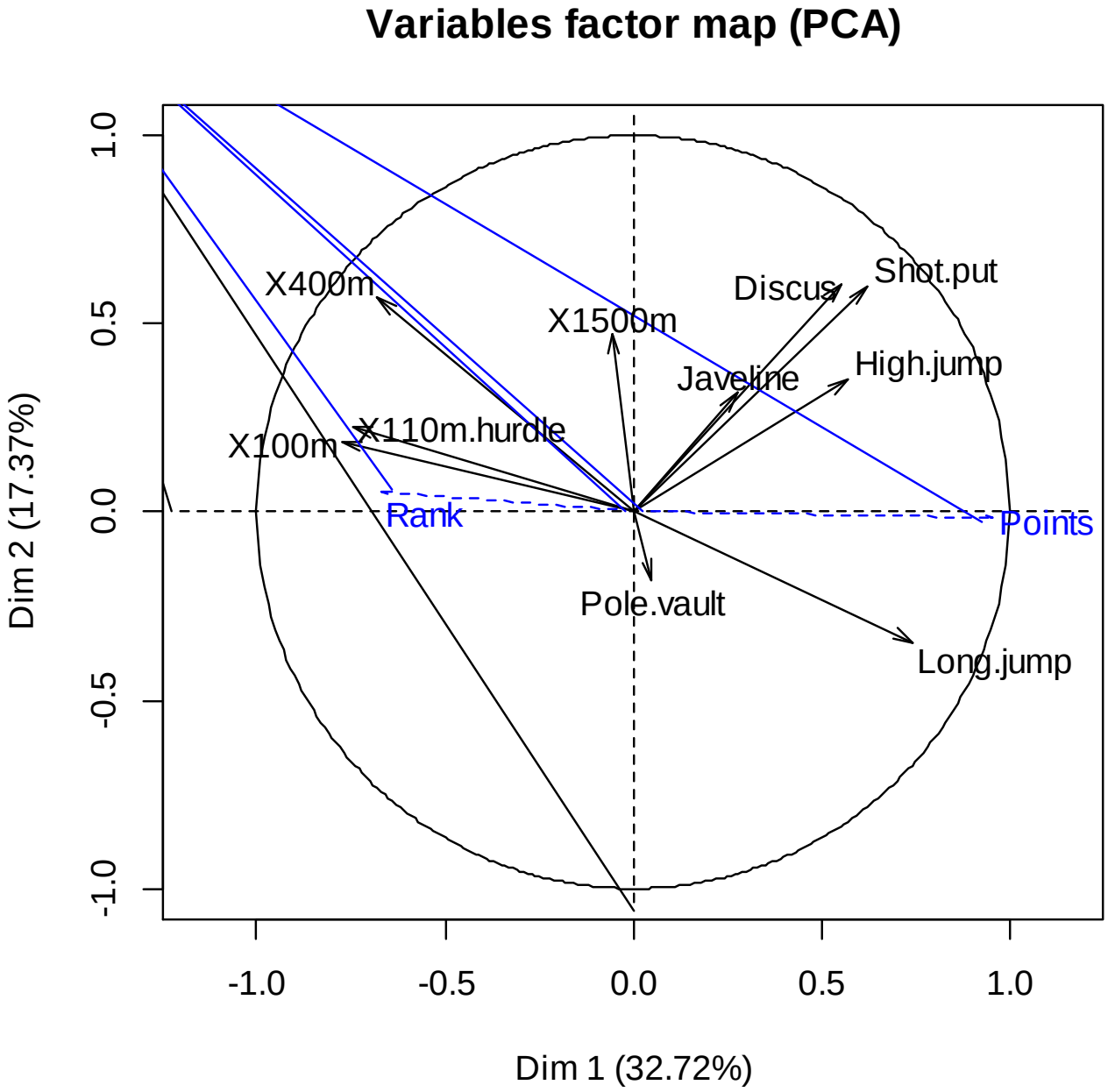
- Lancer FactoMineR puis Rcommander

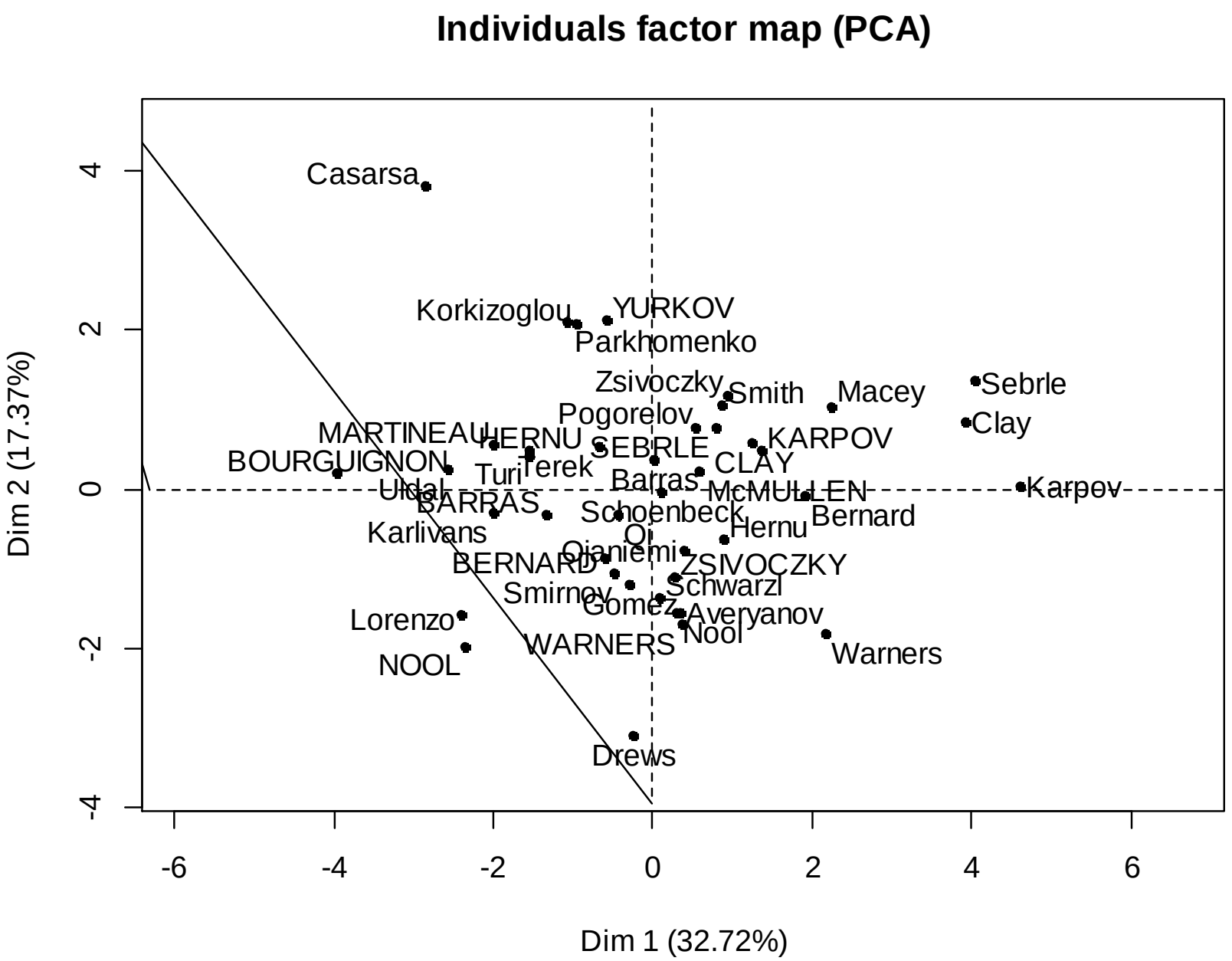
Tutoriel

- http://factominer.free.fr/classical-methods/analyse-en-composantes-principales.html

- Tutoriel:
  - http://factominer.free.fr/classical-methods/analyse-des-correspondances-multiples.html

- data(decathlon)
- data(tea)


Variables factor map (PCA)
Dim 2 (17.37%)
Dim 1 (32.72%)
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
X400m
Discus
Shot.put
X1500m
Javeline
High.jump
X100m
X110m.hurdle
Rank
Points
Pole.vault
Long.jump


Individuals factor map (PCA)
Dim 2 (17.37%)
Dim 1 (32.72%)
Casarsa
Korkizoglou
YURKOV
Parkhomenko
Zsivoczky
Smith
Macey
Sebrle
Pogorelov
Clay
MARTINEAU
HERNU
SEBRLE
KARPOV
BOURGUIGNON
Terek
CLAY
Barras
Karpov
McMULLEN
BARRAS
Schoenbeck
Bernard
Karlivans
Hernu
Qi
Qianiemi
BERNARD
ZSIVOCZKY
Schwarzl
Smirnov
Gomez
Averyanov
Lorenzo
Nool
WARNERS
Warners
NOOL
Drews

- Shot put : lancer de poids
- Discuss throw : lancer de disque
- Pole vault : saut à la perche

- Les deux premières dimensions contiennent 50% de la variance.
- La variable *"X100m"* est négativement corrélée à la variable *"long.jump"*. Quand un ahtlète réalise un temps faible au 100m, il peut sauter loin. Il faut faire attention ici qu'une petite valeur pour les variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* et*"X1500m"* correspond à un score élevé : plus un athlète court rapidement, plus il gagne de points.
- Le premier axe oppose les athlètes qui sont "bons partout" comme Karpov pendant les Jeux Olympiques à ceux qui sont "mauvais partout" comme Bourguignon pendant le Décastar.
- Le deuxième axe oppose les athlètes qui sont forts (variables *"Discus"* et *"Shot.put"*) à ceu qui ne le sont pas. Les variables *"Discus"*, *"Shot.put"* et *"High.jump"* ne sont pas très corrélées aux variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* et *"Long.jump"*. Cela signifie que force et vitesse ne sont pas très corrélées.
- A l'issue de cette première approche, on peut diviser le premier plan factoriel en quatre parties : les athlètes rapides et puissants (comme Sebrle), les athlètes lents (comme Casarsa), les athlètes rapides mais faibles (comme Warners) et les ahtlètes ni forts ni rapides, relativement parlant (comme Lorenzo).
- Source: http://factominer.free.fr/classical-methods/analyse-en-composantes-principales.html

Analyse des Correspondances Multiples

- La technique favorite de Pierre Bourdieu…
- S'applique sur des variables qualitatives
  - Elle est donc beaucoup plus fréquente en sociologie et dans la plupart des sciences sociales

Etude des variables

- L'étude de la liaison entre des variables qualitative implique de se situer au niveau des modalités plus que des variables
  - On analysera ainsi le nuage des modalités plutôt que des variables
- On résume l'ensemble de variables qualitatives par un petit nombre de variables numériques

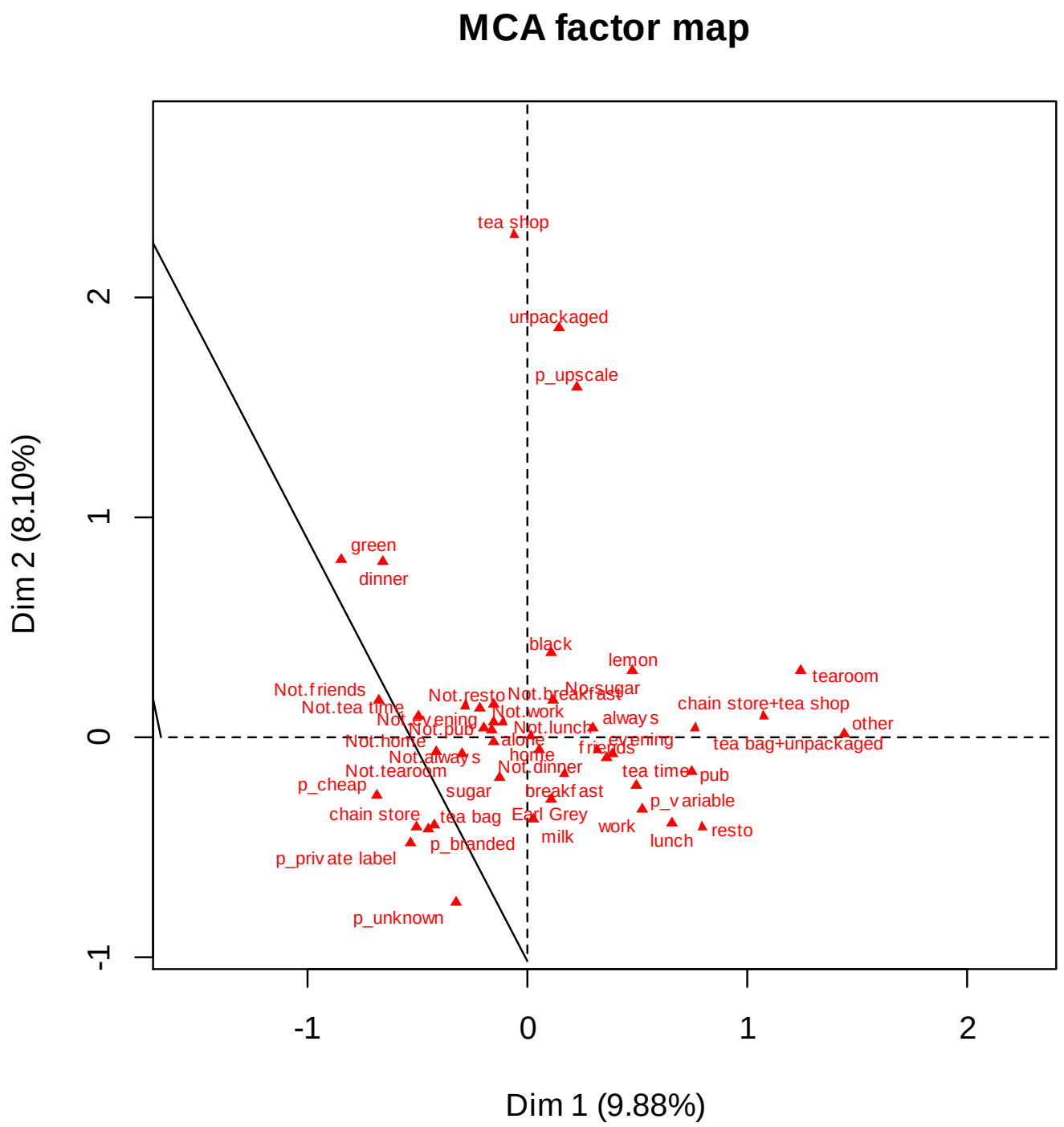
Etude des individus

- L'ACM doit permettre de réaliser une typologie des individus.
- Les individus sont d'autant plus proche qu'ils possèdent un grand nombre de modalités en commun
- Des classes d'individus se ressemblent d'autant plus que leurs profils de répartition sur l'ensemble des modalités sont proches

Source :Escofier, Pagès, 1998

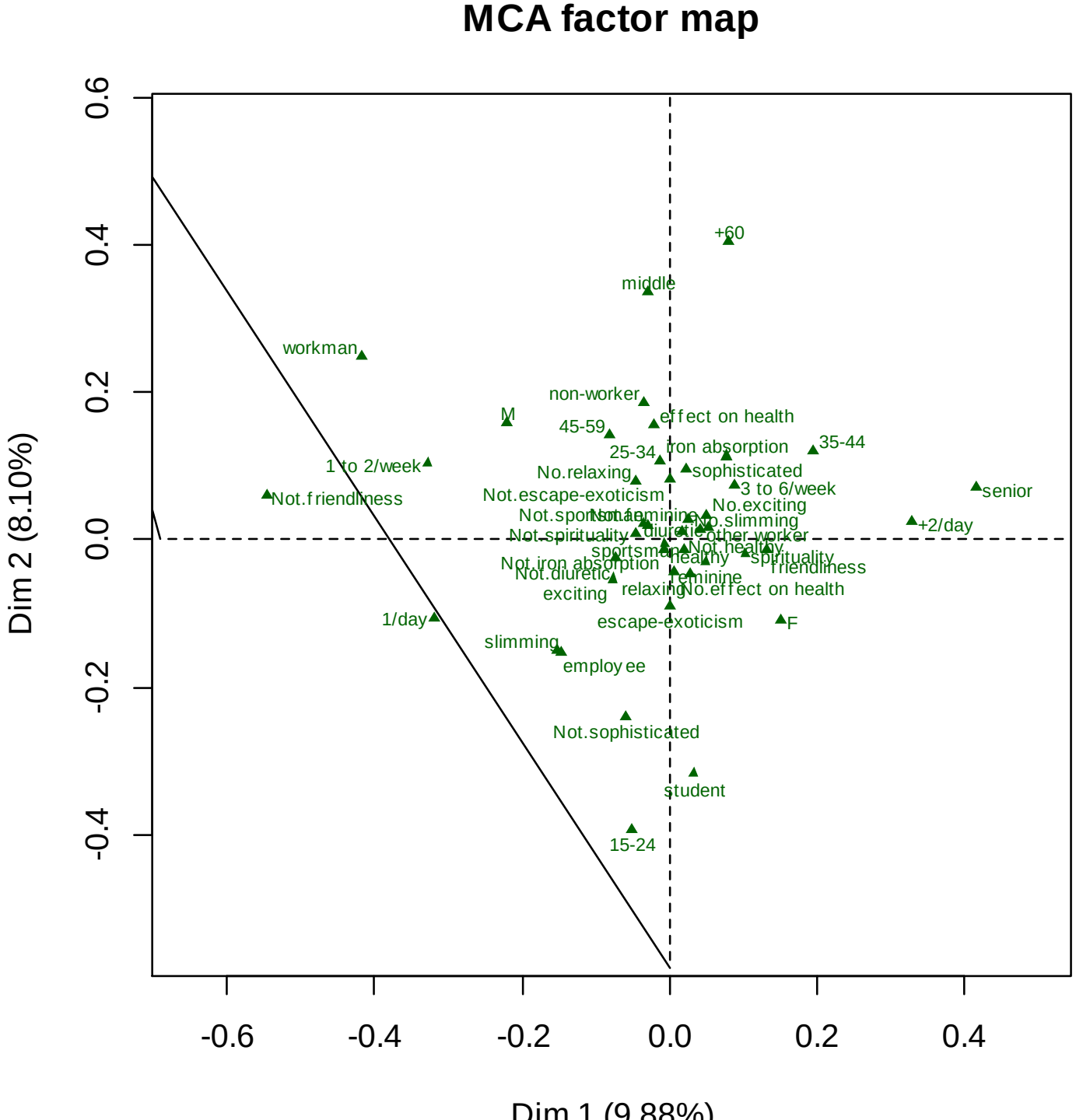
Application sur FactominerR

- res.mca = MCA(tea, quanti.sup=19, quali.sup=c(20:36))
- plot.MCA(res.mca, cex=0.5)
- plot.MCA(res.mca, invisible=c("var","quali.sup"), cex=0.5)
- plot.MCA(res.mca, invisible=c("ind","quali.sup"), cex=0.6)
- plot.MCA(res.mca, invisible=c("ind", "var"),cex=0.6)


MCA factor map
tea shop
unpackaged
p_upscale
green
dinner
black
lemon
tearoom
Not.friends
Not.resto
Not.breakfast
No sugar
Not.tea time
Not.work
chain store+tea shop
Not.evening
Not.pub
Not.lunch
always
other
Not.home
alone
evening
tea bag+unpackaged
Not.always
home
friends
Not.tearoom
Not.dinner
tea time
pub
p_cheap
sugar
breakfast
p_variable
chain store
tea bag
Earl Grey
work
resto
milk
lunch
p_branded
p_private label
p_unknown
Dim 2 (8.10%)
Dim 1 (9.88%)
-1
0
1
2

- La première dimension oppose *"tea room", "chain store+tea shop", "tea bag+unpackaged", "pub", "resto", "work"* à *"not friends", "not resto", "not work", "not home"*. Elle oppose les buveurs de thé réguliers aux buveurs occasionnels.

- La deuxième dimension oppose *« tea shop », "unpackaged"* et *"upscale price"* aux autres modalités.


MCA factor map
Dim 2 (8.10%)
Dim 1 (9.88%)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.4
-0.2
0.0
0.2
0.4
+60
middle
workman
non-worker
M
effect on health
45-59
25-34
iron absorption
35-44
1 to 2/week
No.relaxing
sophisticated
Not.friendliness
Not.escape-exoticism
3 to 6/week
senior
No.exciting
Not.sportsman
Not.feminine
No.slimming
+2/day
Not.spirituality
diuretic
other worker
Not.healthy
sportsman
healthy
spirituality
Not.iron absorption
Not.diuretic
feminine
friendliness
exciting
relaxing
No.effect on health
1/day
escape-exoticism
F
slimming
employee
Not.sophisticated
student
15-24

Obtenir les contributions et les coordonnées sur les axes

- dimdesc(res.mca)
- res$eig
- res$var
- res$ind

Bibliographie

- Escofier Brigitte, Pagès Jérôme, *Analyses factorielles simples et multiples, objectifs, méthodes et interprétation*, Dunod, 1998.