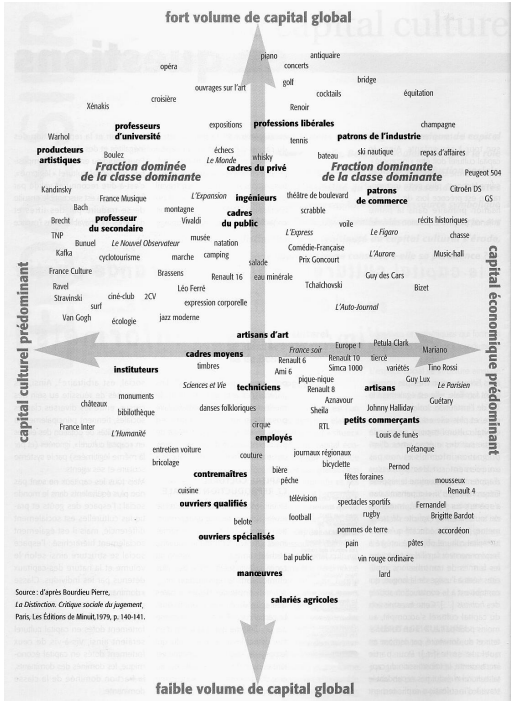
Analyse des données

1. Explorer la dimensionnalité des données
2. Analyse en composantes principales
3. Analyse factorielle des correspondances
4. Analyse des correspondances multiples
5. Le package `FactoMineR`



Explorer la dimensionnalité des données

L'analyse géométrique des données (AGD) correspond à une famille de méthodes d'analyse des données, développées par des mathématiciens français (Jean-Paul Benzécri).
Comme beaucoup de méthodes statistiques, l'AGD a au départ partie liée avec la psychologie et la psychométrie : une ou plusieurs dimensions de l'intelligence ? Mais aussi à la biométrie (de Quételet et son homme moyen au facteur des premiers généticiens).
Mais se développe rapidement en sociologie (chez Bourdieu notamment) et dans bien d'autres domaines (analyse sensorielle, TAL, …)
Dites aussi « analyse des données », « analyse des données à la française », « analyse factorielle », …

Explorer la dimensionnalité des données

L'idée consiste à dégager, d'un tableau de données défini par un certain nombre de variables (colonnes) et d'individus (lignes), les structures sous-jacentes à ces données : réalisme épistémologique ?

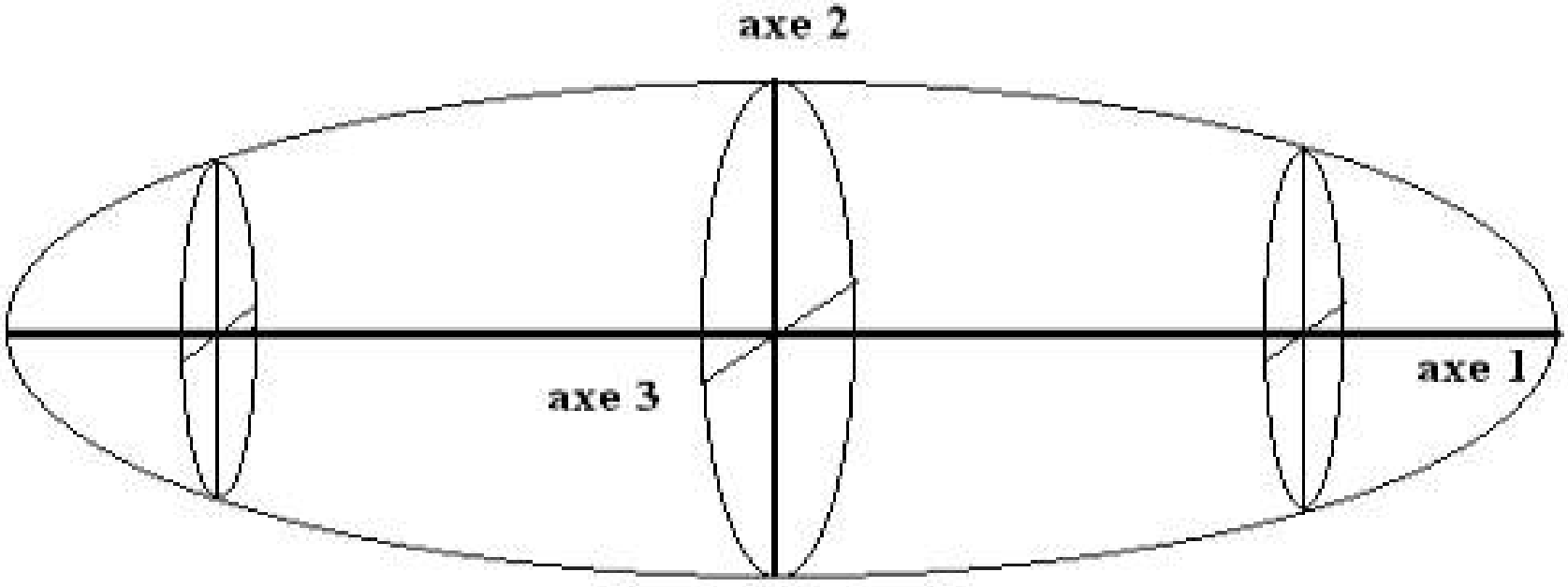
Pour cela, on procède à partir de l'idée de distance entre deux points, dans un univers à *N* dimensions (*N* étant le nombre de variables considérées).

Le statut des variables est symétrique : pas de notion de causalité.

La structure des données n'est pas modélisée : pas d'hypothèse sur la distribution des variables.

Explorer la dimensionnalité des données

Explorer la dimensionnalité des données



Explorer la dimensionnalité des données

Différents scénarios d'usage :

- beaucoup de cas : réduire à un faible nombre de cas typiques
- beaucoup de variables homogènes : réduire à un faible nombre de dimensions
- beaucoup de variables hétérogènes : réduire à un faible nombre de dimensions
- beaucoup de cas et de variables : réduire à un faible nombre de cas et/ou de dimensions

Dans tous les cas : réduction d'information.

Explorer la dimensionnalité des données

Les méthodes relevant de l'AGD sont nombreuses, et de nouvelles continuent d'apparaître (voir par exemple l'ACM spécifique ou l'analyse spécifique de classe).
Ici, quelques méthodes seulement sont présentées, mais ce sont les plus fondamentales : ACP, AFC et ACM.

Analyse en composantes principales

L'ACP :

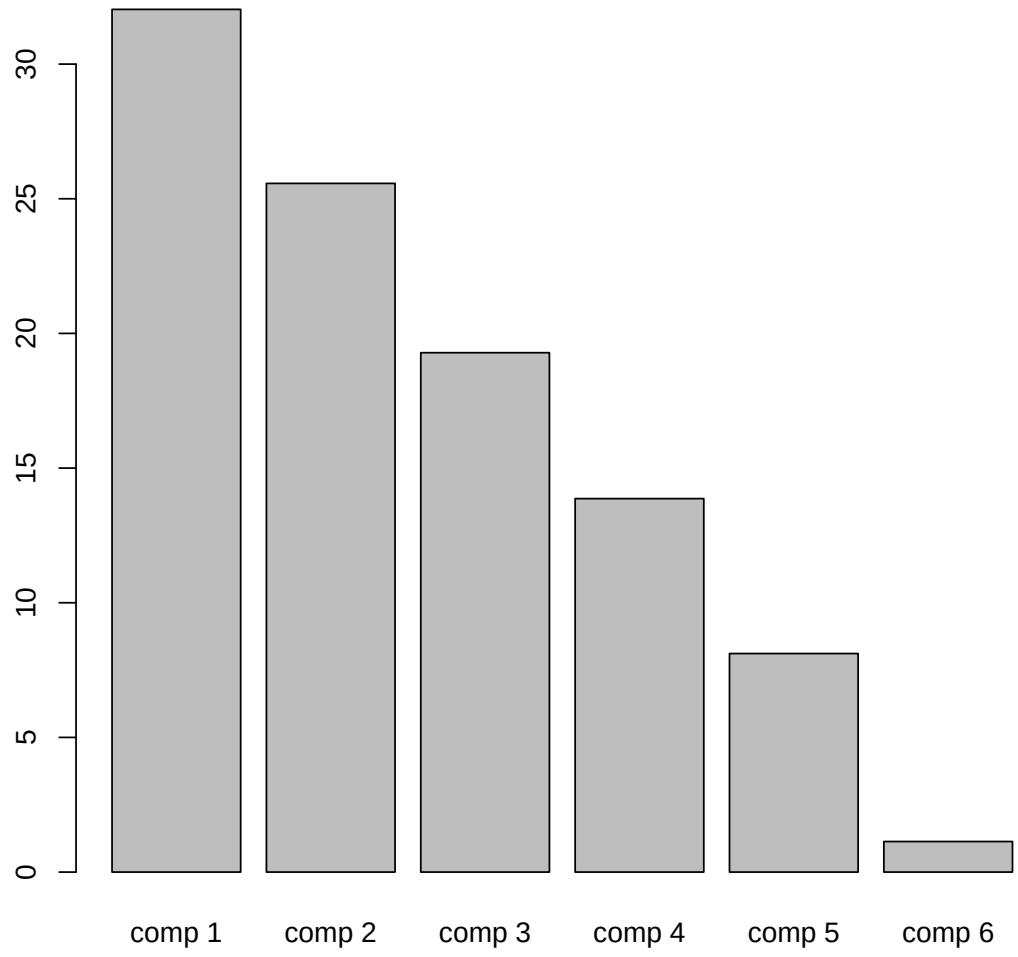
- porte sur des variables actives continues
- décrit un espace des variables, organisé en dimensions appelées composantes principales
- décrit un espace des individus, organisé selon les mêmes dimensions mais non directement commensurable
- repose sur la métrique du coefficient de corrélation
- peut prendre en compte des variables illustratives, continues ou qualitatives, mais aussi des individus supplémentaires.

Analyse en composantes principales

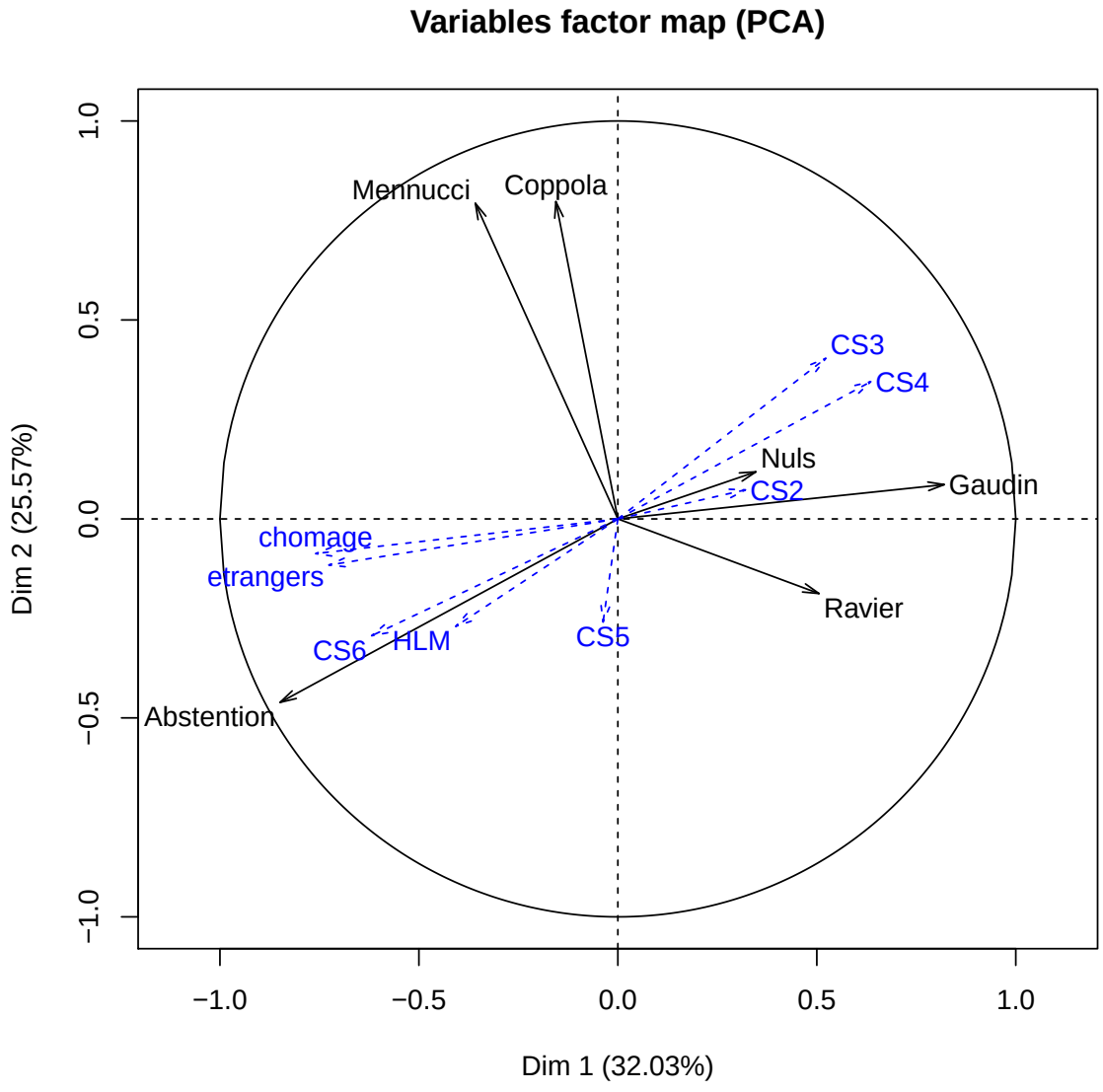
L'ACP consiste à projeter le nuage de points à N dimensions (N étant le nombre de variables continues caractérisant un ensemble d'individus) de manière successive afin de maximiser la variance des n premières composantes principales succesives ($n < N$). Chaque composante principale est ainsi caractérisée par :

- une part de la variance (ou inertie) totale résumée. La somme de l'inertie de toutes les composantes principales est égale à 100 %.
- une transformation linéaire des variables
- un vecteur de corrélations avec les variables.
- un vecteur de coordonnées des individus sur cette composante principale.

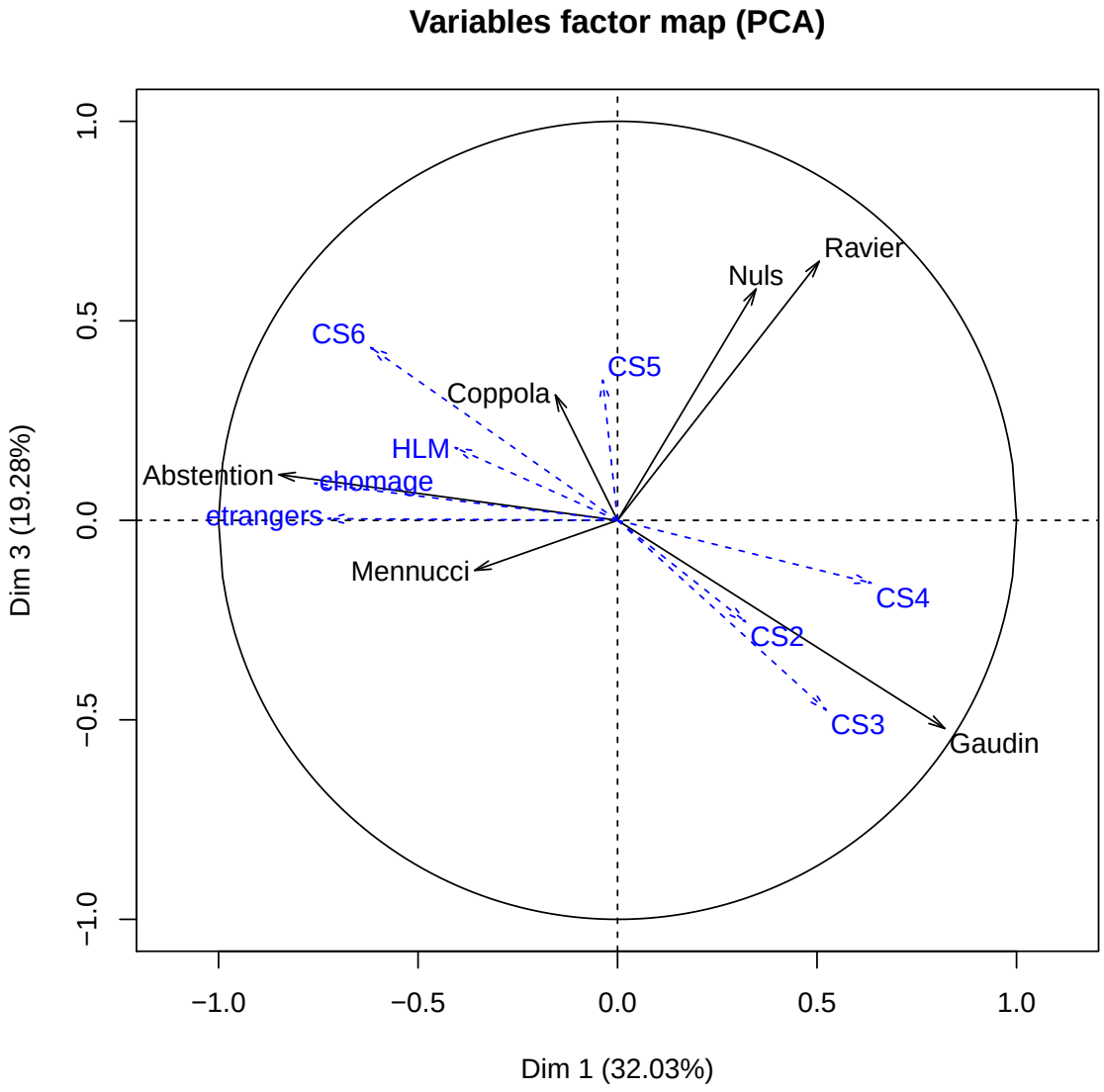
L’inertie des premières composantes principales



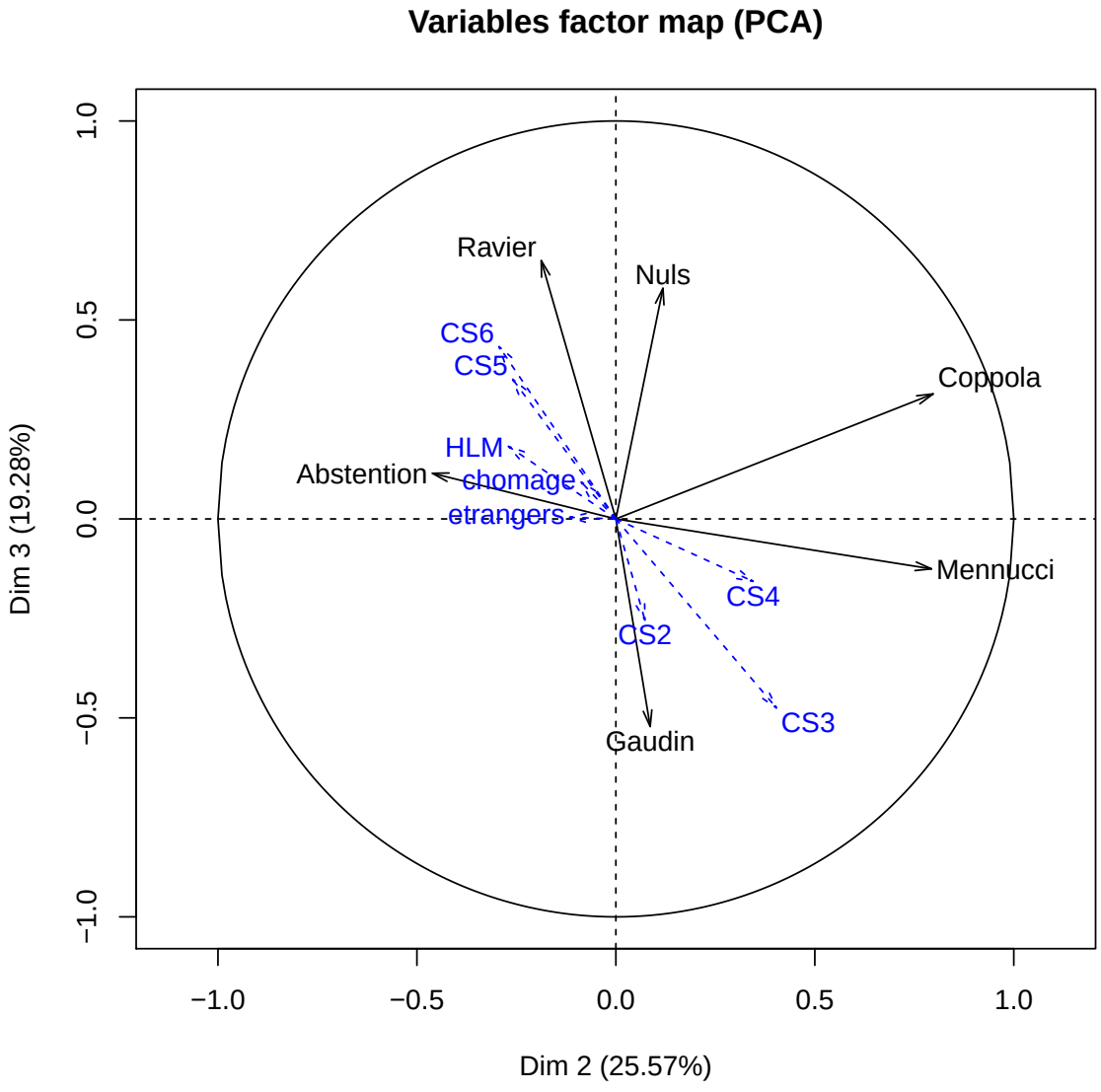
Le cercle des corrélations



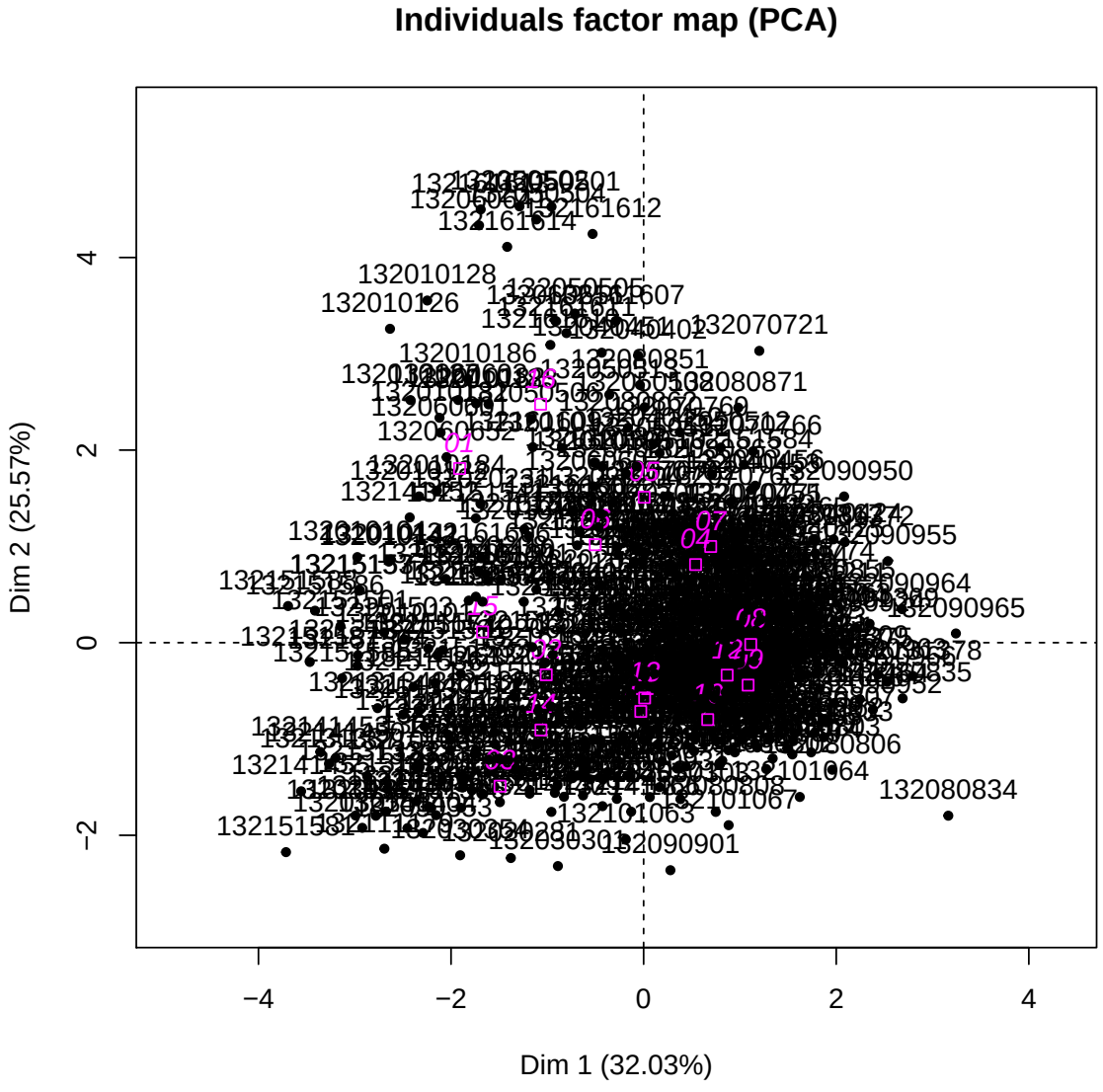
Le cercle des corrélations



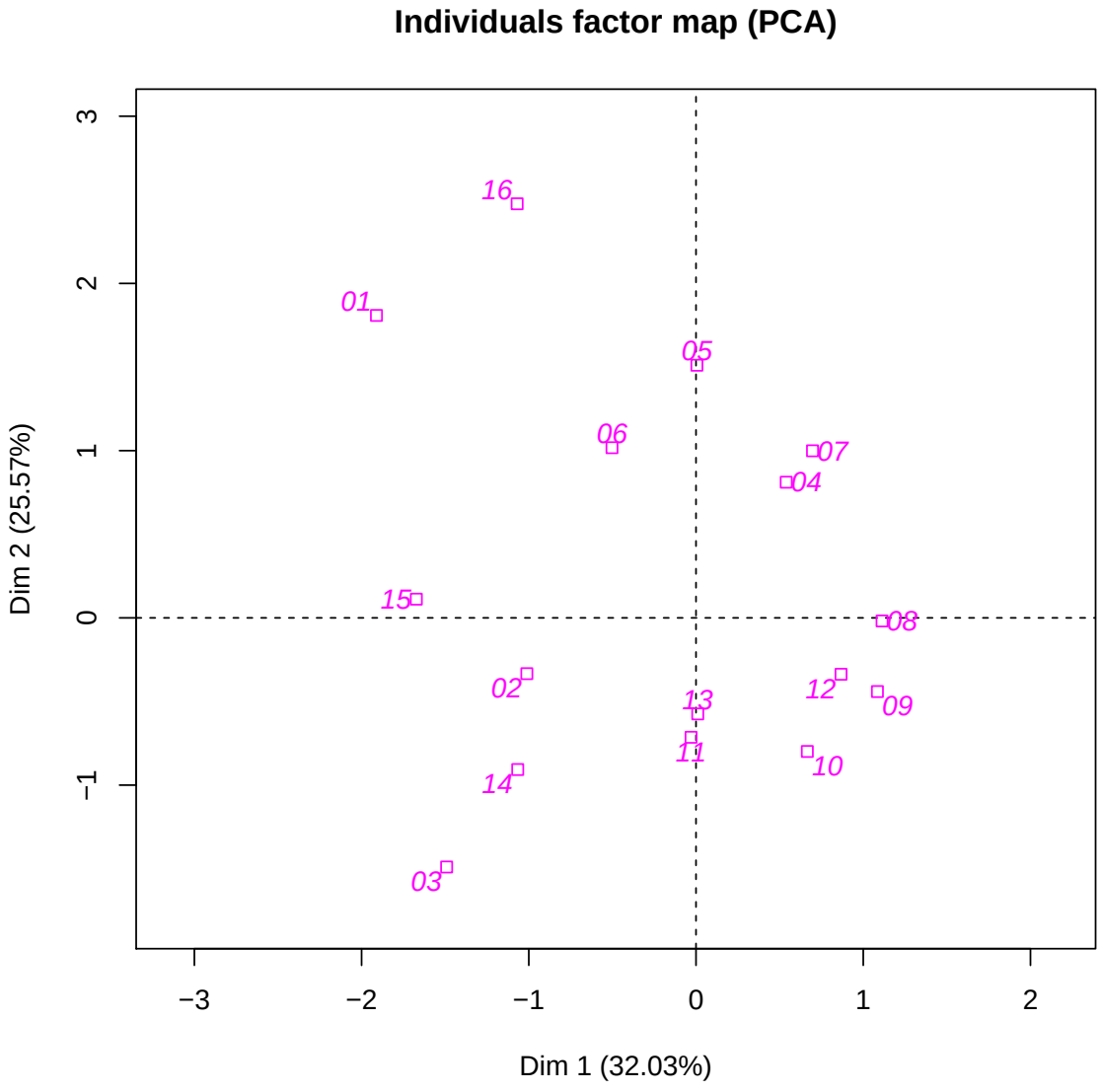
Le cercle des corrélations



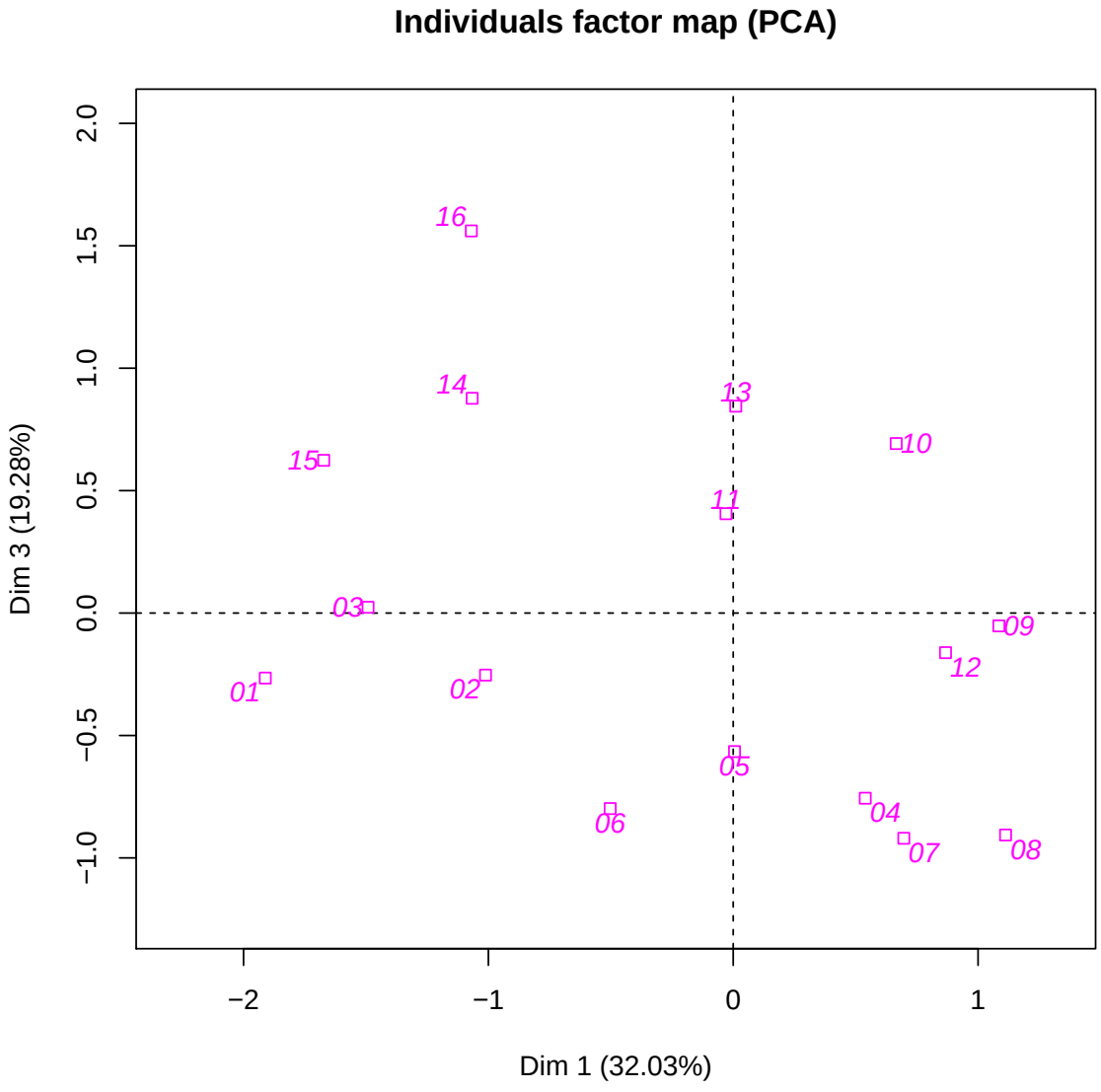
Le nuage des individus



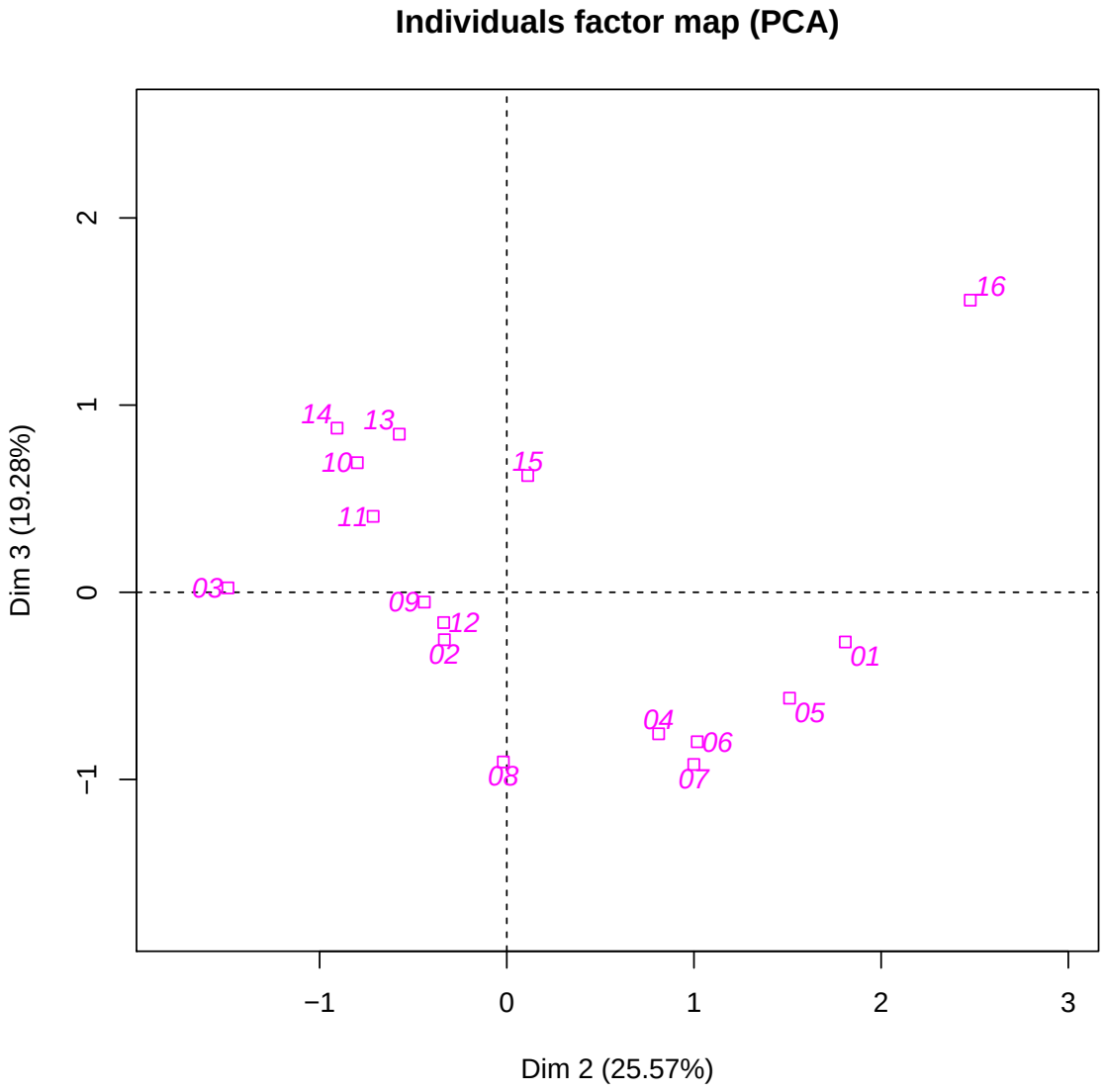
Le nuage des individus



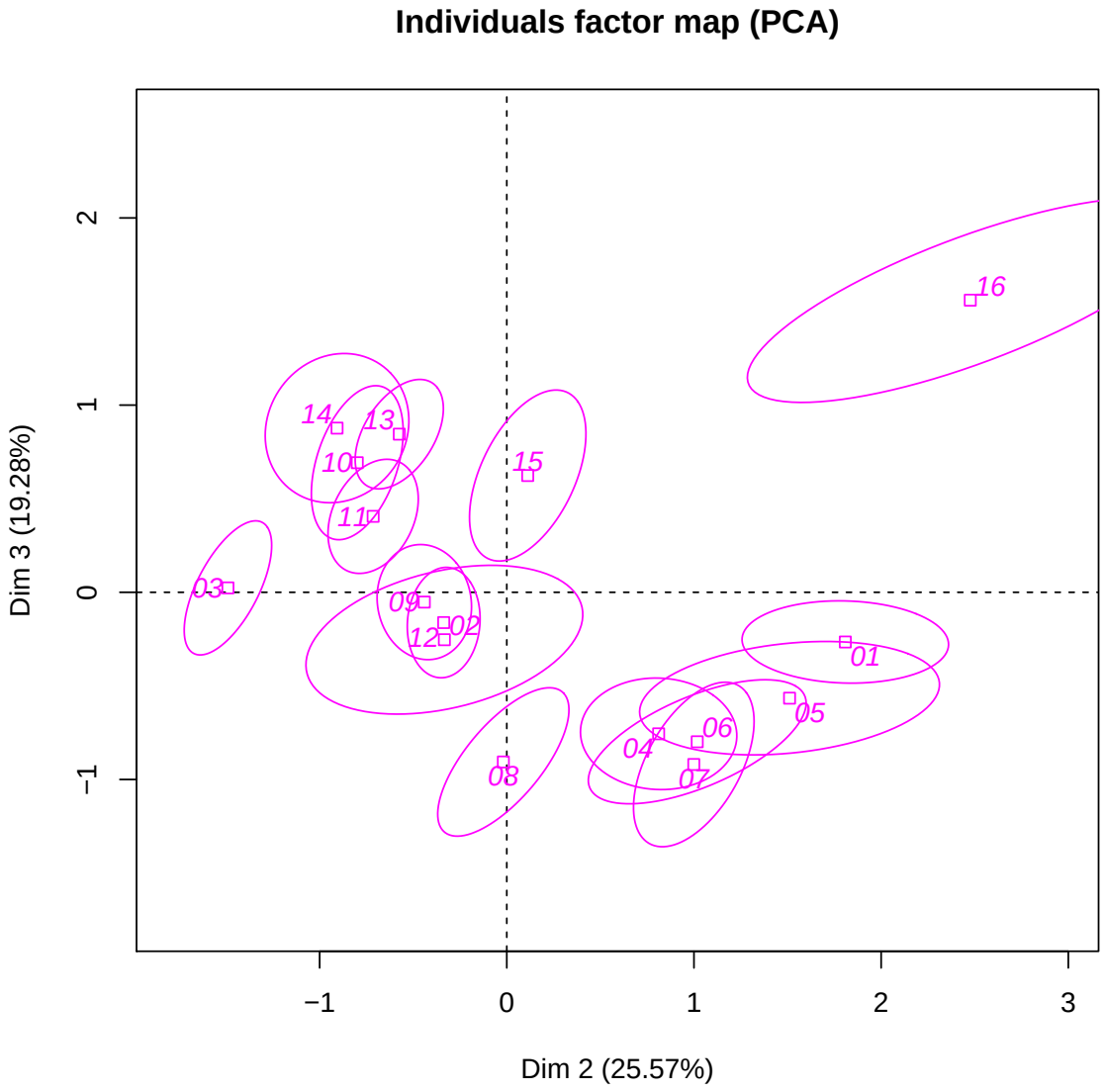
Le nuage des individus



Le nuage des individus



Le nuage des individus



Caractériser les dimensions

	correlation	p.value
Gaudin	0.82	0.00
CS4	0.64	0.00
CS3	0.52	0.00
Ravier	0.51	0.00
Nuls	0.35	0.00
CS2	0.32	0.00
Coppola	-0.16	0.00
Mennucci	-0.36	0.00
HLM	-0.41	0.00
CS6	-0.62	0.00
etrangers	-0.73	0.00
chomage	-0.76	0.00
Abstention	-0.85	0.00

TABLE: Dimension 1

Analyse factorielle des correspondances

L'AFC :

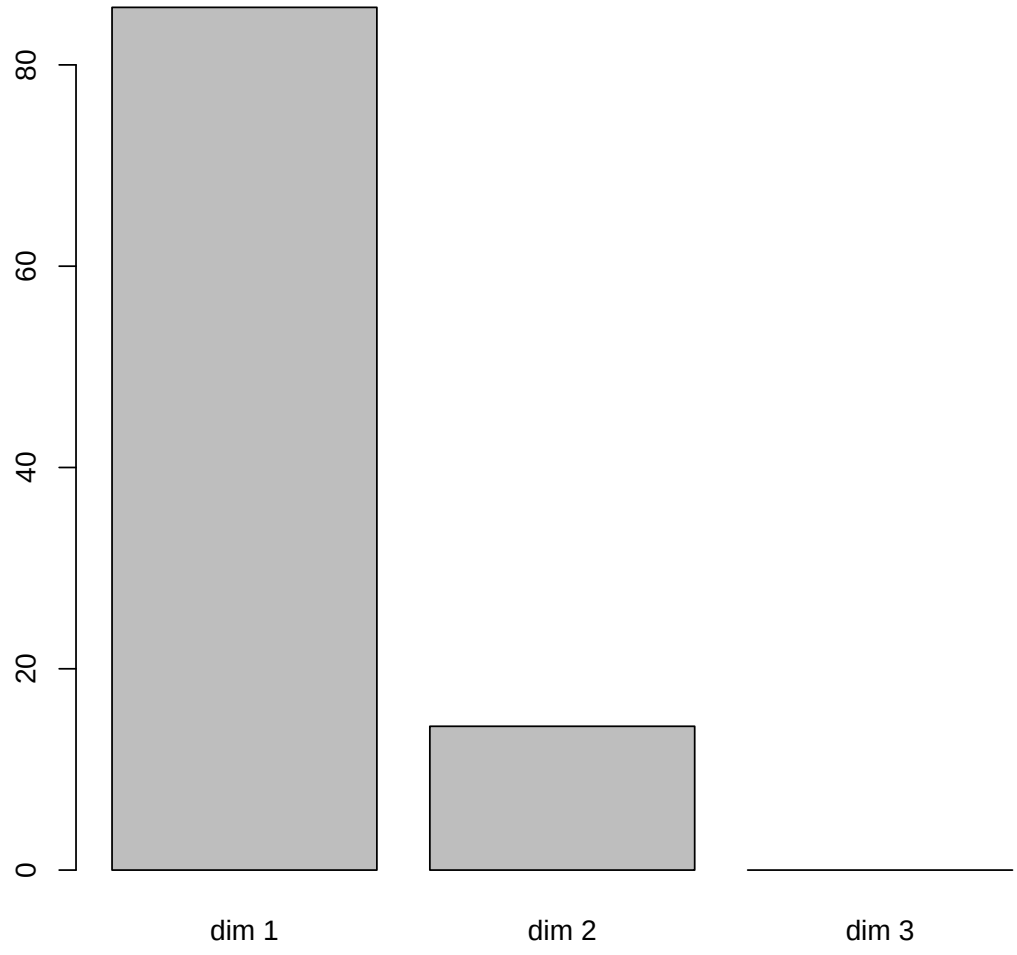
- porte sur un tableau de contingence, c'est-à-dire le croisement de deux variables qualitatives. Lignes et colonnes sont symétriques.
- décrit un espace des modalités. Les individus sont subsumés sous les modalités.
- repose sur la métrique du χ^2.
- peut prendre en compte des variables illustratives (sous la forme de colonnes supplémentaires).

Analyse factorielle des correspondances

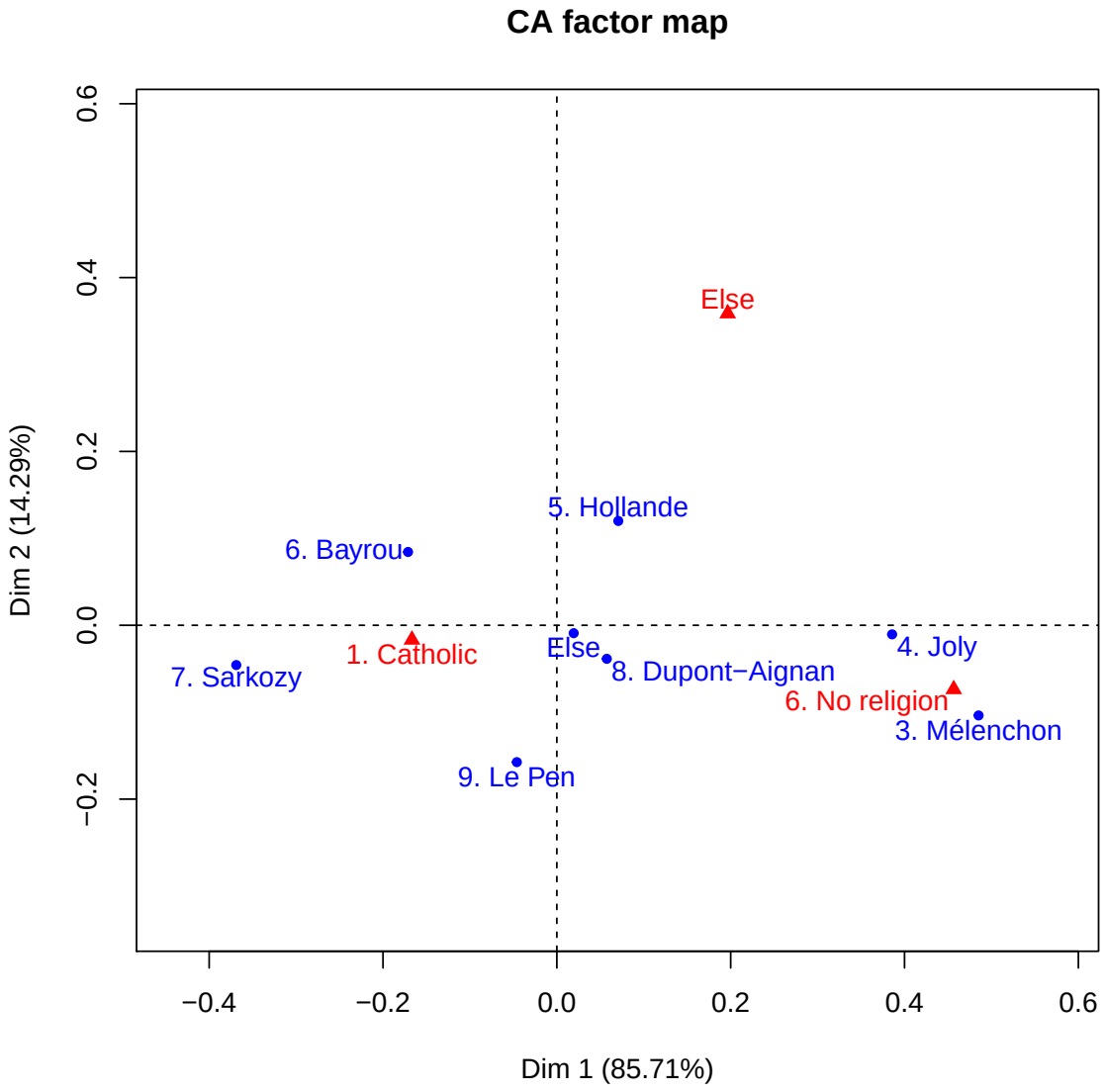
L'AFC consiste, à partir d'un tableau de contingences, à rechercher quelles sont les lignes les plus similaires (resp. différentes) du point de vue de leur association à des colonnes.
Les dimensions dégagées sont ici appelées facteurs. Chaque facteur est caractérisé par :

- une part de la variance (ou inertie) totale résumée. La somme de l'inertie de toutes les composantes principales est égale à 100 %.
- un vecteur de coordonnées de chaque modalité (pour les lignes comme pour les colonnes).

L’inertie des premiers facteurs



Le nuage des modalités



Caractériser les facteurs

	coord
7. Sarkozy	-0.37
6. Bayrou	-0.17
9. Le Pen	-0.05
Else	0.02
8. Dupont-Aignan	0.06
5. Hollande	0.07
4. Joly	0.39
3. Mélenchon	0.49

TABLE: Dimension 1 – lignes

	coord
1. Catholic	-0.17
Else	0.20
6. No religion	0.46

TABLE: Dimension 1 – colonnes

Analyse des correspondances multiples

L'ACM :

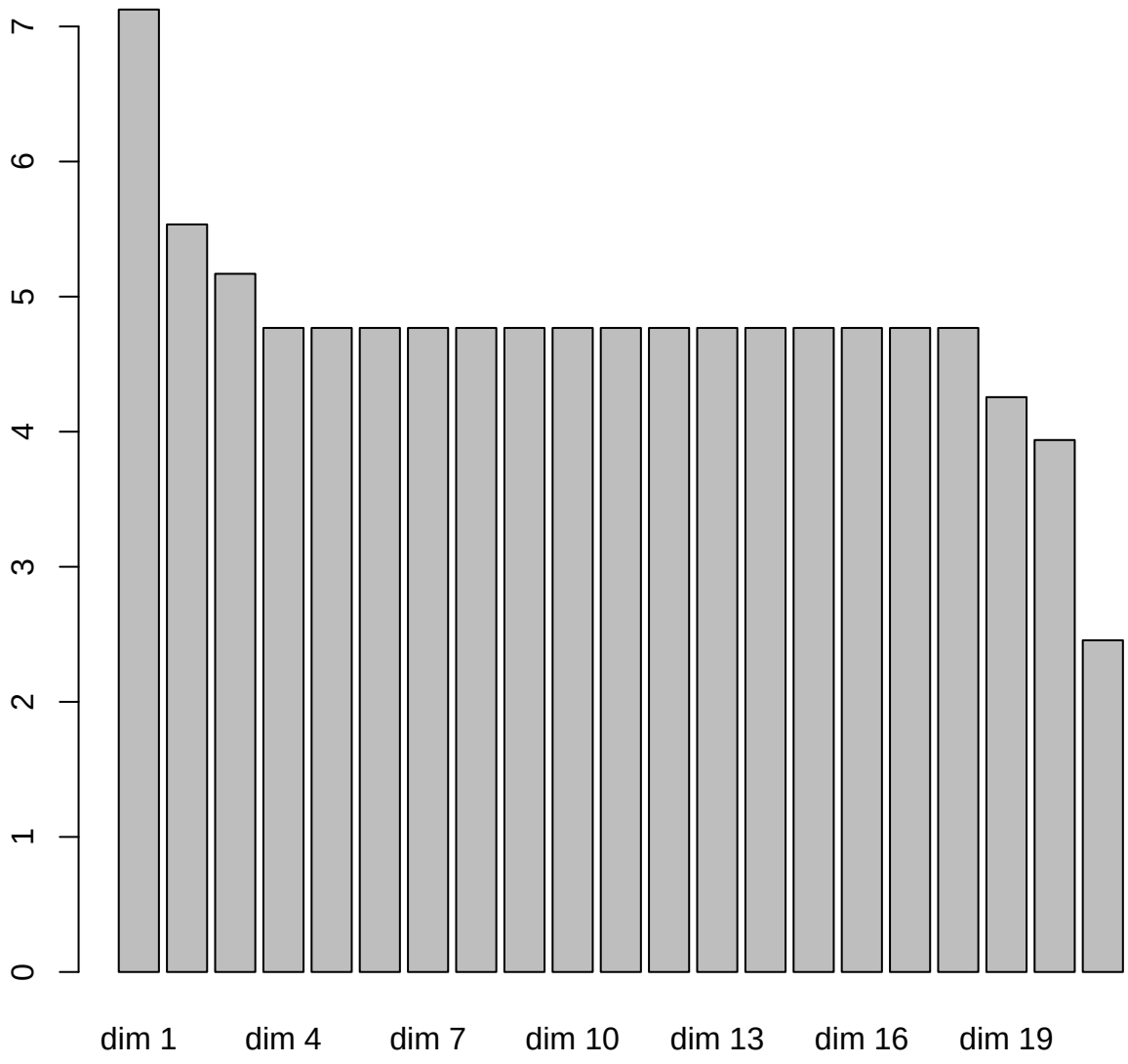
- porte sur des variables actives catégorielles (c'est une généralisation de l'AFC)
- décrit un espace commun des modalités et des individus, organisé en dimensions appelées dimensions, axes ou facteurs.
- repose sur la métrique du χ^2.
- peut prendre en compte des variables illustratives, qu'elles soient qualitatives ou continues, ainsi que des individus supplémentaires.

Analyse des correspondances multiples

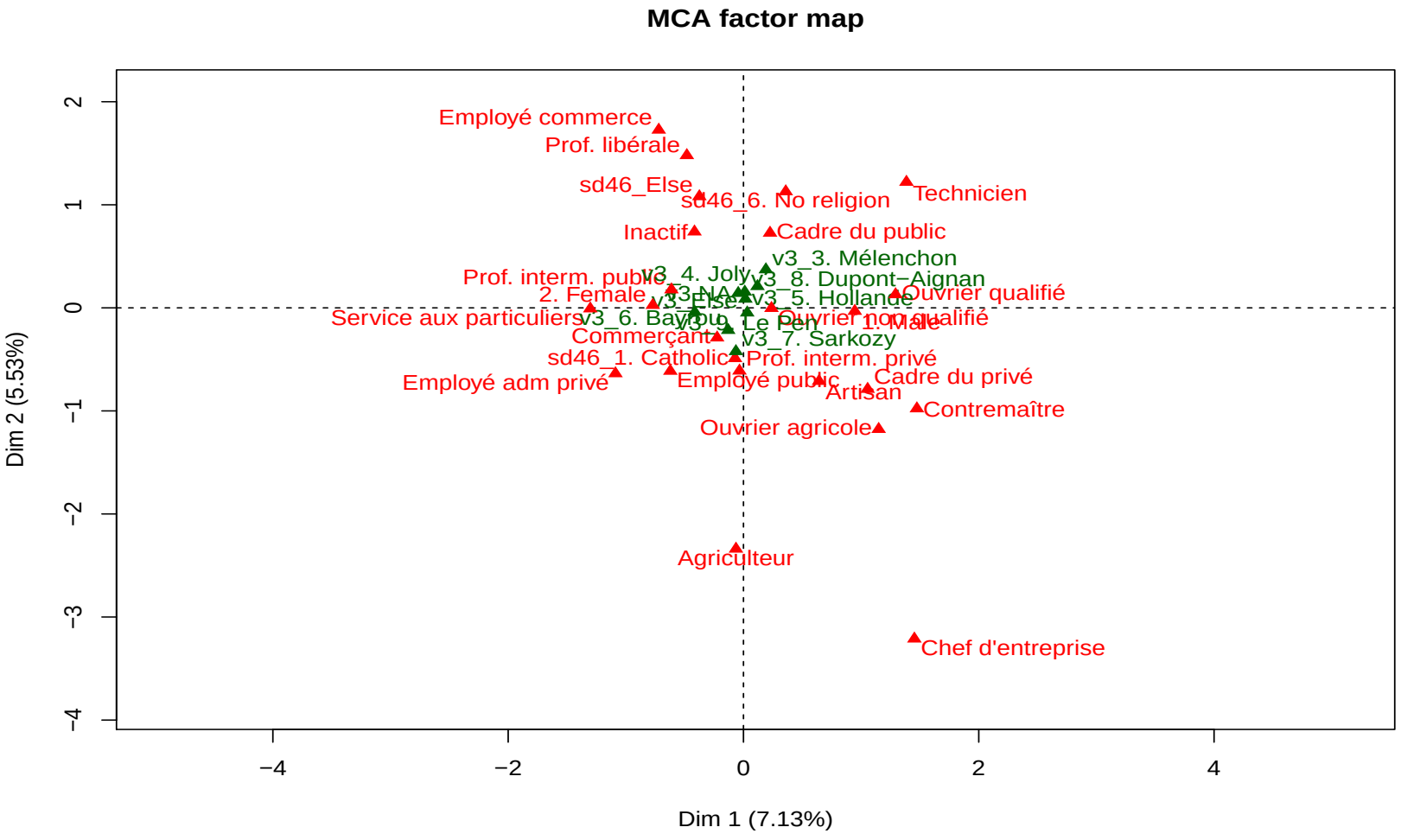
L'ACM consiste à rechercher, à partir d'un tableau d'individus en lignes et de variables qualitatives en colonnes, à rechercher quelles sont les individus les plus proches les uns des autres, c'est-à-dire qui partagent le plus de modalités.
De fait, on se concentre généralement sur les modalités plutôt que sur les individus eux-mêmes.
Les dimensions dégagées, appelées axes ou facteurs, sont caractérisées par :

- une part de la variance (ou inertie) totale résumée. La somme de l'inertie de toutes les composantes principales est égale à 100 %.
- un vecteur de coordonnées pour chaque modalité.
- des tests d'association avec les modalités (`v.test`).
- des tests d'association avec les variables (η^2).

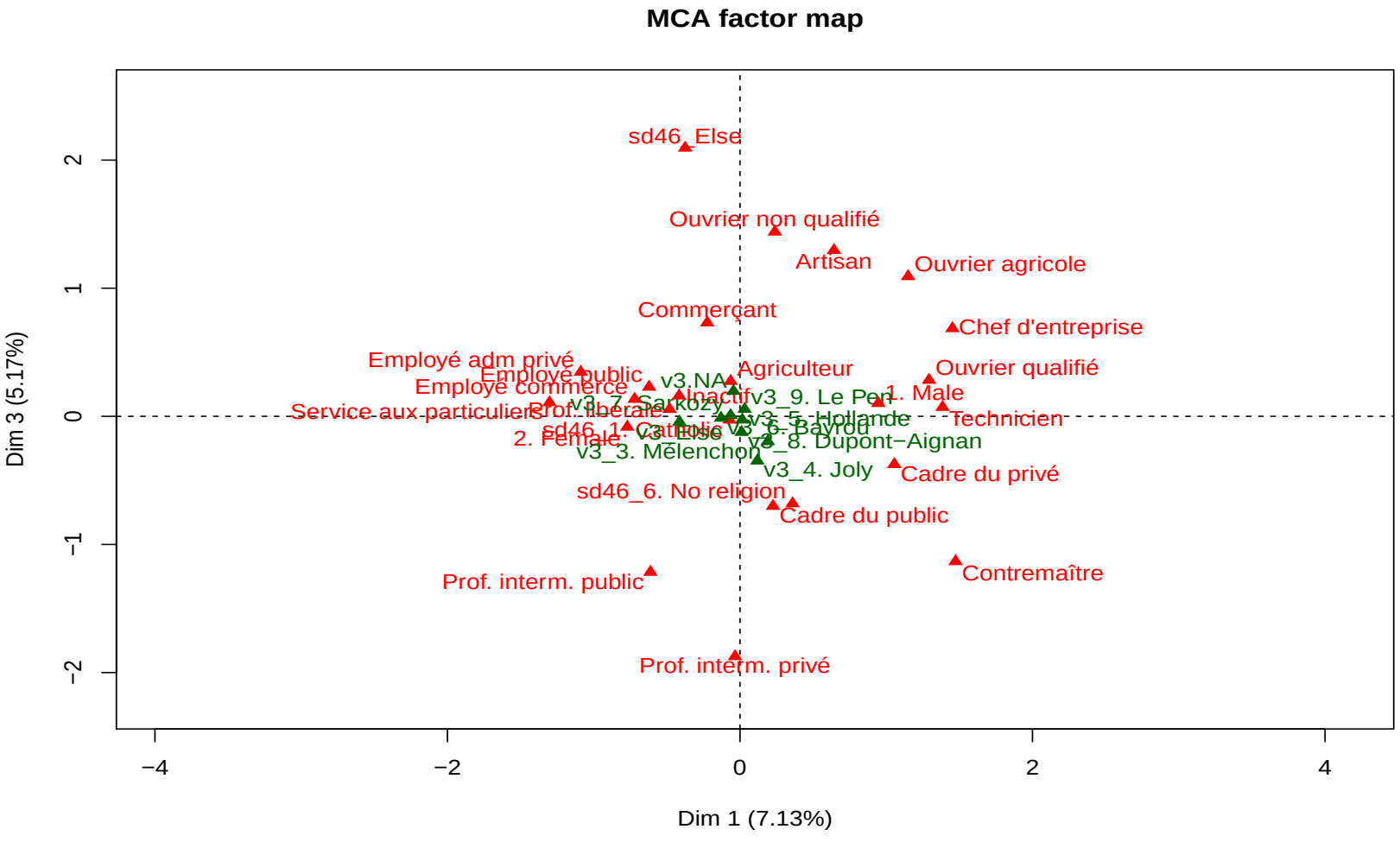
L’inertie des premiers facteurs



Le nuage des modalités



Le nuage des modalités



Caractériser les facteurs

	Estimate	p.value
1. Male	0.61	0.00
Ouvrier qualifié	0.80	0.00
Cadre du privé	0.64	0.00
Technicien	0.87	0.00
Contremaître	0.93	0.00
sd46_6. No religion	0.33	0.00
Artisan	0.34	0.00
Chef d'entreprise	0.92	0.00
sd46_1. Catholic	0.02	0.00
Ouvrier agricole	0.70	0.00
Ouvrier non qualifié	0.06	0.00
v3_3. Mélenchon	0.16	0.00
Cadre du public	0.05	0.01
Prof. libérale	-0.45	0.05
v3_Else	-0.27	0.03
sd46_Else	-0.20	0.00
Employé commerce	-0.62	0.00
Inactif	-0.40	0.00
Prof. interm. public	-0.54	0.00
Employé public	-0.55	0.00
Employé adm privé	-0.88	0.00
Service aux particuliers	-1.03	0.00
2. Female	-0.61	0.00

TABLE: Dimension 1

Le package FactoMineR

Il existe plusieurs packages proposant des méthodes d'AGD dans R. On utilisera, tout au long de cette séance, le package FactoMineR, français, complet et bien documenté (http://factominer.free.fr).

À noter

Il existe une interface graphique pour utiliser FactoMineR. Il faut installer le package Rcmdr (voir procédure sur le site de FactoMineR).

Le package `FactoMineR`

Les principales fonctions sont `PCA`, `CA` et `MCA`. Comme dans le cas des régressions, on stocke le résultat dans un objet qu'on peut ensuite exploiter.
Les sorties graphiques de `FactoMineR` se sont améliorées mais demeurent perfectibles. Il est tout-à-fait possible d'utiliser les données brutes pour créer ses propres graphiques, voir par exemple ce tutoriel, ce package ou encore ce package.

Le package `FactoMineR`

Pour apprendre à utiliser `FactoMineR`, pratiquons des exercices. Pour l'ACP, voir le script `acpMarseille.R` ; pour l'AFC et l'ACM, le script `acFES.R`.