Analyse Exploratoire des Données Multidimensionnelles

DU Dataviz

Magali Champion

Université de Paris

02-03/03/2023

Introduction

Qu'est-ce que l'analyse exploratoire?

L'**analyse exploratoire des données** (AED) est utilisée par les spécialistes des données pour analyser et étudier les ensembles de données puis résumer leurs principales caractéristiques, souvent en employant des méthodes de visualisation des données. (IBM Cloud Education)

Motivations :

- découvrir des règles, relations, dépendances à travers une grande quantité de données, Apprentissage non-supervisé
- utiliser un ensemble de données pour prédire des informations, des comportements. Apprentissage supervisé

Comment trouver un diamant dans un tas de charbon sans se salir les mains?

CEO de SAS

Introduction

Apprentissage non-supervisé

On considère :

- p variables explicatives $(X^1, ..., X^p)$,
- un n-échantillon $(x_i^1, ..., x_i^p)_{1 \leq i \leq n}$ de $(X^1, ..., X^p)$.

Motivation : découvrir des règles, relations, dépendances dans les données $(x_i^1, ..., x_i^p)_{1 \leq i \leq n}$.

Introduction

Apprentissage supervisé

On considère :

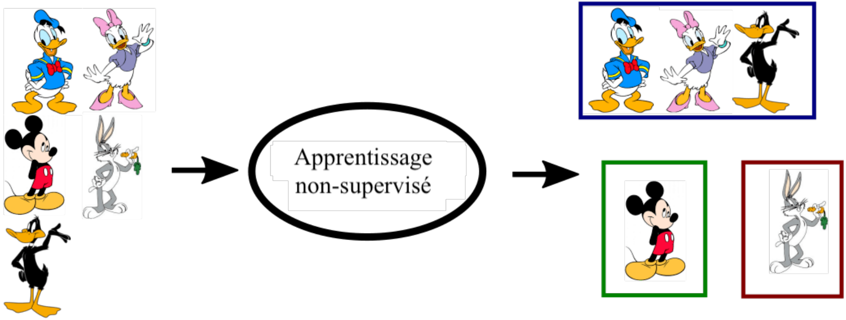
- p variables explicatives $(X^1, ..., X^p)$,
- un vecteur d'observations Y à expliquer,
- un n-échantillon $(x_i^1, ..., x_i^p, y_i)_{1 \leq i \leq n}$ de $(X^1, ..., X^p, Y)$.

Motivation : utiliser les observations $(x_i^1, ..., x_i^p, y_i)_{1 \leq i \leq n}$ pour prédire des informations, des comportements.

On parle d'apprentissage supervisé car les $(y_i)_{1 \leq i \leq n}$ *permettent de guider le processus d'estimation.*

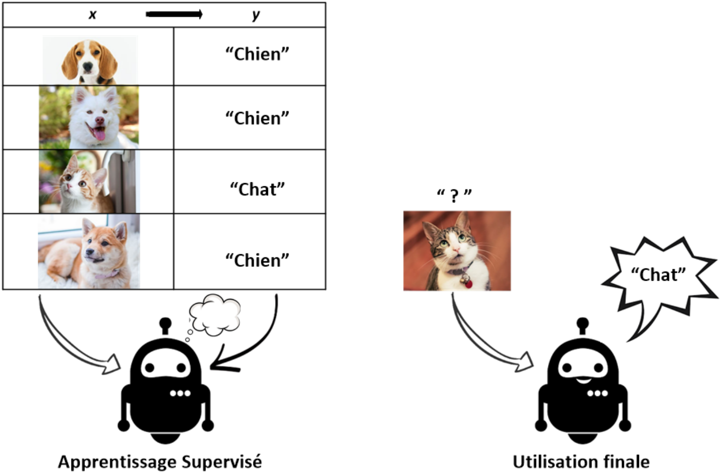
Introduction

Apprentissage supervisé vs non-supervisé



Introduction

Apprentissage supervisé vs non-supervisé



Introduction

Exemples d'application

- Entreprise et relation clients :
  - création de profils clients
  - ciblage de clients potentiels
- Finances et assurances :
  - minimisation de risques financiers
  - détection de fraudes
- Biomédical :
  - analyse du génome
  - identification de sous-groupes de patients
- Internet :
  - détection de spam
  - mise au point de systèmes de recommandation

Introduction

Données et nature des variables

On considère une matrice de données X et un vecteur d'observations Y à expliquer. Les observations portent sur p variables, mesurées sur n individus. Il existe plusieurs situations qui nécessitent l'utilisation d'outils différents selon :

- la nature des variables : discrètes (catégorielles ou ordinales) ou continues
- la dimension des données : univariées, bivariées, multivariées

Introduction

Données et nature des variables

On considère une matrice de données X et un vecteur d'observations Y à expliquer. Les observations portent sur p variables, mesurées sur n individus. Il existe plusieurs situations qui nécessitent l'utilisation d'outils différents selon :

- la nature des variables : discrètes (catégorielles ou ordinales) ou continues
- la dimension des données : univariées, bivariées, multivariées

```
##      Sex Wr.Hnd W.Hnd Pulse Exer Smoke Height    Age
## 1 Female   18.5 Right    92 Some Never 173.00 18.250
## 2   Male   19.5  Left   104 None Regul 177.80 17.583
## 3   Male   20.0 Right    35 Some Never 165.00 23.667
## 4 Female   18.0 Right    64 Some Never 172.72 21.000
## 5   Male   17.7 Right    83 Freq Never 182.88 18.833
## 6 Female   17.0 Right    74 Freq Never 157.00 35.833
```

Section 1

Statistiques descriptives

Table de données

A titre d'exemple, nous utiliserons le jeu de données `enquete`, qui contient des données relatives à 237 étudiants :

```
data <- read.csv2("enquete.csv",sep=",",dec=".")
head(data)
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse  Clap Exer Smoke Heigh
## 1 Female   18.5   18.0 Right  R on L    92  Left Some Never 173.0
## 2   Male   19.5   20.5  Left  R on L   104  Left None Regul 177.8
## 3   Male   20.0   20.0 Right Neither    35 Right Some Never 165.0
## 4 Female   18.0   17.7 Right  L on R    64 Right Some Never 172.7
## 5   Male   17.7   17.7 Right  L on R    83 Right Freq Never 182.8
## 6 Female   17.0   17.3 Right  R on L    74 Right Freq Never 157.0
##      Age
## 1 18.250
## 2 17.583
## 3 23.667
## 4 21.000
## 5 18.833
## 6 35.833
```

Table de données

Nature des variables

Les analyses descriptives effectuées dépendent de

- la nature des variables : discrètes (catégorielles ou ordinales) ou continues
- la dimension des données : univariées, bivariées, multivariées

```
str(data)
```

```
## 'data.frame':    168 obs. of  12 variables:
##  $ Sex   : chr  "Female" "Male" "Male" "Female" ...
##  $ Wr.Hnd: num  18.5 19.5 20 18 17.7 17 20 18.5 17 19.5 ...
##  $ NW.Hnd: num  18 20.5 20 17.7 17.7 17.3 19.5 18.5 17.2 20.2 ...
##  $ W.Hnd : chr  "Right" "Left" "Right" "Right" ...
##  $ Fold  : chr  "R on L" "R on L" "Neither" "L on R" ...
##  $ Pulse : int  92 104 35 64 83 74 72 90 80 66 ...
##  $ Clap  : chr  "Left" "Left" "Right" "Right" ...
##  $ Exer  : chr  "Some" "None" "Some" "Some" ...
##  $ Smoke : chr  "Never" "Regul" "Never" "Never" ...
##  $ Height: num  173 178 165 173 183 ...
##  $ M.I   : chr  "Metric" "Imperial" "Metric" "Imperial" ...
```

Analyses univariées

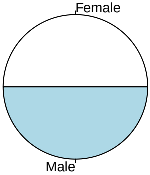
Indicateurs statistiques

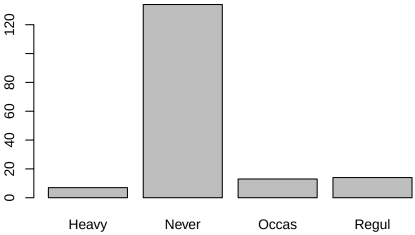
```
summary(data)
```

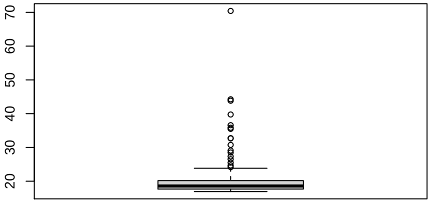
```
##      Sex                Wr.Hnd          NW.Hnd          W.Hnd
##  Length:168         Min.   :13.0   Min.   :12.50   Length:168
##  Class :character   1st Qu.:17.5   1st Qu.:17.50   Class :charact
##  Mode  :character   Median :18.5   Median :18.50   Mode  :charact
##                     Mean   :18.8   Mean   :18.73
##                     3rd Qu.:20.0   3rd Qu.:20.00
##                     Max.   :23.2   Max.   :23.50
##      Fold               Pulse             Clap                Exer
##  Length:168         Min.   : 35.00   Length:168          Length:16
##  Class :character   1st Qu.: 66.75   Class :character   Class :ch
##  Mode  :character   Median : 72.00   Mode  :character   Mode  :ch
##                     Mean   : 74.02
##                     3rd Qu.: 80.00
##                     Max.   :104.00
##     Smoke               Height          M.I                 Age
##  Length:168         Min.   :152.0   Length:168          Min.   :16
##  Class :character   1st Qu.:165.0   Class :character   1st Qu.:17
```

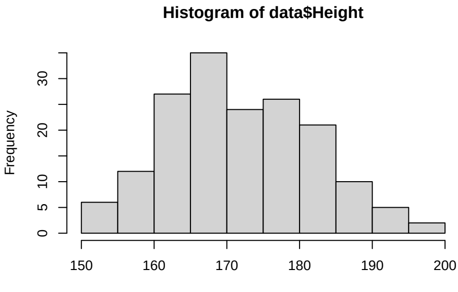
Analyses univariées

Graphiques





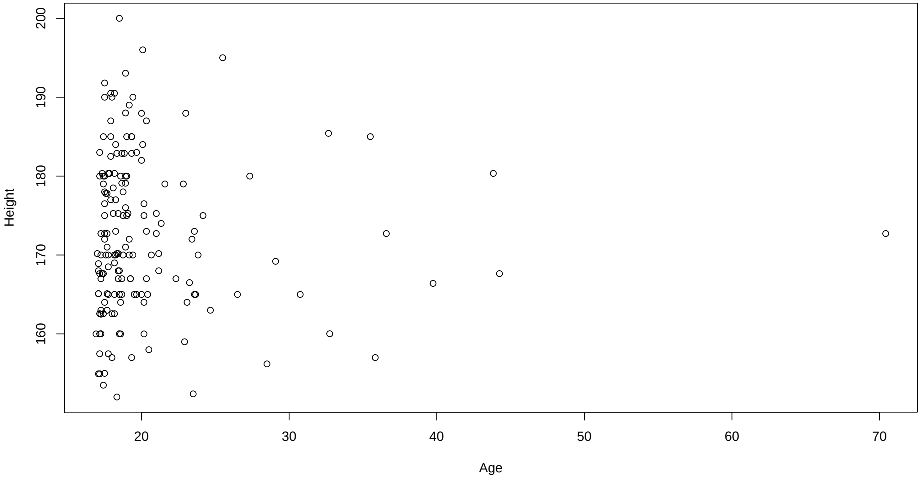




Analyses bivariées

Variables quantitatives

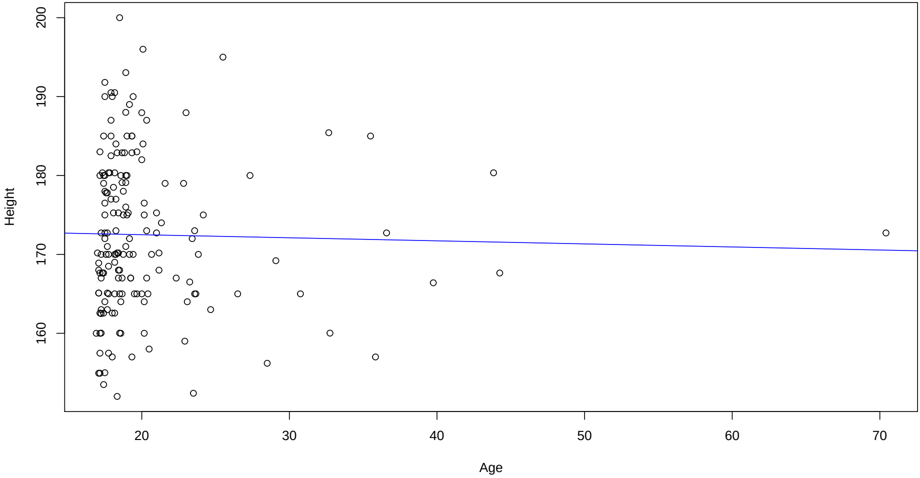
```
plot(data$Age,data$Height,xlab="Age",ylab="Height")
```



Analyses bivariées

Variables quantitatives

```
plot(data$Age,data$Height,xlab="Age",ylab="Height")
abline(lm(data$Height ~ data$Age), col = "blue")
```



Analyses bivariées

Variables quantitatives

Le lien entre les variables quantitatives peut être mesuré grâce à :

- la covariance

$$Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

- la corrélation

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1],$$

où $\bar{\cdot}$ représente la moyenne et $\sigma_{\cdot}$ l'écart-type.

```
cor(data$Age,data$Height)
```

```
## [1] -0.02372612
```

Analyses bivariées

Variables qualitatives

Table de contingence

```
table(data$W.Hnd,data$Sex)
```

```
##
##         Female Male
##   Left       5    7
##   Right     79   77
```

Test de χ_2 d'indépendance

- permet de tester si les variables qualitatives ont de l'influence l'une sur l'autre,
- la significativité du test est mesurée par la *p*-valeur.

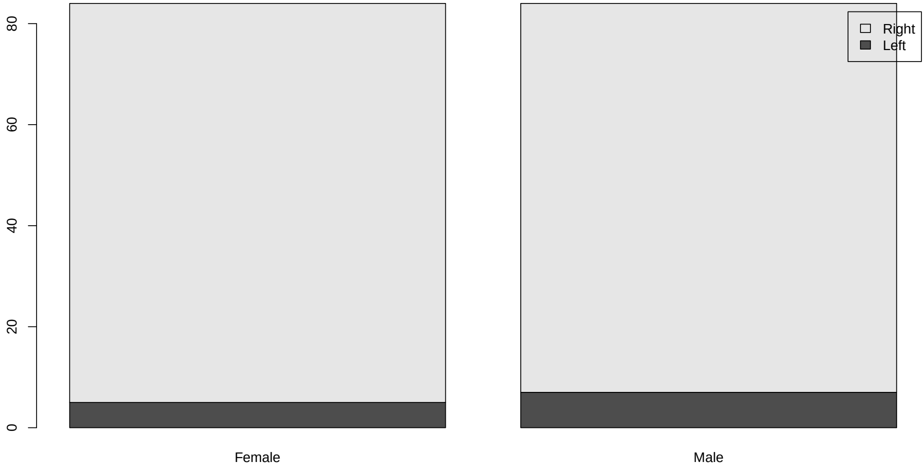
```
test <- chisq.test(table(data$W.Hnd,data$Sex))
test$p.value
```

```
## [1] 0.7645034
```

Analyses bivariées

Variables qualitatives

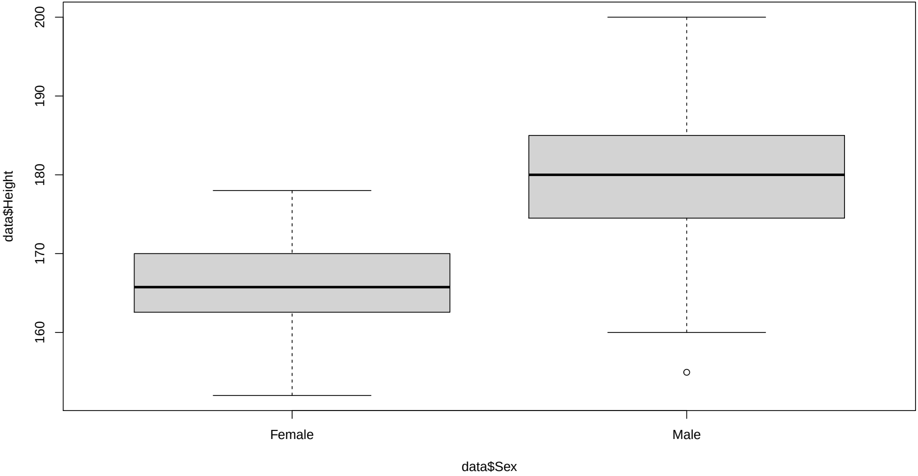
```
barplot(table(data$W.Hnd,data$Sex),legend=TRUE)
```



Analyses bivariées

Variables quantitatives-qualitatives

```
boxplot(data$Height~data$Sex)
```



Pour aller plus loin

Qu'est-ce que l'analyse factorielle?

L'**analyse factorielle** a pour objectif de rechercher des relations complexes entre des variables en résumant l'information en un petit nombre seulement de facteurs. Elle est particulièrement adaptée au cadre de la grande dimension.

Objectifs :

- Meilleure compréhension des données
- Visualisation des individus/variables
- Réduction de dimension

Il existe différentes méthodes d'analyse factorielle : l'analyse en composantes principales (ACP), l'analyse factorielle des correspondances (AFC), l'analyse des correspondances multiples (ACM), l'analyse factorielle de données mixtes (AFDM), l'analyse factorielle multiple (AFM) et l'analyse factorielle multiple hiérarchique (AFMH).

Section 2

Analyse en Composantes Principales

(ACP)

Analyse en Composantes Principales

Qu'est-ce que c'est?

L'**Analyse en Composantes Principales** construit des facteurs (composantes principales) qui résument l'information contenue dans un jeu de données sous la forme de combinaisons linéaires de variables.

Objectifs :

- Meilleure compréhension des données
- Visualisation des individus/variables
- Réduction de dimension

Méthode descriptive pour l'analyse multivariée de variables ***quantitatives***.

Analyse en Composantes Principales

Principe

Construction de :

- une matrice A de taille $p \times r$ ($r << p$) contenant en colonne les coefficients des combinaisons linéaires des anciennes variables (les vecteurs engendrant le nouvel espace),
- une matrice Z de taille $n \times r$ contenant les r nouvelles variables telles que :

$$Z = XA.$$

Analyse en Composantes Principales

Principe

Construction de :

- une matrice A de taille $p \times r$ ($r << p$) contenant en colonne les coefficients des combinaisons linéaires des anciennes variables (les vecteurs engendrant le nouvel espace),
- une matrice Z de taille $n \times r$ contenant les r nouvelles variables telles que :

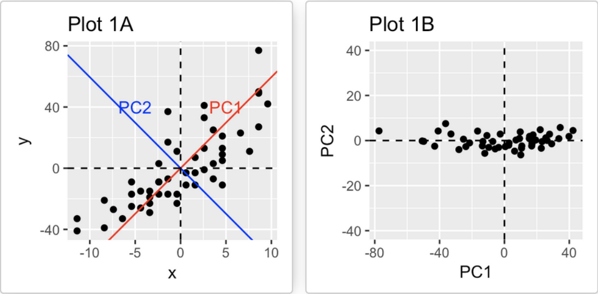
$$Z = XA.$$

Les composantes principales $Z^1, ..., Z^r$ sont construites de telle sorte à garder le plus d'information possible contenue dans $X^1, ..., X^p$: la variance des coordonnées des n individus sur chaque nouvel axe doit être maximale.

Analyse en Composantes Principales

Illustration I

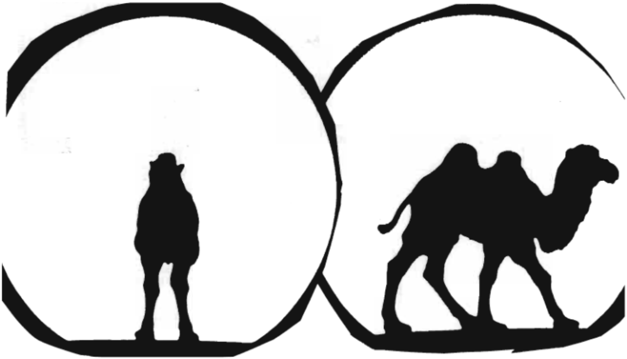
Ici, on voit que les composantes principales $Z^1, \ldots, Z^r$ sont construites de telle sorte à garder le plus d'information possible contenue dans $X^1, \ldots, X^p$. Le nuage de points se répartit bien sur l'axe, gardant la diversité initiale du nuage, ce qui ne serait pas le cas si tous les points étaient projetés au même endroit (variance nulle).



Analyse en Composantes Principales

Illustration II

Le point de vue du photographe est-il le bon pour que l'on puisse savoir s'il s'agit d'un chameau ou d'un dromadaire?



Analyse par Composantes Principales

Initialisation

Normalisation des données

Afin de limiter l'effet de trop “grosses” variables présentes, les variables $X^1, \ldots, X^p$ sont **centrées et réduites**.

```
X <- scale(X,center=TRUE,scale=TRUE)
```

Ceci a pour effet :

- de pouvoir comparer des variables à échelles différentes (*avantage*),
- de lisser le signal (*inconvénient*).

Remarque : centrer les données n'ayant aucun effet sur la forme du nuage, on les centre systématiquement. La standardisation des données n'est effectuée que dans le cas de variables d'unités différentes.

Analyse par Composantes Principales

Algorithme

Etape 1 : construction du 1er axe

Le 1er axe Z^1 est choisi comme étant la combinaison linéaire de $X^1, ...X^p$ de variance maximale :

$$Z^1 = X\alpha_1,$$

avec $\|\alpha_1\| = 1$ et $Var(X\alpha_1)$ maximale parmi les vecteurs de la forme $X\alpha$.

- $\alpha_1 \in \mathbb{R}^p$ représente la direction du 1er axe principal,
- $X\alpha_1 \in \mathbb{R}^n$ est l'ensemble des coordonnées du nuage de points sur cet axe.

Etape 2 : construction du 2ième axe

Le 2nd axe Z^2 est choisi comme étant la combinaison linéaire de $X^1, ...X^p$ de variance maximale :

$$Z^2 = X\alpha_2,$$

avec $\|\alpha_2\| = 1$ et $Var(X\alpha_2)$ maximale parmi les vecteurs de la forme $X\alpha$.

On y ajoute la **contrainte** :

$$\langle Z^2, Z^1 \rangle = 0.$$

Analyse par Composantes Principales

Algorithme

Etape k : construction du k-ième axe

De manière plus générale, Z^k est choisi comme étant la combinaison linéaire de $X^1, ...X^p$ de variance maximale :

$$Z^k = X\alpha_k,$$

où

$$\alpha_k = \underset{\alpha\in\mathbb{R}^p}{argmax} \;\; Var(X\alpha).$$

sous les contraintes :

$$\|\alpha_k\| = 1 \text{ et } \forall \ell \in [\![1, k-1]\!], \;\; {}^t\alpha_k {}^tXX\alpha_\ell = 0.$$

Par construction, tous les axes sont orthogonaux et ils sont ordonnés du plus informatif Z^1 au moins informatif Z^r.

Analyse par Composantes Principales

Algorithme

Etape k : construction du k-ième axe

De manière plus générale, Z^k est choisi comme étant la combinaison linéaire de $X^1, ...X^p$ de variance maximale :

$$Z^k = X\alpha_k,$$

où

$$\alpha_k = \underset{\alpha \in \mathbb{R}^p}{argmax} \ {}^t\alpha {}^tXX\alpha.$$

sous les contraintes :

$$\|\alpha_k\| = 1 \text{ et } \forall \ell \in [\![1, k-1]\!], \ {}^t\alpha_k {}^tXX\alpha_\ell = 0.$$

Par construction, tous les axes sont orthogonaux et ils sont ordonnés du plus informatif Z^1 au moins informatif Z^r.

Analyse par Composantes Principales

Algorithme

Etape k : construction du k-ième axe

De manière plus générale, Z^k est choisi comme étant la combinaison linéaire de $X^1, ...X^p$ de variance maximale :

$$Z^k = X\alpha_k,$$

où

$$\alpha_k = \underset{\alpha \in \mathbb{R}^p}{argmax} \ \ {}^t\alpha {}^tXX\alpha = {}^t\alpha\Sigma\alpha,$$

avec Σ la matrice de covariance empirique de X, sous les contraintes :

$$\|\alpha_k\| = 1 \text{ et } \forall \ell \in [\![1, k-1]\!], \ \ {}^t\alpha_k {}^tXX\alpha_\ell = 0.$$

Par construction, tous les axes sont orthogonaux et ils sont ordonnés du plus informatif Z^1 au moins informatif Z^r.

En pratique, d'un point de vue algorithmique :

- soit on trouve α_1 puis on projette tous les individus (qui sont des points de $\mathbb{R}^p$) sur $(\alpha_1)^\perp$. On relance alors la résolution du problème d'optimisation pour trouver α_2,...
- soit on utilise le fait que les α_k correspondent aux vecteurs propres de Σ (qui est diagonalisable car symétrique), ordonnés par ordre décroissant de leur valeur propre associée. Ceci permet d'obtenir tous les α_k d'un seul coup!

Analyse par Composantes Principales

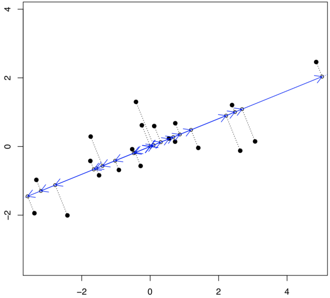
Choix du nombre de composantes

Déterminer le nombre r d'axes à retenir est une problématique centrale pour faire de la réduction de dimension. Il existe de nombreux critères basés sur :

- la part d'inertie :

$$r = \underset{k<p}{\operatorname{argmin}} \; \{\mathcal{I}_k > \tau\},$$

  où $\mathcal{I}_k$ est l'inertie de la composante k, qui mesure la dispersion des points autour du centre de gravité dans un nuage de points.

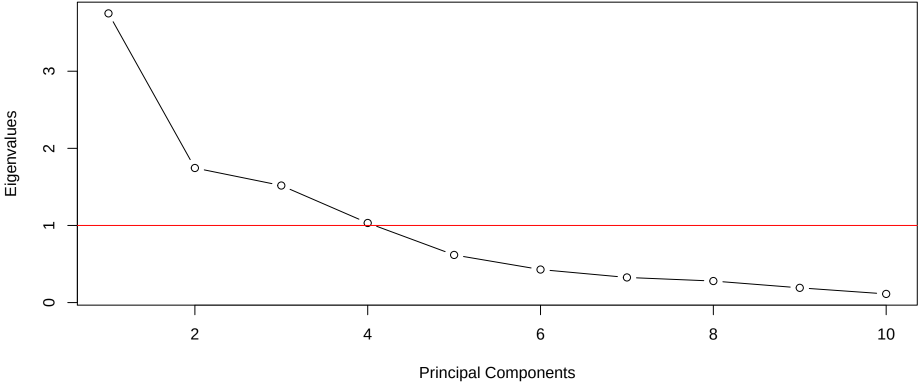


Analyse par Composantes Principales

Choix du nombre de composantes

Déterminer le nombre r d'axes à retenir est une problématique centrale pour faire de la réduction de dimension. Il existe de nombreux critères basés sur :

- la règle de Kaiser : on ne conserve que les valeurs propres supérieures à leur moyenne.

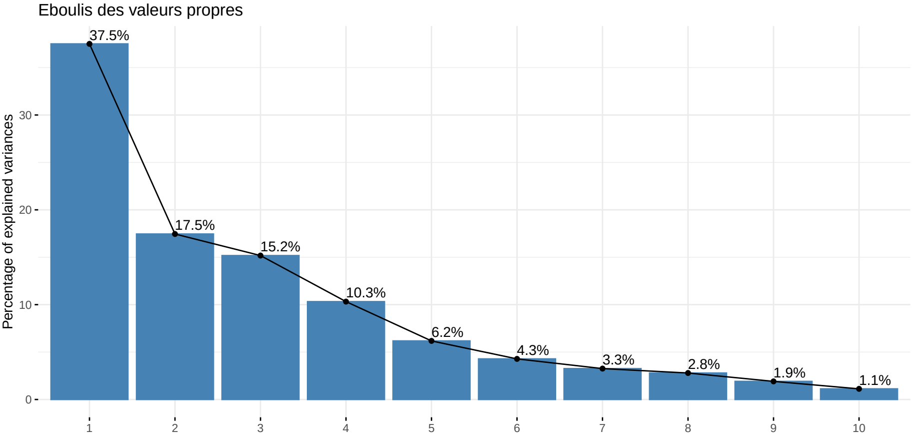


Analyse par Composantes Principales

Choix du nombre de composantes

Déterminer le nombre r d'axes à retenir est une problématique centrale pour faire de la réduction de dimension. Il existe de nombreux critères basés sur :

- l'éboulis des valeurs propres : graphique présentant la décroissance des valeurs propres. On cherche un coude dans le graphe pour déterminer r.



Analyse en Composantes Principales

Application sur R

A titre d'exemple, nous utiliserons le jeu de données `decathlon2`, qui contient des données relatives à des sportifs participant à des épreuves de décathlon :

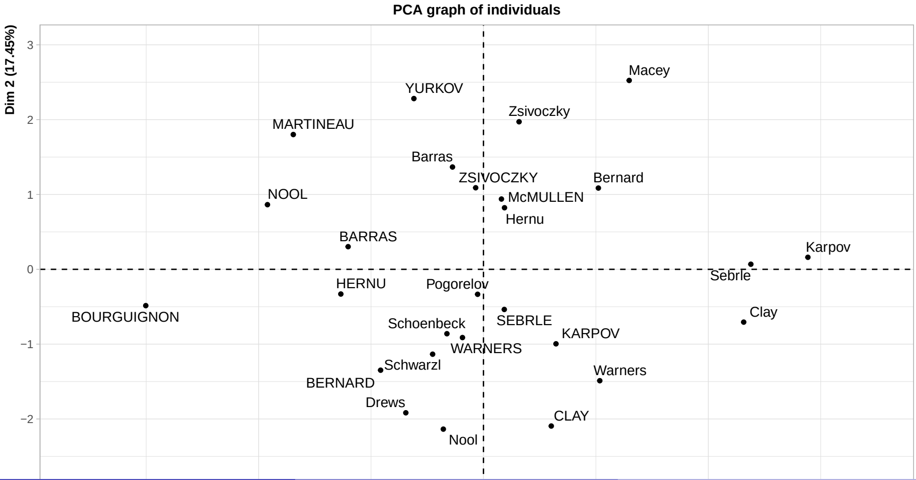
```
library(factoextra)
data("decathlon2")
head(decathlon2)
```

```
##           X100m Long.jump Shot.put High.jump X400m X110m.hurdle D
## SEBRLE    11.04      7.58    14.83      2.07 49.81        14.69
## CLAY      10.76      7.40    14.26      1.86 49.37        14.05
## BERNARD   11.02      7.23    14.25      1.92 48.93        14.99
## YURKOV    11.34      7.09    15.19      2.10 50.42        15.31
## ZSIVOCZKY 11.13      7.30    13.48      2.01 48.62        14.17
## McMULLEN  10.83      7.31    13.76      2.13 49.91        14.38
##           Pole.vault Javeline X1500m Rank Points Competition
## SEBRLE          5.02    63.19  291.7    1   8217    Decastar
## CLAY            4.92    60.15  301.5    2   8122    Decastar
## BERNARD         5.32    62.77  280.1    4   8067    Decastar
## YURKOV          4.72    63.44  276.4    5   8036    Decastar
```

Analyse en Composantes Principales

Application sur R

```r
library(FactoMineR)
decathlon2 <- decathlon2[,1:10] # variables quantitatives
pca <- PCA(decathlon2)
```



Analyse par Composantes Principales

Interprétation I

Les axes factoriels sont interprétés par rapport aux variables bien représentées en utilisant les contributions ou cercle des corrélations.

```
pca$var$contrib
```

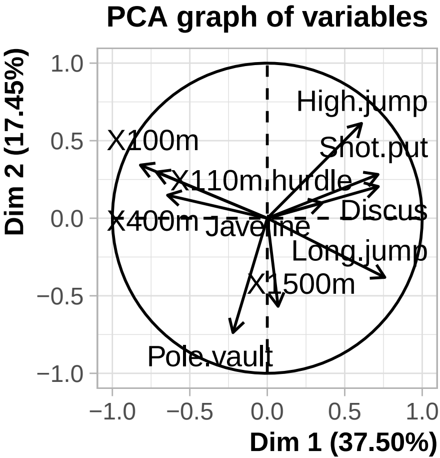
```
##                   Dim.1      Dim.2       Dim.3      Dim.4      Di
## X100m        17.8849964  6.7327182  0.670277238  0.9949816  7.819
## Long.jump    15.3581652  8.3394260  0.002582856  3.3309545 11.256
## Shot.put     13.6357518  4.5605826 14.793387670  0.1262838 12.566
## High.jump     9.8737797 21.4165036  0.001397716  0.4917303 14.623
## X400m        11.0544924  1.2622988 17.525552828  7.0513559  6.424
## X110m.hurdle 13.6869799  5.0732713 11.426291715  2.4733503  4.196
## Discus       13.7048617  2.3939535  4.814385760 15.3171947 18.654
## Pole.vault    1.3073595 31.1704550 10.704533859  6.1303166 11.160
## Javeline      3.3640178  0.5562982 31.863162259 22.8412641  2.881
## X1500m        0.1295955 18.4944926  8.198428099 41.2425682 10.415
```

Analyse par Composantes Principales

Interprétation I

Les axes factoriels sont interprétés par rapport aux variables bien représentées en utilisant les contributions ou cercle des corrélations.

```
plot(pca,choix = "var")
```

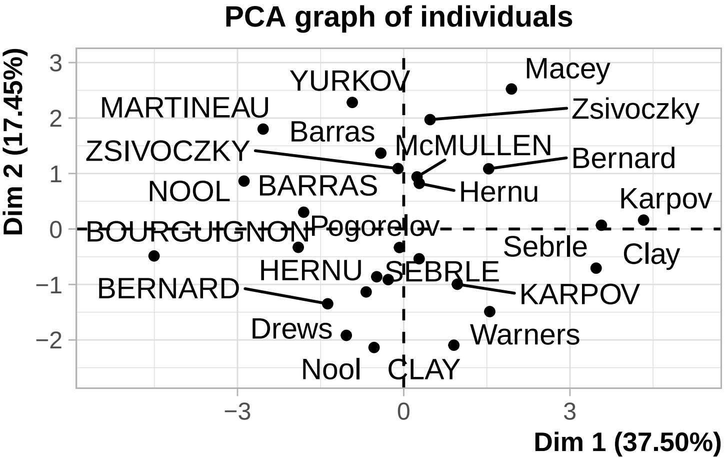


Analyse par Composantes Principales

Interprétation II

Les contributions des individus permettent d'identifier ceux qui ont une grande influence sur l'ACP. Ces individus sont à étudier parfois séparément.

```
plot(pca,choix = "ind")
```



Section 3

Analyse Factorielle des Correspondances Simples

(AFC)

Analyse Factorielle des Correspondances Simples

Qu'est-ce que c'est?

L'**Analyse Factorielle des Correspondances Simples** construit des facteurs (composantes principales) qui résument l'information contenue dans une table de contingence.

Objectifs :

- Visualisation des correspondances entre les modalités d'une même variable
- Représentation simultanée des modalités de 2 variables
- Analyse les liens entre les 2 variables

*Méthode descriptive pour l'analyse de variables **qualitatives**.*

Analyse Factorielle des Correspondances Simples

Application sur R

A titre d'exemple, nous utiliserons le jeu de données `housetasks`, qui contient des données relatives au partage de tâches ménagères au sein d'un couple:

```
data("housetasks")
head(housetasks)
```

```
##            Wife Alternating Husband Jointly
## Laundry     156          14       2       4
## Main_meal   124          20       5       4
## Dinner       77          11       7      13
## Breakfeast   82          36      15       7
## Tidying      53          11       1      57
## Dishes       32          24       4      53
```

Analyse Factorielle des Correspondances Simples

Principe

On dispose d'une table de contingence indiquant la répartition d'individus selon deux variables qualitatives. L'**Analyse Factorielle des Correspondances Simples** (AFC) consiste à :

- définir la **distance** entre 2 modalités par la distance du χ_2 entre profils lignes (ou colonnes).
- effectuer une **ACP** sur les profils lignes (ou colonnes), centrée sur la distribution marginale correspondante en remplaçant la distance classique par la distance du χ_2.

Analyse Factorielle des Correspondances Simples

Définitions

On note $X = (f_{i,j})$ le tableau de contingence à $1 \leq i \leq n$ modalités lignes et $1 \leq j \leq p$ modalités colonnes.

- Les **profils** lignes (resp. colonnes) sont définis comme les fréquences conditionnelles aux modalités des variables lignes (resp. colonnes).
- Le **profil moyen** est défini comme la distribution marginale en colonne (resp. ligne).
- La **distance** du χ_2 entre les modalités lignes i et i' (resp. j et j') est définie comme :

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}} \right),$$

$$d^2(y_j, y_{j'}) = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{i,j}}{f_{.j}} - \frac{f_{i,j'}}{f_{.j'}} \right).$$

Analyse Factorielle des Correspondances Simples

Profils lignes

```r
pl = rbind(housetasks, apply(housetasks, 2, sum))
rownames(pl)[14] = "Profil moyen"
round(100*prop.table(as.matrix(pl),margin=1),2)[c(1:5,14),]
```

```
##               Wife Alternating Husband Jointly
## Laundry      88.64        7.95    1.14    2.27
## Main_meal    81.05       13.07    3.27    2.61
## Dinner       71.30       10.19    6.48   12.04
## Breakfeast   58.57       25.71   10.71    5.00
## Tidying      43.44        9.02    0.82   46.72
## Profil moyen 34.40       14.56   21.85   29.19
```

Analyse Factorielle des Correspondances Simples

Profils colonnes

```
pc = cbind(housetasks, apply(housetasks, 1, sum))
colnames(pc)[5] = "Profil moyen"
head(round(100*prop.table(as.matrix(pc),margin=2),2))
```

```
##             Wife Alternating Husband Jointly Profil moyen
## Laundry    26.00        5.51    0.52    0.79        10.09
## Main_meal  20.67        7.87    1.31    0.79         8.77
## Dinner     12.83        4.33    1.84    2.55         6.19
## Breakfeast 13.67       14.17    3.94    1.38         8.03
## Tidying     8.83        4.33    0.26   11.20         7.00
## Dishes      5.33        9.45    1.05   10.41         6.48
```

Analyse Factorielle des Correspondances Simples

Principe II

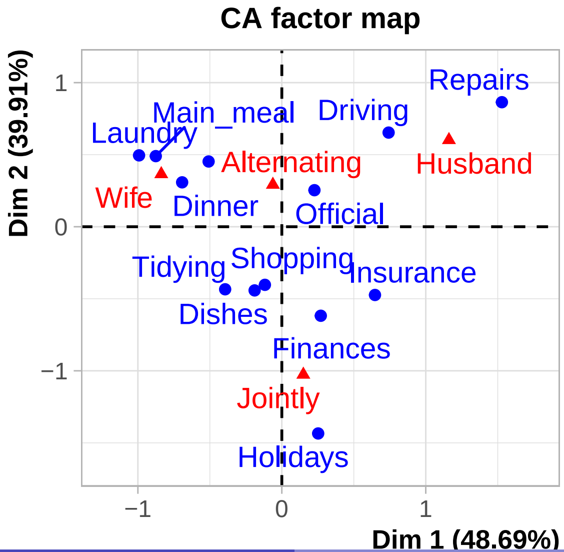
1. Transformation du tableau de contingence afin de récupérer les profils et colonnes.
2. ACP sur la tableau des profils lignes avec la distance du χ_2 :
   - maximisation sur chaque axe de la distance de chaque modalité de la variable ligne au profil ligne moyen
   - association de chaque modalité ligne i un point M_i, barycentre des p facteurs pondéré par les fréquences conditionnelles $(f_{i,j}/f_{\cdot j})_{1\leq j\leq p}$
3. Représentation des modalités lignes dans les plans factoriels centrés sur le profil moyen.
4. Choix des axes à l'aide des mêmes méthodes que pour l'ACP.

Remarque : si $p > n$, on travaille plutôt sur les profils colonnes.

Analyse Factorielle des Correspondances Simples

Application sur R

```
CA(as.table(as.matrix(housetasks)))
```



Analyse Factorielle des Correspondances Simples

Interprétation I

La contribution des modalités lignes aux axes factoriels est donnée par le code suivant :

```
res.ca <- CA(as.table(as.matrix(housetasks)),graph = FALSE)
res.ca$row$contrib
```

```
##                 Dim 1      Dim 2       Dim 3
## Laundry    18.2867003  5.5638913  7.96842443
## Main_meal  12.3888433  4.7355230  1.85868941
## Dinner      5.4713982  1.3210221  2.09692603
## Breakfeast  3.8249284  3.6986131  3.06939857
## Tidying     1.9983518  2.9656441  0.48873403
## Dishes      0.4261663  2.8441170  3.63429434
## Shopping    0.1755248  2.5151584  2.22335679
## Official    0.5207837  0.7956201 36.94038942
## Driving     8.0778371  7.6468564 18.59638635
## Finances    0.8750075  5.5585460  0.06175066
## Insurance   6.1470616  4.0203590  5.25263863
## Repairs    40.7300940 15.8806509 16.59639139
```

Analyse Factorielle des Correspondances Simples

Interprétation II

La qualité de représentation des modalités lignes sur les axes factoriels est d'autant bonne que son cosinus carré est proche de 1.

```
res.ca$row$cos2
```

```
##                 Dim 1      Dim 2       Dim 3
## Laundry    0.73998741 0.18455213 0.075460467
## Main_meal  0.74160285 0.23235928 0.026037873
## Dinner     0.77664011 0.15370323 0.069656660
## Breakfeast 0.50494329 0.40023001 0.094826699
## Tidying    0.43981243 0.53501508 0.025172490
## Dishes     0.11811778 0.64615253 0.235729693
## Shopping   0.06365362 0.74765514 0.188691242
## Official   0.05304464 0.06642648 0.880528877
## Driving    0.43201860 0.33522911 0.232752289
## Finances   0.16067678 0.83666958 0.002653634
## Insurance  0.57601197 0.30880208 0.115185951
## Repairs    0.70673575 0.22587147 0.067392778
## Holidays   0.02979239 0.96235977 0.007847841
```

Analyse Factorielle des Correspondances Simples

Interprétation III

Une modalité ligne (resp. colonne) n'intervient dans l'interprétation d'un axe factoriel que si :

- sa contribution à la construction de l'axe est $\geq 1/n$ (resp. $\geq 1/p$)
- **ET** elle est bien représentée sur l'axe, i.e. son cosinus carré est proche de 1.

Remarque : Si deux modalités bien représentées sur un plan factoriel sont proches, leurs distributions sont comparables. Les individus prenant ces modalités se comportent de manière comparable.

Analyse Factorielle des Correspondances Simples

Modalités atypiques

Les **modalités atypiques** ont de fortes contributions, éloignées du centre mais relativement mal représentées sur les axes factoriels. Elles sont dues à la distance du χ_2 qui a tendance à sur-représenter les modalités de faible effectif.

Que faire?

- les éliminer de l'analyse,
- les traiter comme modalités illustratives,
- les regrouper avec des modalités comparables,
- les ventiler sur les autres modalités en les attribuant de manière aléatoire une autre modalité aux individus concernés (*uniquement si ce sont des modalités de faible effectif*).

Section 4

Analyse Factorielle des Correspondances Multiples

(ACM)

Analyse Factorielle des Correspondances Multiples

Qu'est-ce que c'est?

L'**Analyse Factorielle des Correspondances Multiples** (ACM) construit des facteurs (composantes principales) qui résument l'information contenue dans plusieurs tableaux de contingence. Il s'agit donc d'une généralisation de l'Analyse Factorielle des Correspondances Simples.

Objectifs :

- Visualisation des correspondances entre les modalités d'une même variable
- Représentation simultanée des liens entre plusieurs variables
- Mettre en évidence des profils-types

Méthode descriptive pour l'analyse de $p > 2$ variables ***qualitatives***.

Analyse Factorielle des Correspondances Multiples

Application sur R

A titre d'exemple, nous utiliserons le jeu de données `poison`, qui contient des données recoltées sur 55 enfants d'une école primaire suite à une intoxication alimentaire :

```
library(FactoMineR)
data(poison)
head(poison)
```

```
##   Age Time   Sick Sex   Nausea Vomiting Abdominals   Fever   Diar
## 1   9   22 Sick_y   F Nausea_y  Vomit_n     Abdo_y Fever_y Diarrh
## 2   5    0 Sick_n   F Nausea_n  Vomit_n     Abdo_n Fever_n Diarrh
## 3   6   16 Sick_y   F Nausea_n  Vomit_y     Abdo_y Fever_y Diarrh
## 4   9    0 Sick_n   F Nausea_n  Vomit_n     Abdo_n Fever_n Diarrh
## 5   7   14 Sick_y   M Nausea_n  Vomit_y     Abdo_y Fever_y Diarrh
## 6  72    9 Sick_y   M Nausea_n  Vomit_n     Abdo_y Fever_y Diarrh
##     Fish   Mayo Courgette   Cheese   Icecream
## 1 Fish_y Mayo_y   Courg_y Cheese_y Icecream_y
## 2 Fish_y Mayo_y   Courg_y Cheese_n Icecream_y
## 3 Fish_y Mayo_y   Courg_y Cheese_y Icecream_y
```

Analyse Factorielle des Correspondances Multiples

Principe

On dispose d'un jeu de données de n individus et p variables qualitatives qui correspond à la donnée de plusieurs tableaux de contingence sur les mêmes individus.

L'**Analyse Factorielle des Correspondances Multiples** est basée sur l'étude du *tableau disjonctif complet* ou de la table de Burt associé :

```
##   Sick_y Sick_n F M Nausea_y Nausea_n Vomit_n Vomit_y Abdo_y Abdo
## 1      1      0 1 0        1        0       1       0      1
## 2      0      1 1 0        0        1       1       0      0
## 3      1      0 1 0        0        1       0       1      1
## 4      0      1 1 0        0        1       1       0      0
## 5      1      0 0 1        0        1       0       1      1
## 6      1      0 0 1        0        1       1       0      1
##   Fever_n Diarrhea_y Diarrhea_n Potato_y Potato_n Fish_y Fish_n M
## 1       0          1          0        1        0      1      0
## 2       1          0          1        1        0      1      0
## 3       0          1          0        1        0      1      0
## 4       1          0          1        1        0      1      0
## 5       0          1          0        1        0      1      0
```

Analyse Factorielle des Correspondances Multiples

Principe

On dispose d'un jeu de données de n individus et p variables qualitatives qui correspond à la donnée de plusieurs tableaux de contingence sur les mêmes individus.

L'**Analyse Factorielle des Correspondances Multiples** est basée sur l'étude du tableau disjonctif complet ou de la *table de Burt* associé :

```
##          Nausea_n Nausea_y Vomit_n Vomit_y Abdo_n Abdo_y Fever_n
## Nausea_n       43        0      28      15     18     25      19
## Nausea_y        0       12       5       7      0     12       1
## Vomit_n        28        5      33       0     17     16      18
## Vomit_y        15        7       0      22      1     21       2
## Abdo_n         18        0      17       1     18      0      17
## Abdo_y         25       12      16      21      0     37       3
##          Diarrhea_n Diarrhea_y Potato_n Potato_y Fish_n Fish_y Ma
## Nausea_n         20         23        1       42      1     42
## Nausea_y          0         12        2       10      0     12
## Vomit_n          17         16        3       30      1     32
## Vomit_y           3         19        0       22      0     22
## Abdo_n           17          1        0       18      0     18
```

Analyse Factorielle des Correspondances Multiples

Pré-traitement des données

L'ACM est sensible aux modalités de faible effectif qui perturbent l'analyse :

- nuages de points très concentrés et très eloignés des autres,
- petits effectifs ayant un grand poids dans l'analyse,
- instabilité des axes factoriels.

Pour rendre l'analyse plus robuste, on procède à un **apurement** : les modalités dont l'effectif est insuffisant ($< 2\%$ de l'effectif total) sont

- ventilées aléatoirement dans les autres modalités,
- gardées comme modalités illustratives.

Analyse Factorielle des Correspondances Multiples

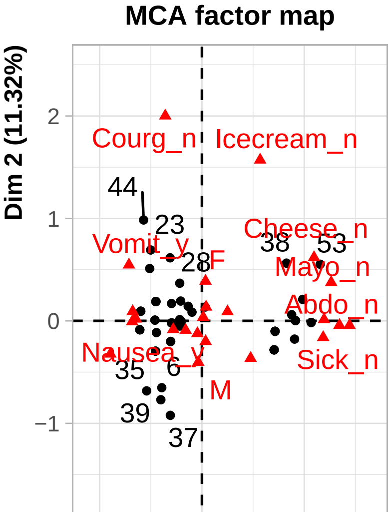
Remarques

- Nombre maximal de composantes correspondant au nombre total de modalités non-ventilées - nombre de variables.
- Choix du nombre de composantes à l'aide de la règle de Kaiser ou la règle du coude en traçant l'éboulis des valeurs propres.
- Décroissance des valeurs propres moins fortes que dans l'ACP ou l'AFC, ce qui implique un plus grand nombre d'axes à retenir.
- Représentation simultanée des n individus et p modalités (comme pour l'AFC).
- Axes factoriels expliqués à l'aide des contirbutions et cosinus carrés (comme pour l'AFC).

Analyse Factorielle des Correspondances Multiples

Application sur R

```r
res.mca <- MCA(poison,graph = FALSE)
plot(res.mca, autoLab = "yes")
```



Section 5

Positionnement Multidimensionnel

(MDS)

Positionnement Multidimensionnel

Qu'est-ce que c'est?

La méthode de **Positionnement Multidimensionnel** construit des axes permettant de représenter graphiquement les relations qui existent entre individus d'un même jeu de données.

Objectifs :

- Construction d'une carte spatiale des individus
- Visualisation des similarités entre individus
- Meilleure compréhension du jeu de données

*Méthode descriptive pour l'analyse multivariée de variables **quantitatives**.*

Positionnement Multidimensionnel

Principe

La méthode de positionnement multidimensionnel est basée sur une **matrice de proximité** $\Delta = (\delta_{i,j})_{1 \leq i,j \leq n}$, qui mesure les similarités/dissimilarités entre individus et a les propriétés suivantes :

- elle est positive, symétrique et à diagonale constante (similarité) ou nulle (dissimilarité),
- ses coefficients sont d'autant plus grand que les individus sont semblables ou différents.

Objectif : déduire de Δ des points dans l'espace.

Positionnement Multidimensionnel

Types d'algorithmes

- **MDS métrique** :
  - basée sur une matrice de distances Δ de type euclidien
  - configuration calculée directement à partir de Δ par des méthodes classiques d'algèbre linéaire
  - adaptée aux données quantitatives
- **MDS non-métrique**
  - basée sur une matrice de distances Δ de type non-euclidien
  - configuration estimée à partir de Δ de manière à conserver l'ordre de similarité (des + semblables aux - semblables)
  - adaptée aux données qualitatives

Positionnement Multidimensionnel

Application sur R - MDS métrique

A titre d'exemple, nous utiliserons le jeu de données `iris`, qui contient des données concernant 150 iris.

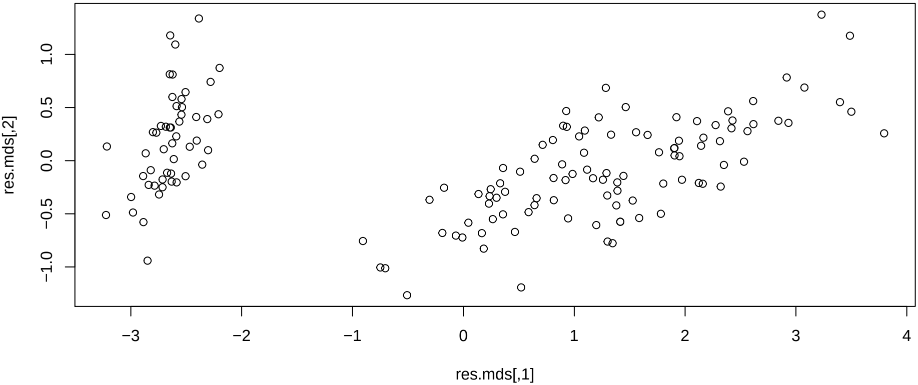
```
data(iris)
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ..
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ..
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1
```

Positionnement Multidimensionnel

Application sur R - MDS métrique

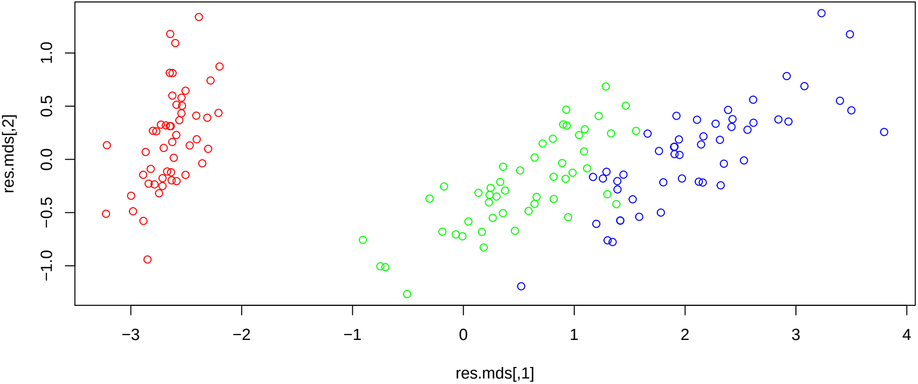
```r
d <- dist(iris[,-5])
res.mds <- cmdscale(d)
plot(res.mds)
```



Positionnement Multidimensionnel

Application sur R - MDS métrique

```
d <- dist(iris[,-5])
res.mds <- cmdscale(d)
plot(res.mds,col=rainbow(3)[iris$Species])
```



Positionnement Multidimensionnel

Application sur R - MDS non-métrique

A titre d'exemple, nous utiliserons le jeu de données swiss, qui contient des données relatives à 47 villes de Suisse :

```
library(MASS)
data(swiss)
str(swiss)
```

```
## 'data.frame':    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 5
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9
```

Positionnement Multidimensionnel

Application sur R - MDS non-métrique

```r
d <- dist(swiss)
res.mds <- isoMDS(d)
```

```
## initial  value 5.463800
## iter   5 value 4.499103
## iter   5 value 4.495335
## iter   5 value 4.492669
## final  value 4.492669
## converged
```

Positionnement Multidimensionnel

Application sur R - MDS non-métrique

```r
plot(res.mds$points)
text(res.mds$points, labels=rownames(swiss))
```