Analyse en Composantes Principales

N. Jégou

Université Rennes 2

Master 1 Géographie

Plan du cours

- Introduction
- Nuages N_p et N_n
- La méthode
- Interprétation

Bibliographie

- Ouvrages
  - Pagès J., Statistique générale pour utilisateurs : 1) Méthodologie, PUR (2010)
  - Pagès J., Analyse Factorielle multiple avec R EDP Sciences (2013)
  - Cornillon *et al.*, Statistique avec R PUR (2012)
- Vidéos - et Tutoriels R sur la page d'Agrocampus Ouest

  http://math.agrocampus-ouest.fr/infoglueDeliverLive/enseignement/support2cours/videos
  - Cours d'ACP

    https://www.youtube.com/watch?v=TAaAr9OM8rc&list=PLD5F63A877B376200
  - Utilisation de R

    https://www.youtube.com/watch?v=1QPRsg3Bxok

Motivations

L'Analyse en Composantes Principales (ACP) est la méthode de base en statistique exploratoire multidimensionnelle (ou analyse des données)

- Multidimensionnelle : l'analyse porte sur plusieurs variables
- Exploratoire : descriptive (par opposition à inférentielle)

Il s'agit de résumer l'information portant sur plusieurs variables en

- faisant émerger des liaisons entre variables
- formant des groupes d'individus se ressemblant

Les données en ACP

- En ACP les données se présentent dans un tableau X à n lignes et p colonnes où
  - chaque ligne représente un individu
  - chaque colonne représente une variable
- Les variables sont quantitatives : la matrice X est constituée de valeurs numériques

Les données en ACP

X est une matrice $n \times p$ de valeurs numériques :

$$X = \begin{bmatrix} x_{11} & . & . & . & . & . & x_{1p} \\ x_{21} & . & . & . & . & . & x_{2p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ x_{n1} & . & . & . & . & . & x_{np} \end{bmatrix}$$

Les données en ACP

Un individu est un élément de $\mathbb{R}^p$
Le $i^{\text{ème}}$ individu :

$$X = \begin{bmatrix} x_{11} & . & . & . & . & . & x_{1p} \\ x_{21} & . & . & . & . & . & x_{2p} \\ . & . & . & . & . & . & . \\ x_{i1} & . & . & . & x_{ij} & . & x_{ip} \\ . & . & . & . & . & . & . \\ x_{n1} & . & . & . & . & . & x_{np} \end{bmatrix}$$

Les données en ACP

Une variable est un élément de $\mathbb{R}^n$

La $j^{\text{ème}}$ variable :

$$X = \begin{bmatrix} x_{11} & . & . & . & x_{1j} & . & x_{1p} \\ x_{21} & . & . & . & . & . & x_{2p} \\ . & . & . & . & . & . & . \\ . & . & . & . & x_{ij} & . & . \\ . & . & . & . & . & . & . \\ x_{n1} & . & . & . & x_{nj} & . & x_{np} \end{bmatrix}$$

Données Températures

- On dispose des $p = 12$ températures mensuelles pour $n = 35$ villes Européennes
- Sont par ailleurs renseignées les variables
  - température moyenne annuelle
  - amplitude de température
  - latitude
  - longitude
  - région (qualitative à 4 modalités)

Données Températures

```
> don <- read.table("temperat.csv",sep=";",
+ dec=".",header=TRUE,row.names=1)

> dim(don)
[1] 35 17

> names(don)
[1] "Janvier" "Fevrier" "Mars" "Avril" "Mai" "Juin"
[7] "Juillet" "Aout" "Septembre" "Octobre" "Novembre"
[12] "Decembre"
[13] "Moyenne" "Amplitude" "Latitude" "Longitude" "Region"

> rownames(don)
[1] "Amsterdam" "Athenes" "Berlin" "Bruxelles"
[5] "Budapest" "Copenhague" "Dublin" "Helsinki"
[9] "Kiev" "Cracovie" "Lisbonne" "Londres"
[13] "Madrid" "Minsk" "Moscou" "Oslo"
[17] "Paris" "Prague" "Reykjavik" "Rome"
[21] "Sarajevo" "Sofia" "Stockholm" "Anvers"
[25] "Barcelone" "Bordeaux" "Edimbourg" "Francfort"
[29] "Geneve" "Genes" "Milan" "Palerme"
[33] "Seville" "St. Petersbourg" "Zurich"
```

Données Températures

- Nous ne considérons ici que les températures mensuelles ($p = 12$)
- Les individus sont les villes
- Un individu est décrit par ses $p = 12$ valeurs : c'est un élément de $\mathbb{R}^{12}$

- Les variables sont les températures mensuelles
- Une variable est décrite par ses valeurs sur les $n = 35$ individus
- Une variable est un élément de $\mathbb{R}^{35}$

Données centrées

- Moyennes par colonnes :

$$\begin{bmatrix} x_{11} & . & . & . & {\color{green}x_{1j}} & . & x_{1p} \\ x_{21} & . & . & . & {\color{green}.} & . & x_{2p} \\ . & . & . & . & {\color{green}.} & . & . \\ . & . & . & . & {\color{green}x_{ij}} & . & . \\ . & . & . & . & {\color{green}.} & . & . \\ x_{n1} & . & . & . & {\color{green}x_{nj}} & . & x_{np} \end{bmatrix}$$

$$\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow & {\color{green}\downarrow} & \downarrow & \downarrow \\ \bar{x}_1 & . & . & . & {\color{green}\bar{x}_j} & . & \bar{x}_p \end{matrix}$$

```
> apply(don[,1:12],FUN=mean,MARGIN=2)

Janvier    Fevrier    Mars       Avril     Mai       Juin
1.34571    2.21714    5.228571   9.28285   13.9114   17.414286

Juillet    Aout       Septembre  Octobre   Novembre  Decembre
19.622857  18.98000   15.631429  11.00285  6.065714  2.880000
```

Données centrées

- Centrage des données :

$$X = \begin{bmatrix} x_{11}-\bar{x}_1 & . & . & x_{1j}-\bar{x}_j & . & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & . & . & . & . & x_{2p}-\bar{x}_p \\ . & . & . & . & . & . \\ . & . & . & x_{ij}-\bar{x}_j & . & . \\ . & . & . & . & . & . \\ x_{n1}-\bar{x}_1 & . & . & x_{nj}-\bar{x}_j & . & x_{np}-\bar{x}_p \end{bmatrix}$$

- A Paris, la température en janvier est plus élevée que la moyenne, pas en août :

```
> don["Paris",1:12][c("Janvier","Aout")]-apply(don[,1:12],FUN=mean,MARGIN=2)[c("Janvier","Aout")]

         Janvier       Aout
Paris    2.354286      -0.28
```

Ecart-type

- On peut calculer l'écart-type pour chaque variable :

$$\begin{array}{ccccccc} \left[\begin{array}{ccccccc} x_{11} & . & . & . & x_{1j} & . & x_{1p} \\ x_{21} & . & . & . & . & . & x_{2p} \\ . & . & . & . & . & . & . \\ . & . & . & . & x_{ij} & . & . \\ . & . & . & . & . & . & . \\ x_{n1} & . & . & . & x_{nj} & . & x_{np} \end{array}\right] \\ \begin{array}{ccccccc} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \sigma_1 & . & . & . & \sigma_j & . & \sigma_p \end{array} \end{array}$$

- Il y a plus de variabilité de température en janvier qu'en mai :

```
> apply(don[,1:12],FUN=sd,MARGIN=2)[c("Janvier","Mai")]
Janvier      Mai
5.502157     3.273582
```

Données centrées-réduites

- Centrage puis réduction :

$$X = \begin{bmatrix} (x_{11}-\bar{x}_1)/\sigma_1 & . & . & (x_{1j}-\bar{x}_j)/\sigma_j & . & (x_{1p}-\bar{x}_p)/\sigma_p \\ (x_{21}-\bar{x}_1)/\sigma_1 & . & . & . & . & (x_{2p}-\bar{x}_p)/\sigma_p \\ . & . & . & . & . & . \\ . & . & . & (x_{ij}-\bar{x}_j)/\sigma_j & . & . \\ . & . & . & . & . & . \\ (x_{n1}-\bar{x}_1)/\sigma_1 & . & . & (x_{nj}-\bar{x}_j)/\sigma_j & . & (x_{np}-\bar{x}_p)/\sigma_p \end{bmatrix}$$

- A Reykjavik, la température en mai est beaucoup plus froide que la moyenne :

```
> scale(don[,1:12])["Reykjavik",c("Mai","Decembre")]
Mai                 Decembre
-2.2640122          -0.5395164
```

Objectifs

- Nous considérons X centrée-réduite (ACP normée)
- Le tableau X peut être analysé à travers ses lignes (les individus) ou à travers ses colonnes (les variables)
- $\Rightarrow$ résumer l'information en gardant à l'esprit cette dualité

Objectifs

- Nous considérons X centrée-réduite (ACP normée)
- Le tableau X peut être analysé à travers ses lignes (les individus) ou à travers ses colonnes (les variables)
- $\Rightarrow$ résumer l'information en gardant à l'esprit cette dualité
- Typologie des individus
  - Il existe une variabilité de températures entre les individus
  - $\Rightarrow$ former des groupes d'individus semblables
  - Termes clé : ressemblance

Objectifs

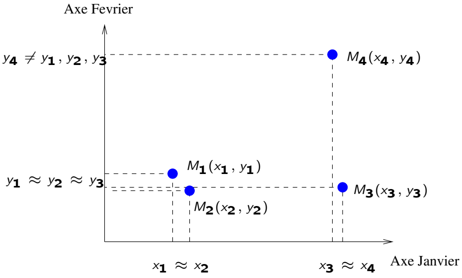
- Nous considérons X centrée-réduite (ACP normée)
- Le tableau X peut être analysé à travers ses lignes (les individus) ou à travers ses colonnes (les variables)
- $\Rightarrow$ résumer l'information en gardant à l'esprit cette dualité
- Typologie des individus
  - Il existe une variabilité de températures entre les individus
  - $\Rightarrow$ former des groupes d'individus semblables
  - Termes clé : ressemblance
- Typologie des variables
  - Il existe des variables liées entre elles
  - $\Rightarrow$ former des groupes de variables liées
  - Termes clé : liaison - corrélation

Objectifs

- Nous considérons X centrée-réduite (ACP normée)
- Le tableau X peut être analysé à travers ses lignes (les individus) ou à travers ses colonnes (les variables)
- $\Rightarrow$ résumer l'information en gardant à l'esprit cette dualité
- Typologie des individus
  - Il existe une variabilité de températures entre les individus
  - $\Rightarrow$ former des groupes d'individus semblables
  - Termes clé : ressemblance
- Typologie des variables
  - Il existe des variables liées entre elles
  - $\Rightarrow$ former des groupes de variables liées
  - Termes clé : liaison - corrélation
- Dualité : Quelles (groupes de) variables expliquent le plus la variabilité inter-individus ?

Nuage N_p des individus : n points de $\mathbb{R}^p$

- Un individu (ville - ligne) est un point de $\mathbb{R}^p$ (espace à p dimensions)
- Nuage N_p des individus : nuage de n points dans $\mathbb{R}^p$
- La "Ville" moyenne est le centre de gravité G du nuage
- Analogie avec la géométrie de $\mathbb{R}^2$, $\mathbb{R}^3$
  Chaque axe est associé à une variable :



Information

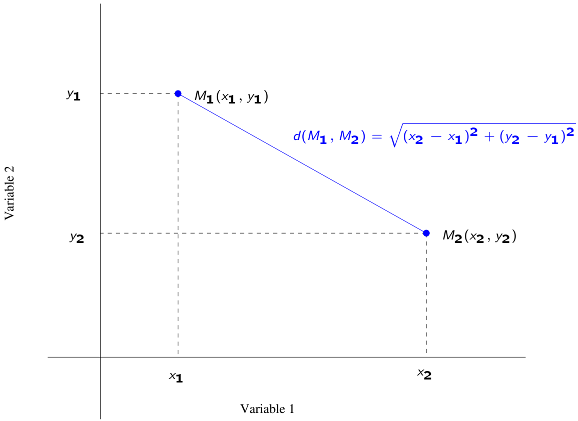
- Identification des groupes de points proches
- Identification de points isolés

$\Rightarrow$ dans quelles directions (i.e sur quelles variables) ?

- Identification de la forme du nuage
- Des directions d'allongements en particulier

$\Rightarrow$ concept clé : distances entre points

Rappel : Distance dans $\mathbb{R}^2$



Distance dans $\mathbb{R}^p$

- Analogie pour calculer la distance entre points de $\mathbb{R}^p$:

$$X = \begin{bmatrix} x_{11} & . & . & . & . & . & x_{1p} \\ x_{21} & . & . & . & . & . & x_{2p} \\ \color{red}{x_{i1}} & \color{red}{.} & \color{red}{.} & \color{red}{.} & \color{red}{x_{ij}} & \color{red}{.} & \color{red}{x_{ip}} \\ . & . & . & . & . & . & . \\ \color{red}{x_{l1}} & \color{red}{.} & \color{red}{.} & \color{red}{.} & \color{red}{x_{lj}} & \color{red}{.} & \color{red}{x_{lp}} \\ . & . & . & . & . & . & . \\ x_{n1} & . & . & . & . & . & x_{np} \end{bmatrix}$$

- Distance entre individu i et individu l :

$$d^2(i,l) = \sum_{j=1}^{p} (x_{ij} - x_{lj})^2$$

"Distance" entre villes

- Amsterdam est plus proche de Paris que d'Athènes en terme de profil de températures :

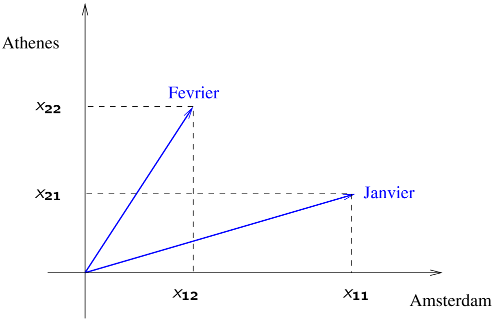
```
> sum((don["Amsterdam",1:12]-don["Paris",1:12])^2)

[1] 21.89

> sum((don["Amsterdam",1:12]-don["Athenes",1:12])^2)

[1] 786.72
```

- Une quantification de l'information sur l'ensemble des distances : la somme (des carrés) des distances au centre de gravité :

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)^2$$

Nuage N_n des variables : p vecteurs de $\mathbb{R}^n$

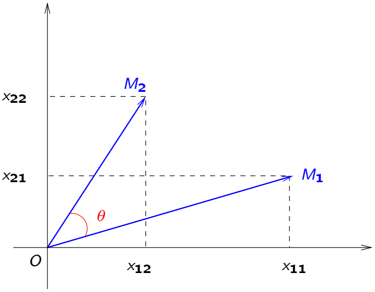
- Une variable (mois - colonne) est ici considérée comme un vecteur de $\mathbb{R}^n$
- Nuage N_n des variables : p vecteurs dans $\mathbb{R}^n$
- Chaque axe est associé à un individu (ville) :



Rappel : Produit scalaire

- La norme d'un vecteur correspond à sa longueur
- Le produit scalaire de deux vecteurs prend en compte longueurs et l'angle qu'ils forment

$$\langle \overrightarrow{OM}_1, \overrightarrow{OM}_2 \rangle = \|\overrightarrow{OM}_1\| \times \|\overrightarrow{OM}_2\| \cos(\theta) = x_{11}x_{12} + x_{21}x_{22}$$



$$\overrightarrow{OM}_1 = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$$

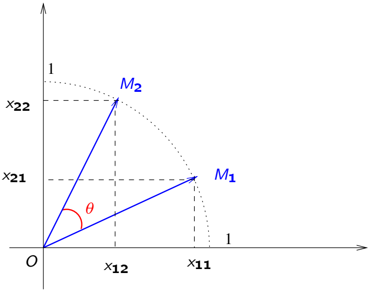
$$\overrightarrow{OM}_2 = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$$

Norme : $\|\overrightarrow{OM}_1\| = \sqrt{x_{11}^2 + x_{21}^2}$

Rappel : Produit scalaire

Pour des vecteurs de norme 1, la produit scalaire donne une mesure de l'angle (via le cos) :

$$\langle \overrightarrow{OM}_1, \overrightarrow{OM}_2 \rangle = \cos(\theta) = x_{11}x_{12} + x_{21}x_{22}$$



$$\overrightarrow{OM_1} = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$$

$$\overrightarrow{OM_2} = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$$

Norme : $\|\overrightarrow{OM_1}\| = \|\overrightarrow{OM_2}\| = 1$

Coefficient de corrélation

- Rappel (coefficient de) corrélation de 2 variables :

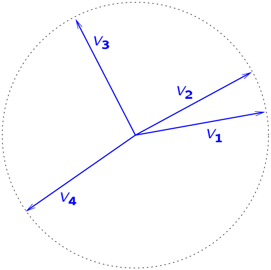
$$\mathbf{cor}(X_j, X_k) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij}-\bar{x}_j}{\sigma_j}\right)\left(\frac{x_{ik}-\bar{x}_k}{\sigma_k}\right)$$

- C'est le produit scalaire des deux colonnes centrées-réduites associées (à $1/n$ près) :

$$X = \begin{bmatrix} . & (x_{1k}-\bar{x}_k)/\sigma_k & . & \leftrightarrow & . & (x_{1j}-\bar{x}_j)/\sigma_j & . \\ . & . & . & \leftrightarrow & . & . & . \\ . & . & . & \leftrightarrow & . & . & . \\ . & (x_{ik}-\bar{x}_k)/\sigma_k & . & \leftrightarrow & . & (x_{ij}-\bar{x}_j)/\sigma_j & . \\ . & . & . & \leftrightarrow & . & . & . \\ . & (x_{nk}-\bar{x}_k)/\sigma_k & . & \leftrightarrow & . & (x_{nj}-\bar{x}_j)/\sigma_j & . \end{bmatrix}$$

Interprétation

- X centrée-réduite $\Rightarrow$ les colonnes ont même norme ($\equiv$ norme 1)
- Les p colonnes sont alors dans une (hyper)sphère (de rayon 1)
- L'angle formé par les vecteurs colonnes renseignent la corrélation sur les variables



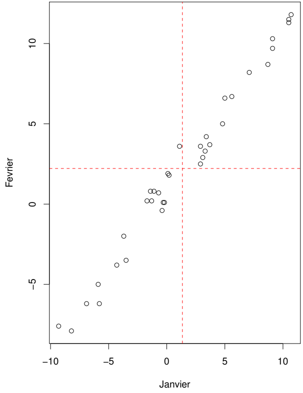
$\mathbf{cor}(V_1, V_2) \approx \mathbf{1}$

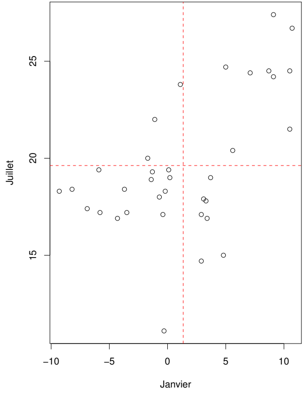
$\mathbf{cor}(V_1, V_4) \approx \mathbf{cor}(V_2, V_4) \approx -\mathbf{1}$

$\mathbf{cor}(V_1, V_3) \approx \mathbf{cor}(V_2, V_3) \approx \mathbf{cor}(V_4, V_3) \approx \mathbf{0}$

Interprétation

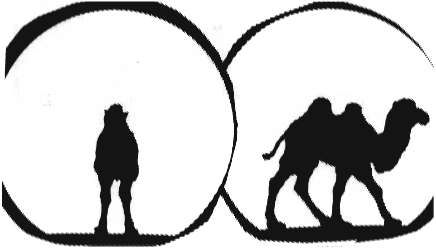
```
> cor(don[,1:12])["Janvier","Fevrier"]
[1] 0.9900015

> cor(don[,1:12])["Janvier","Juillet"]
[1] 0.5739173
```





Vers une représentation simplifiée

- Quelle est la meilleure projection ?



- La plus "grande" des deux

$\Rightarrow$ Séparer les points au maximum

Inertie

- L'inertie $\mathbf{I}$ des données est (à $1/n$ près) la somme des carrés des cellules de X centrée-réduite

$$\mathbf{I} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\frac{x_{ij}-\bar{x}_j}{\sigma_j}\right)^2$$

- C'est la somme (à $1/n$ près) des carrés des distances au centre de gravité pour tous les individus
- Quantification de l'information portée par les données

$\Rightarrow$ renseigne sur la "forme" du nuage des individus

Décomposition de l'inertie

- Idée : construction d'une suite de p axes permettant de restituer la forme du nuage
- Construction itérative
- On en déduit des représentations planes simples à interpréter
- Principe de réduction de la dimension

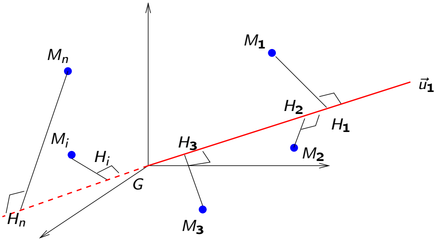
- Basé sur la décomposition de l'inertie

Décomposition de l'inertie

- 1^{er} axe : Axe principal de variabilité du nuage

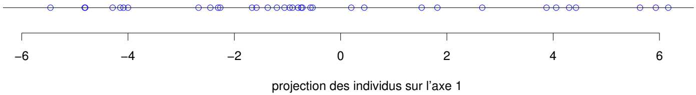
- Direction de $\mathbb{R}^p$ qui maximise l'inertie projetée :

  On cherche $\vec{u}_1$ telle que $\sum_{i=1}^{n} GH_i^2$ maximum



Décomposition de l'inertie

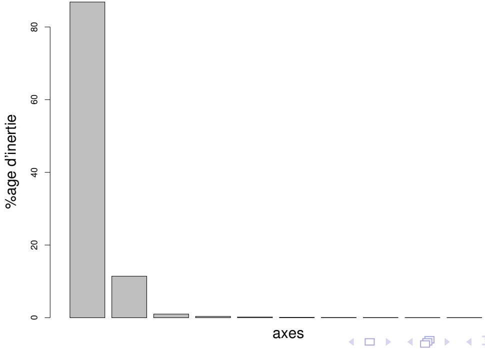
- Projection orthogonale des points sur l'axe 1 :



- On cherche ensuite un axe $\vec{u}_2$, orthogonal à $\vec{u}_1$, qui maximise l'inertie projetée
- C'est le second axe de variabilité du nuage
- Ce 2nd axe présente moins de variabilité que le précédent

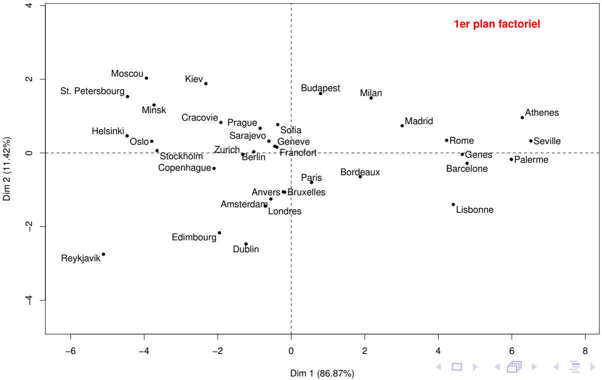
Décomposition de l'inertie

- On itère le procédé en cherchant $\vec{u}_3$ orthogonal au plan $\vec{u}_1, \vec{u}_2$ qui maximise l'inertie projetée
- ...
- Jusqu'à obtenir p axes orthogonaux
- La part d'inertie projetée sur chaque axe donne la part de variabilité restituée :



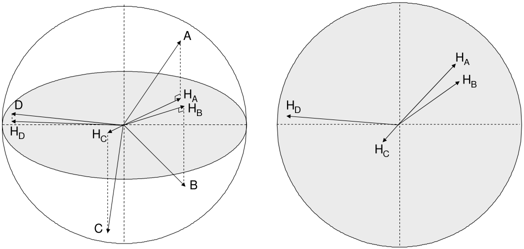
Plan factoriel

- On privilégie les représentations planes en projetant les individus sur les plans formés par les axes
- La projection orthogonale sur le plan formé par $\vec{u}_1$ et $\vec{u}_2$ est la meilleure représentation plane du nuage des individus
- Il concentre 98% de l'inertie



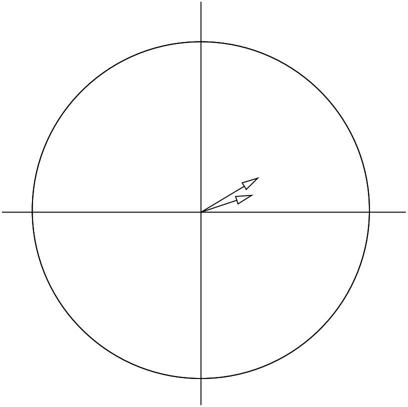
Cercle des corrélations

- Les axes factoriels sont
  - des combinaisons linéaires des colonnes de X
  - sont des vecteurs de $\mathbb{R}^n$
  - orthogonaux 2 à 2
- Les cercles de corrélations représentent les projections des colonnes de X sur les plans formés par ces axes



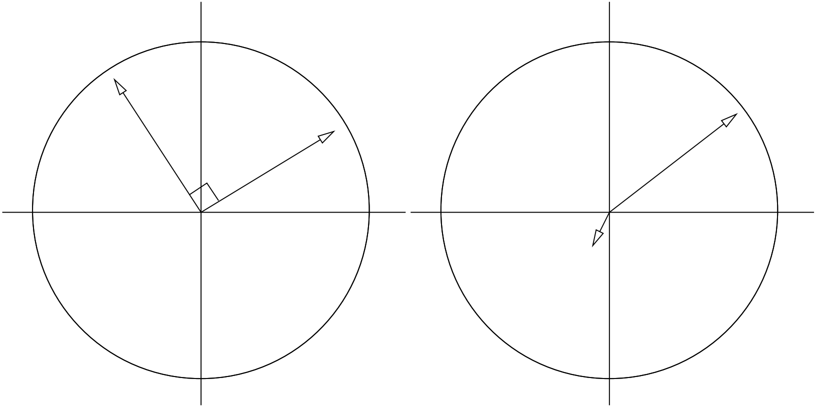
Aide à l'interprétation

Aucune interprétation



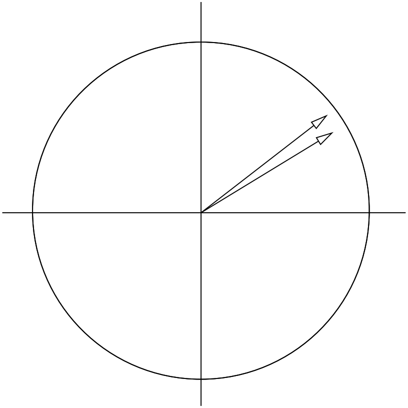
Aide à l'interprétation

Non corrélation

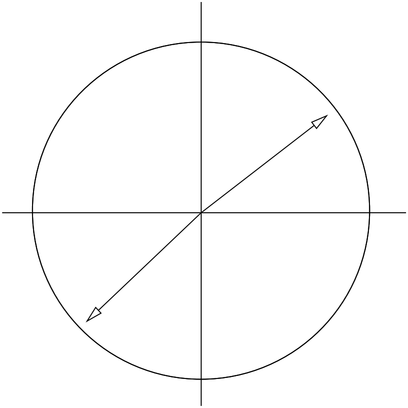


Aide à l'interprétation

Corrélation positive

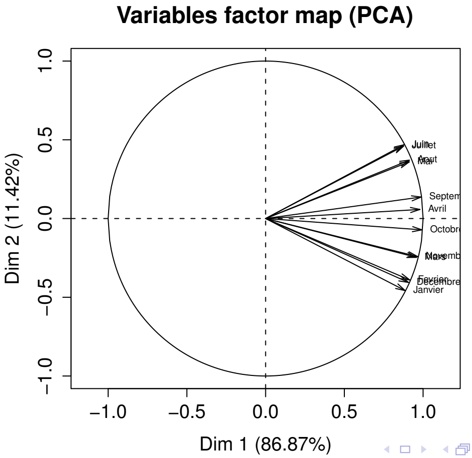


Corrélation négative



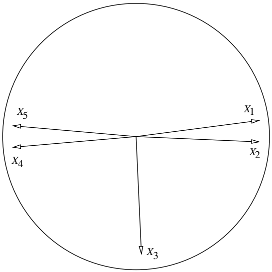
Exemple : effet taille

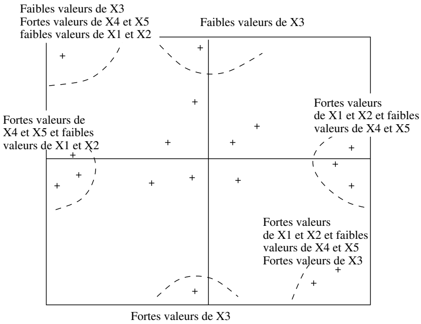
- Toutes les variables sont corrélées positivement : effet taille
- $\Rightarrow$ la plupart des villes sont ou chaudes ou froides toute l'année



Aide à l'interprétation

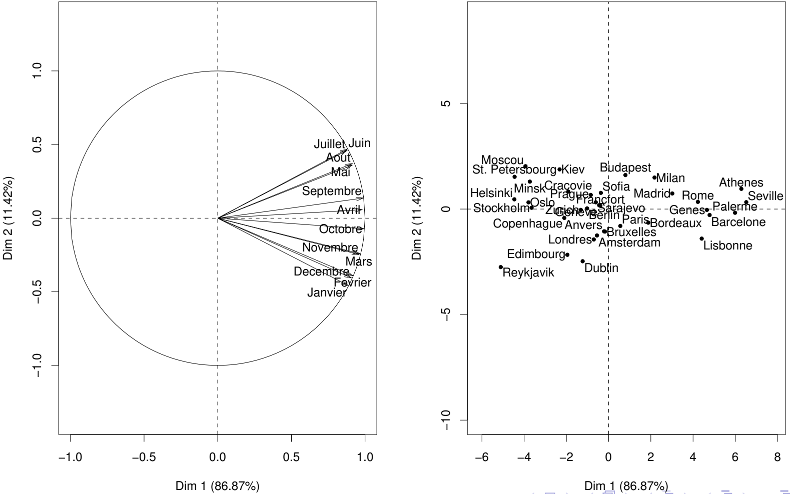
Variables $\rightarrow$ Individus





Package FactoMineR

```
> library(FactoMineR)
> res.pca <- PCA(don[,1:12])
```



Données températures

- Le premier plan principal explique la (quasi)totalité de l'information : 98.25%. Inutile d'analyser d'autres axes
- Typologie des variables
  - Effet taille
  - Axe 2 : opposition été/hiver
- Typologie des individus
  - Villes chaudes toute l'année : Seville, Athènes,...
  - Villes froides toute l'année : Helsinki, St-Petersbourg...
  - Villes très froides l'hiver : Moscou, Kiev,...
  - Villes particulièrement fraiches l'été : Reykjavic, Edimbourg...

Individus supplémentaires (illustratifs)

- Ils ne servent pas à calculer les axes
- Ils sont représentés (projetés) après

- Exemple : centre de gravité d'un groupe d'individus

```
> summary(don[,"Region"])
Est       Nord       Ouest       Sud
8         8          9           10
```

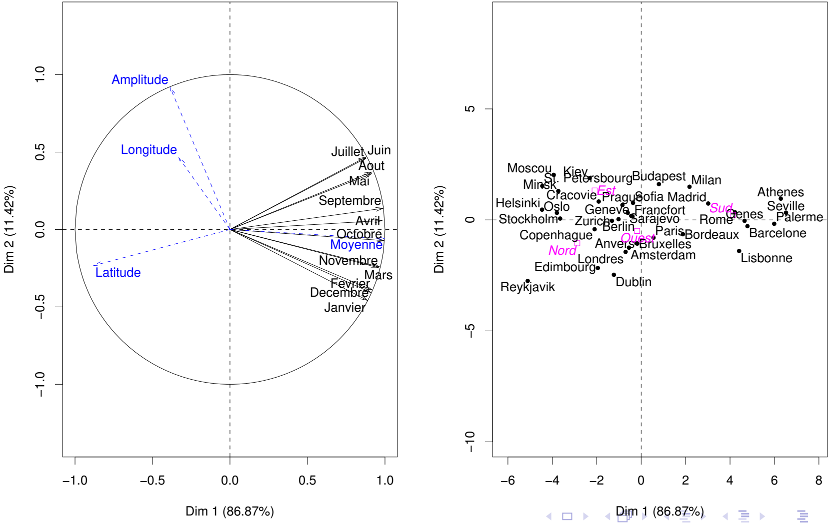
Variables supplémentaires (illustratives)

- Elles ne servent pas à calculer les axes
- Elles sont représentées (projetées) après sur les cercles

- Exemples
  - variables résultant des autres (moyennes...)
  - variables aidant à l'interprétation
  - en régression pour voir l'effet de variables explicatives sur une variable à expliquer

```
> colnames(don)[-c(1:12,17)]
[1] "Moyenne" "Amplitude" "Latitude" "Longitude"
```

Exemple températures



Ajouts aux interprétations

- Le premier axe est très corrélé à la température moyenne
- La latitude est très corrélée au le premier axe qui sépare les villes chaudes (au sud) des villes froides (à l'est)
- L'amplitude corrélée au second axe de variabilité qui résulte d'une oppsition été/hiver : séparation des villes de fortes amplitudes (Moscou, St Petersbourg,..), des villes aux faibles amplitudes (Reykjavic, Edimbourg, Dublin,...)