Analyse de données

L. BELLANGER

Plan

Plan ch. II

Brève introduction …

- **But de l'*ACP***
  - Etudier les ***liaisons entre plusieurs variables quantitatives*** et permettre une description synthétique du tableau, essentiellement sous forme de ***cartes*** de telle sorte que:
    - Les représentations graphiques (nuage de points et nuage de variables) soient optimales : déformation des distances par projection dans un sous-espace réduite.
- **Principe**
  - À partir de p variables initiales **continues**, construire k ($\leq p$) autres variables, appelées **composantes principales (*CP*)**, combinaisons linéaires des variables initiales, telles que :
    - les CP sont **ordonnées** selon l'information (variance) qu'elles restituent,
      - la 1ère étant celle qui restitue le plus d'information
    - la part d'information restituée par chaque CP est connue, et des critères permettent de décider combien de CP il est pertinent de conserver
    - les CP sont des vecteurs orthogonaux, i.e. des variables ***non corrélées*** entre elles
  - Origine : Karl Pearson (1901) – Harold Hotelling (1933)

Brève introduction …

- *Intérêt de l'ACP*
  - Présenter synthétiquement les données sous forme de cartes
  - Simplifier et schématiser les liaisons entre variables ; détecter des liaisons entre variables
  - Localiser les regroupements d'individus ou de variables
  - Détecter des individus exceptionnels ou aberrants, d'éventuels groupes isolés d'individus
  - Construire des variables synthétiques non-corrélées (régression sur CP, …)

5

1. Principe de l'*ACP*

- *Projection des individus dans un sous-espace*
  - ***Recherche des directions privilégiées*** *du nuage – selon des axes d'allongement maximum*

    le «chameau» cf. diapo suivante

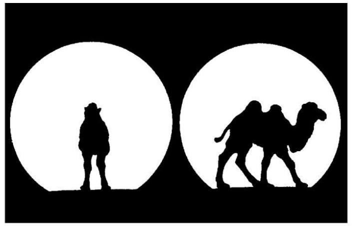
  - **Transformation des axes** (variables) originaux en un nouveau système d'axes factoriels orthogonaux de variance maximum et d'importance décroissante

- **Objectif final**
  - **Réduction** de la dimension et **Visualisation** dans des espaces à 2 ou 3 dimensions **pour mieux comprendre la structure des données**

6

1. Principe de l'*ACP* :
projeter la réalité sur un plan

Figure de J.P. Fenelon



- Nous avons l'habitude de dessiner ou photographier la réalité.
- Nous naturellement passons d'un espace à 3 dimensions à un espace à 2 dimensions.
- Selon le point de vue, l'information retenue ne sera pas la même.
- L'ACP nous propose un point de vue permettant de voir au mieux les individus d'un tableau.

7

1. Principe de l'*ACP* :
résumer les données

- Lorsqu'on projette les données sur un plan, on obtient un graphique déformé de la réalité.
- Le rôle de l'ACP est de trouver des espaces de dimensions plus petites minimisant ces déformations.
- On utilise un espace à 2 dimensions (un plan). Ce plan est appelé le plan principal. Il est constitué de deux droites perpendiculaires.
- La méthode consiste à calculer la première droite D1 de façon à **maximiser** les carrés des distances de projection des points sur la droite.

- Puis une 2éme droite D2 perpendiculaire à la première.

8

1. Principe de l'*ACP* : les composantes principales

- Les droites D1 et D2 sont des caractères synthétiques obtenus par des combinaisons linéaires avec les variables d'origines.
- Ces droites sont appelées **composantes principales**, ou **axes principaux**.
- La première composante principale doit "**capturer**" le maximum d'inertie du tableau des données. La variance des individus doit être maximale.
- Il reste un résidu non expliqué par cette première composante. C'est sur ce résidu qu'est calculée la deuxième composante principale.

9

1. Principe de l'*ACP* : propriétés des composantes principales

- La première composante principale "**capture**" le maximum d'inertie du tableau des données.
- La deuxième composante principale est un complément, une correction de la première.
- La deuxième composante principale doit avoir une corrélation linéaire nulle avec la première (orthogonalité).
- Il n'y a pas de redondance d'information entre deux composantes principales.
- On calcule les autres composantes de la même manière.

10

1. Principe de l'*ACP*

1ère présentation

- *Projection des individus dans un sous-espace*

Selon que l'on regarde les individus ou les variables dans $\mathbf{X}$, on peut se poser 2 questions complémentaires :

- les **individus** sont-ils **proches** ou **éloignés** ?
  - Regard sur les distances entre indiv.
  - Recherche d'une représentation qui les déforme le moins possible
- les **variables**, généralement **corrélées**, peuvent-elles être transformées pour donner d'autres variables aux propriétés plus intéressantes ?

11

1. Principe de l'*ACP*

1ère présentation

- *Projection des individus dans un sous-espace*

Principe de la méthode :

- Obtenir une représentation approchée du nuage des n individus dans un sous-espace de faible dimension $k(< p)$.
- Pour y parvenir, nous allons **projeter** les individus sur un sous-espace de dimension faible (k = dim de ce sous-espace) choisie suivant le critère suivant :
  - ***Les distances en projection devront être le moins déformées possibles (mais rétrécies !)***

Ce qui se traduit par :

- ***Le sous-espace F_k de dim k recherché est tq la moyenne des carrés des distances entre projections soit la plus grande possible.***

ou

- ***L'inertie du nuage projeté sur le ss-espace F_k doit être maximale.***

12

1. Principe de l'*ACP*

$1^{ère}$ présentation

- *Projection des individus dans un sous-espace*

Ces 2 critères sont justifiés puisque les distances ne peuvent que ↘ en projection orthogonale :

- f_1 et f_2 : projections resp. de x_1 et x_2 sur F_2 ;
- a_1 et a_2: coord. resp. de f_1 et f_2 sur le 1^{er} axe Δ_1;
- b_1 et b_2: coord. resp. de f_1 et f_2 sur le 2ième axe Δ_2.

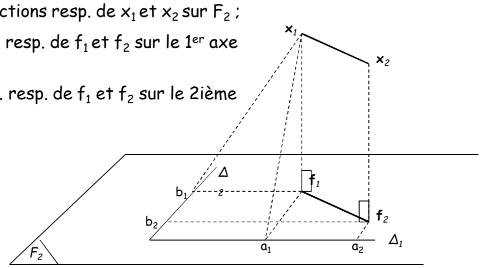


Fig. Projection du nuage des indiv. ds un sous-espace de dim 2

1. Principe de l'*ACP*

$1^{ère}$ présentation

- *Projection des individus dans un sous-espace*

2 résultats importants :

1. **L'inertie du nuage projeté** vaut : $\boxed{tr(\mathbf{SQP}) = tr(\mathbf{S_I P})}$
   où $\mathbf{P}$ est l'opérateur de projection $\mathbf{Q}$-orthogonal sur F_k et $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{DX}$
   Rappel: $\mathbf{Q}$ métrique dans l'espace des individus inclus dans $\mathbb{R}^p$,
   en général $\mathbf{Q} = \mathbf{I}_p$ ou $\mathbf{Q} = diag(\mathbf{S})^{-1}$
2. $I_{G \oplus G'} = I_G + I_{G'}$
   où G et G' sont 2 sous-espaces orthogonaux de l'espace des individus.

Il découle le **Théorème fondamental:**

Soit F_k un sous-espace portant l'inertie maximale ;
$\Longrightarrow$ le sous-espace de dimension $k+1$ portant l'inertie maximale est la **somme directe** de F_k et du sous-espace de dimension 1, $\mathbf{Q}$-orthogonal à F_k, portant l'inertie maximale.

Les solutions sont « emboîtées ».

<u>Dem</u> : ouvrages de ref AD, Lebart & Morineau, …

1. Principe de l'*ACP*

$1^{ère}$ présentation

- *Projection des individus dans un sous-espace*
  - Il suffit de **déterminer** le sous-espace vectoriel de dim 1 de $\mathbb{R}^p$ (i.e. **le vecteur directeur de la droite passant par le pt moyen** $\bar{x} \in \mathbb{R}^p$ si on suppose le nuage centré) qui maximise l'inertie du nuage projeté sur cette droite.
    - Soit $\boldsymbol{a} \in \mathbb{R}^p$ *le vecteur directeur* de cette droite Δ, on peut mq : $\mathbf{P}$ $(p \times p)$, le ***projecteur*** $\mathbf{Q}$***-orthogonal sur*** $\boldsymbol{\Delta}$ s'écrit:

$$\mathbf{P} = \boldsymbol{a}\left(\boldsymbol{a}^{\mathrm{T}}\mathbf{Q}\boldsymbol{a}\right)^{-1}\boldsymbol{a}^{\mathrm{T}}\mathbf{Q} = \frac{\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\mathbf{Q}}{\boldsymbol{a}^{\mathrm{T}}\mathbf{Q}\boldsymbol{a}} \text{ puisque } \boldsymbol{a}^{\mathrm{T}}\mathbf{Q}\boldsymbol{a} \in \mathbb{R}$$

    - ***L'inertie du nuage projetée*** sur cette droite est donc :

$$tr(\mathbf{S_I P}) = \frac{\boldsymbol{a}^{\mathrm{T}}\mathbf{QSQ}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\mathbf{Q}\boldsymbol{a}} = \lambda$$

    - Il faut donc rendre maximum cette quantité pour trouver $\boldsymbol{a}$, soit en dérivant l'expression précédente par rapport à $\boldsymbol{a}$:

$$\boxed{\mathbf{SQ}\boldsymbol{a} = \mathbf{S_I}\boldsymbol{a} = \lambda\boldsymbol{a}}$$

1. Principe de l'*ACP*

$1^{ère}$ présentation

- *Projection des individus dans un sous-espace*

En choisissant pour $\boldsymbol{a}$ le vecteur propre associé à la plus grande valeur propre de $\mathbf{S_I} = \mathbf{SQ}$, on obtient le sous-espace vectoriel de dimension 1 sur lequel l'inertie du nuage projeté est **maximale**.

Cette propriété se généralise à l'ordre k **:**

Le sous-espace F_k de dim k est engendré par les k vecteurs propres de $\mathbf{S_I}$ associés aux k plus grandes valeurs propres.

Rappel: Triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ des données

La forme du tableau de données $\mathbf{X}$ centré

un tableau de données $\mathbf{X}$ est une **représentation rectangulaire** à n lignes et p colonnes de la forme :

$$X = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ x_i^1 & & x_i^j & & x_i^p \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix}$$

individu i : vecteur ligne dans $\mathbb{R}^p$

variable j : vecteur colonne dans $\mathbb{R}^n$

2 nuages de points:

- nuage des n individus dans $\mathbb{R}^p$
- nuage des p variables dans $\mathbb{R}^n$

17

Rappel: Triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ des données

Un jeu de données est constitué par un triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ défini par les 3 éléments suivants:

1. $\mathbf{X} = \left[x_i^j\right] \in \mathcal{M}_{n\times p}$ matrice des données brutes n mesures de p variables, quantitatives ou non ;
2. $\mathbf{Q} \in \mathcal{M}_{p\times p}$, métrique Euclidienne sur l'espace $\mathbb{R}^p$ des lignes $\boldsymbol{x}_i \in \mathbb{R}^p$ de $\mathbf{X}$; $\mathbf{Q} = \mathbf{I}_p$ ou $\left(diag(\mathbf{S})\right)^{-1}$
3. $\mathbf{D} \in \mathcal{M}_{n\times n}$, métrique Euclidienne sur l'espace $\mathbb{R}^n$ des colonnes $\boldsymbol{x}^j \in \mathbb{R}^n$ de $\mathbf{X}$, qui sera **toujours diagonale**. $\mathbf{D} = \boldsymbol{diag}(p_1, \dots, p_n)$.

L'espace Euclidien $(\mathbb{R}^n, \mathbf{D})$ (resp. $(\mathbb{R}^p, \mathbf{Q})$) est l'espace des variables (resp. des individus) .

Notation

$r = rang(\mathbf{X}) \leq min(n, p)$

18

1. Principe de l'*ACP*

- ***Résumé des propriétés des élts principaux : schéma de dualité ACP***

$$(\mathbf{X},\mathbf{Q},\mathbf{D}) \Leftrightarrow \begin{tikzcd} {[p]} \arrow[r, "\mathbf{Q}"] & {[p]} \arrow[d, "\mathbf{X}"] \\ {[n]} \arrow[u, "\mathbf{X}^{\mathrm{T}}"] & {[n]} \arrow[l, "\mathbf{D}"] \end{tikzcd}$$

- ***Où***
  - $\mathbf{X}$ centrée,
  - Chaque ligne du tableau est assimilée à un vecteur de $\mathbb{R}^p$
  - Chaque colonne du tableau est assimilée à un vecteur de $\mathbb{R}^n$
  - $\mathbf{Q} = \mathbf{I}_p$ ou $\mathbf{Q} = \boldsymbol{diag}(\mathbf{S})^{-1}$, métrique de $\mathbb{R}^p$ et
  - $\mathbf{D} = \mathbf{I}_n/n$ ou $\mathbf{D} = \boldsymbol{diag}(p_i)$, matrice des poids des indiv., métrique de $\mathbb{R}^n$
- *L'ACP du triplet* $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ *: les tableaux obtenus à partir de la DVS de* $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$.

19

1. Principe de l'*ACP*

2ème présentation : *Généralisation de la DVS*

- La plupart des méthodes de l'AD peuvent être présentées dans un cadre commun : celui de l'extension du théorème de la DVS au cadre d'espaces Euclidiens plus généraux.
- Le choix d'une métrique permettra d'adapter cette technique générale appelée *ACP* du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ au pb posé par le type de données à traiter.
- **Historiquement**, l'*ACP* correspond au triplet :
  - $\mathbf{X} \in \mathcal{M}_{n\times p}$ matrice des variables centrées (éventuellement réduites)
  - $\mathbf{Q} = \mathbf{I}_p$ métrique usuelle sur l'espace des individus,
  - $\mathbf{D} = n^{-1}\mathbf{I}_n$ métrique sur l'espace des variables.

20

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

- *DVS du triplet* $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Rappel : DVS « maigre » dans le cas usuel:

Soit la matrice réelle $\mathbf{X} \in \mathcal{M}_{n \times p}$ de rg r alors il existe :

- Une matrice $\mathbf{U}_{p \times r} = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r]$ orthonormée $(\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathrm{I}_r)$
  dont les colonnes sont les vecteurs propres de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$
  $\Rightarrow$ *Vecteurs principaux de l'espace des lignes*
- Une matrice $\mathbf{V}_{n \times r} = [\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_r]$ orthonormée
  dont les colonnes sont les vecteurs propres de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$
  $\Rightarrow$ *Vecteurs principaux de l'espace des colonnes*

$\mathbf{U}$ et $\mathbf{V}$ matrices colonnes-orthonormales

- Une matrice diagonale $\boldsymbol{\Theta}_{r \times r} = \boldsymbol{\Lambda}^{1/2}$ contenant les r valeurs singulières de $\mathbf{X}$ tq $\mathbf{X}$ se décompose en :

$$\mathbf{X}_{n \times p} = \mathbf{V}_{n \times r}\boldsymbol{\Theta}_{r \times r}(\mathbf{U}_{p \times r})^{\mathrm{T}}$$

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

Dans la DVS, les matrices $\mathbf{X}^{\mathrm{T}}\mathbf{X}_{p \times p}$ et $\mathbf{X}\mathbf{X}^{\mathrm{T}}{}_{n \times n}$ symétriques jouent un rôle fondamental.

$\Rightarrow$ Dans la DVS de $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$, ce rôle va être attribué respectivement à :

$$\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{Q}_{p \times p} = \mathbf{S}\mathbf{Q} = \mathbf{S}_{\mathrm{I}} \text{ et } \mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}\mathbf{D}_{n \times n} = \mathbf{W}\mathbf{D}$$

Matrices non symétriques SAUF:

- ✓ dans le cas où $\mathbf{Q}$ et $\mathbf{D}$ sont de la forme $k\mathbf{I}$ (DVS usuelle) et
- ✓ dans le cas de l'*ACP* usuelle.

$\Rightarrow$ Elles sont alors resp. $\mathbf{Q}$ et $\mathbf{D}$-symétriques.

Le lemme suivant nous assure que les v.p. de telles matrices sont ≥ 0 et que les vecteurs propres sont orthogonaux au sens de la métrique concernée.

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

- *Lemme :* La matrice $\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{Q}$ (resp. $\mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}\mathbf{D}$) est $\mathbf{Q}$-sym. (resp. $\mathbf{D}$-sym), ses r valeurs propres > 0 et ses vecteurs propres forment une base $\mathbf{Q}$-orthonormée de Im $\mathbf{X}^{\mathrm{T}}$ (resp. $\mathbf{D}$-orthonormée de Im $\mathbf{X}$).

Dem: Durand pp. 102

Remarque: La construction effective des vecteurs propres $\{\boldsymbol{u}_i\}$ de $\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{Q}$ passe **d'abord** par:

- ✓ le calcul des vecteurs propres $\{\tilde{\boldsymbol{u}}_i\}$ de $\mathbf{Q}^{1/2}\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{Q}^{1/2}$; puis
- ✓ le calcul de $\boldsymbol{u}_i = \mathbf{Q}^{-1/2}\tilde{\boldsymbol{u}}_i$ car $\tilde{\boldsymbol{u}}_i$ est alors orthonormé au sens usuel. (cf. après relation avec la DVS usuelle)

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

- ***Théorème:*** *DVS de* $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$
  - Soient les matrices réelles $\mathbf{X}_{n \times p}$ de rang r et les métriques $\mathbf{Q}_{p \times p}$ et $\mathbf{D}_{n \times n}$ de $\mathcal{R}^p$ et de $\mathcal{R}^n$. Il existe :
    - Une matrice $\mathbf{U}_{p \times r} = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r]$ $\mathbf{Q}$-orthonormée
      les colonnes $\boldsymbol{u}_i$ sont les vecteurs propres de $\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X} = \mathbf{S}\mathbf{Q} = \mathbf{S}_{\mathrm{I}}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$.
      $\Rightarrow$ *Vecteurs principaux de l'espace des indiv.*
    - Une matrice $\mathbf{V}_{n \times r} = [\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_r]$ $\mathbf{D}$-orthonormée
      les colonnes st les vecteurs propres de $\mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}\mathbf{D} = \mathbf{W}\mathbf{D}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$
      $\Rightarrow$ *Vecteurs principaux de l'espace des variables.*
    - Une matrice diagonale $\boldsymbol{\Theta}_{r \times r} = \boldsymbol{\Lambda}^{1/2}$ contenant les r **valeurs singulières** du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ tq $\mathbf{X}$ se décompose en :

$$\mathbf{X}_{n \times p} = \mathbf{V}_{n \times r}\boldsymbol{\Theta}_{r \times r}(\mathbf{U}_{p \times r})^{\mathrm{T}} = \sum_{k=1}^{r} \sqrt{\lambda_k}\, \boldsymbol{v}_k (\boldsymbol{u}_k)^{\mathrm{T}}$$

1. Principe de l'*ACP*

2ème présentation : *Généralisation de la DVS*

Remarques DVS :

- $[\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r]$ vecteurs principaux de l'espace des individus ; $\|\boldsymbol{u}_k\|_{\mathbf{Q}} = 1$ et si $k \neq k'$ alors $\langle \boldsymbol{u}_k ; \boldsymbol{u}_{k'} \rangle_{\mathbf{Q}} = 0$
- $[\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_r]$ vecteurs principaux de l'espace des variables ; $\|\boldsymbol{v}_k\|_{\mathbf{D}} = 1$ et si $k \neq k'$ alors $\langle \boldsymbol{v}_k ; \boldsymbol{v}_{k'} \rangle_{\mathbf{D}} = 0$
- $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots \geq \sqrt{\lambda_r} > 0$ valeurs singulières
- Cette décomposition n'est pas unique : $\begin{matrix} \boldsymbol{u}_k & \rightarrow & -\boldsymbol{u}_k \\ \boldsymbol{v}_k & \rightarrow & -\boldsymbol{v}_k \end{matrix}$

Les représentations graphiques des projections des lignes ou des colonnes sont définies à une transformation orthogonale près du groupe engendré par les symétries hyperplanes dont les hyperplans sont engendrés par tous les axes principaux excepté un.

1. Principe de l'*ACP*

2ème présentation : *Généralisation de la DVS*

Formules de transition exprimant $\mathbf{U}$ en fonction de $\mathbf{V}$ (resp. $\mathbf{V}$ en fonction de $\mathbf{U}$)

$$\underset{p\times r}{\mathbf{U}} = \mathbf{X}^{\mathrm{T}} \underset{n\times n}{\mathbf{D}} \underset{n\times r}{\mathbf{V}} \underset{r\times r}{\mathbf{\Lambda}}^{-1/2} \quad \text{et} \quad \underset{n\times r}{\mathbf{V}} = \mathbf{X} \underset{p\times p}{\mathbf{Q}} \underset{p\times r}{\mathbf{U}} \underset{r\times r}{\mathbf{\Lambda}}^{-1/2}$$

Corollaire: Décomposition des matrices $\mathbf{S}$ *et* $\boldsymbol{W}$

Comme $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}$ et $\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}$, si $(\mathbf{U}, \mathbf{\Theta}, \mathbf{V})$ une DVS de $\mathbf{X}_{n\times p}$, on a :

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} = \sum_{k=1}^{r} \lambda_k \, \boldsymbol{u}_k (\boldsymbol{u}_k)^{\mathrm{T}}$$

et

$$\mathbf{W} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} = \sum_{k=1}^{r} \lambda_k \, \boldsymbol{v}_k (\boldsymbol{v}_k)^{\mathrm{T}}$$

1. Principe de l'*ACP*

2ème présentation : *Généralisation de la DVS*

Equations aux valeurs propres

Théorème

- Soit $(\mathbf{U}, \mathbf{\Theta}, \mathbf{V})$ une *DVS* de $\mathbf{X}_{n\times p}$ sous la forme $\mathbf{V}_{n\times r}\mathbf{\Theta}_{r\times r}(\mathbf{U}_{p\times r})^{\mathrm{T}}$ où l'espace des variables est muni de la métrique $\mathbf{D}$ et l'espace des individus de la métrique $\mathbf{Q}$ alors :

$$\underbrace{\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}}_{=\mathbf{S}} \, \mathbf{Q} \, \mathbf{U} = \mathbf{U} \underbrace{\mathbf{\Theta}^2_{p\times p}}_{=\mathbf{\Lambda}} \quad \text{et} \quad \underbrace{\mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}}_{=\mathbf{W}} \, \mathbf{D} \, \mathbf{V} = \mathbf{V} \underbrace{\mathbf{\Theta}^2_{n\times n}}_{=\mathbf{\Lambda}}$$

1. Principe de l'*ACP*

2ème présentation : *Généralisation de la DVS*

- ***Relation avec la DVS usuelle***

  - La DVS usuelle de $\mathbf{X}$ est la DVS du triplet $(\mathbf{X}, \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \mathbf{I}_n)$
  - DVS du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ ⇔ DVS $(\mathbf{Z} = \mathbf{D}^{1/2}\mathbf{X}\mathbf{Q}^{1/2}, \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = \mathbf{I}_n)$

au sens où elles ont les mêmes valeurs singulières θ.

$\Rightarrow$ Si $\mathbf{Z} = \mathbf{V}_{\mathbf{Z}}\mathbf{\Lambda}_{\mathrm{r}}^{1/2}(\mathbf{U}_{\mathbf{Z}})^{\mathrm{T}}$ et $\mathbf{X} = \mathbf{V}_{\mathbf{X}}\mathbf{\Lambda}_{\mathrm{r}}^{1/2}(\mathbf{U}_{\mathbf{X}})^{\mathrm{T}}$ sont les 2 décompositions,

ALORS

$$\mathbf{V}_{\mathbf{X}} = \mathbf{D}^{-1/2}\mathbf{V}_{\mathbf{Z}} \text{ et } \mathbf{U}_{\mathbf{X}} = \mathbf{Q}^{-1/2}\mathbf{U}_{\mathbf{Z}}$$

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

- ***Norme de Fröbenius*** (ou d'Hilbert-Schmidt) : rappel

Proposition : soient $\mathbf{D}$ et $\mathbf{Q}$ deux métriques des espaces $\mathcal{M}^n$ et $\mathcal{M}^p$ alors

$$\langle\, ; \rangle_{\mathbf{Q}\otimes\mathbf{D}} : \begin{cases} \mathcal{M}^{n\times p} \rightarrow \mathbb{R} \\ (\mathbf{X};\mathbf{Y}) \longmapsto tr\left(\mathbf{X}\mathbf{Q}\mathbf{Y}^{\mathrm{T}}\mathbf{D}\right) \end{cases}$$

est un produit scalaire sur l'espace vectoriel des matrices $\mathcal{M}^{n\times p}$. La norme induite est appelée norme de Fröbenius ou de Hilbert-Schmidt.

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

- *Rappel : Théorème d'approximation d'Eckart-Young*

– **Théorème d'approximation d'Eckart-Young**

Soient $X \in \mathcal{M}_{n\times p}$ de rang r, $X = V\Lambda^{1/2}U'$ la DVS de (X,Q,D) et k un entier $\leq r$. On note :

$U^{(k)} = [U_1 \dots U_k] \in \mathcal{M}_{p\times k}$ et $V^{(k)} = [V_1 \dots V_k] \in \mathcal{M}_{n\times k}$ les matrices extraites de U et V et $\left(\Lambda^{(k)}\right)^{1/2} = diag\left(\lambda_1^{1/2}, \lambda_2^{1/2}, \dots, \lambda_k^{1/2}\right)$ la diagonale des k premières valeurs propres singulières.

On recherche dans l'E.V. E_k des matrices $n \times p$, la matrice de rang $k \leq rg(X) = r$, la plus proche de X au sens de la norme $\|\|_{Q\otimes D}$:

$$\min_{E_k} \left\|X - X^{(k)}\right\|^2_{Q\otimes D} = \left\|X - Z^{(k)}\right\|^2_{Q\otimes D} = \sum_{i=k+1}^{r} \lambda_i$$

L'optimum est atteint par la DVS incomplète de rang k :

$$\boxed{Z^{(k)} = V^{(k)}\left(\Lambda^{(k)}\right)^{1/2}\left(U^{(k)}\right)'.}$$

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

- **Rappel : Approximation d'une matrice**

Soit $\mathbf{X}$ de rang r, on veut construire une matrice $\widehat{\mathbf{X}}^{(K)}$ de rang $k < r$ la plus proche de $\mathbf{X}$:

↪ il suffit de supprimer les valeurs singulières d'ordre $> k$ dans la DVS de $\mathbf{X}$ donnée par $\mathbf{X} = \mathbf{V}_{n\times r}\mathbf{\Theta}_{r\times r}\left(\mathbf{U}_{p\times r}\right)^{\mathrm{T}}$, on a alors :

$$\widehat{\mathbf{X}}^{(K)} = \sum_{i=1}^{k} \sqrt{\lambda_i}\, \boldsymbol{v}_i (\boldsymbol{u}_i)^{\mathrm{T}}$$

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

- *ACP du triplet (X,Q,D)*

Usuellement : *ACP* = étude du triplet $\left(\mathbf{X}_C, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n\right)$.

Le fait d'envisager des métriques plus générales introduit une distorsion dans la représentation des distances (AFC distance du Chi-deux).

⇒ Les **plans factoriels** de projection sont ceux formés par les couples de vecteurs des bases orthonormée de la DVS du triplet :

- $\left(\boldsymbol{u}_i, \boldsymbol{u}_j\right)$ *pour voir les points lignes (obs.)* et
- $\left(\boldsymbol{v}_i, \boldsymbol{v}_j\right)$ *pour voir les points colonnes (variables).*

=> Quels sont les **meilleurs k plans factoriels** i.e. ceux pour qui les photos sont porteuses d'information ?

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

Définition :

- $X = \left(I_n - \mathbf{1}_n \mathbf{1}_n' D\right) T$ matrice D **centrée** en colonne
- $S = X'DX$ matrice des covariances entre variables
- $W = XQX'$ matrice des produits scalaires entre indiv

A ces matrices sont associées les opérateurs aux valeurs propres-vecteurs propres de la DVS du triplet
Opérateurs en dualité et inertie du triplet (X, Q, D)

- Opératuer des covariances : $SQ = X'DXQ$
- Opérateur des prod. scal. entre ind. : $WD = XQX'D$
- Inertie totale du triplet : $\|X\|^2_{Q\otimes D} = tr\,(XQX'D) = tr\,(X'DXQ)$

<u>**Théorème**</u> : Soit $\mathbf{X}_{n\times p} = \mathbf{V}_{n\times r}\boldsymbol{\Theta}_{r\times r}\left(\mathbf{U}_{p\times r}\right)^{\mathrm{T}}$ une DVS de $\mathbf{X} = \mathbf{X}_C$

ALORS $\|\mathbf{X}\|^2_{\mathbf{Q}\otimes\mathbf{D}}$ correspond à I_g i.e. l'inertie de la matrice $\mathbf{X}$ s'écrit :

$$\boxed{\begin{aligned} I_g &= \|\mathbf{X}\|^2_{\mathbf{Q}\otimes\mathbf{D}} = \langle \mathbf{X}; \mathbf{X}\rangle_{\mathbf{Q}\otimes\mathbf{D}} = tr\left(\mathbf{XQX}^{\mathrm{T}}\mathbf{D}\right) \\ &= tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{DXQ}\right) = tr(\mathbf{S}_I) = \Sigma_{k=1}^{r}\lambda_k \end{aligned}}$$

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

Principe fondamentale de l'analyse factorielle

Si on admet que le meilleur « cliché » uni-dimensionnel est fourni par un axe sur lequel, en projection le nuage des points lignes est **d'inertie maximale**

ALORS

l'axe factoriel $\boldsymbol{u}_1$ est le meilleur axe ensuite $\boldsymbol{u}_2$ est le meilleur second orthog. Au 1er ....

Sous R, 1ere ou 2eme présentation : ex

- <u>1ere présentation</u> (décomposition spectrale) : les fonctions **PC** (avec option « eigen ») et **princomp**
- <u>2eme présentation</u> (DVS) : les fonctions **PC** (avec option svd), **prcomp**, **PCA**, **pca**

1. Principe de l'ACP

2ème présentation : *Généralisation de la DVS*

Définition : ***ACP d'ordre k du triplet*** $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

La matrice k étant supposé de rang r et D-centrée. On appelle ACP d'ordre k, $k \le r$ du triplet (X, Q, D) la DVS incomplète de rang k

$$\widehat{X}^{(k)} = V^{(k)}\left(\Lambda^{(k)}\right)^{1/2}\left(U^{(k)}\right)'$$

tq def. dans le théo d'Eckart-Young.

Les 2 formules de transition s'écrivent à l'ordre k:

$$V^{(k)} = XQU^{(k)}\left(\Lambda^{(k)}\right)^{-1/2} \text{ et } U^{(k)} = X'DV^{(k)}\left(\Lambda^{(k)}\right)^{-1/2}$$

2. Eléments principaux de l'ACP

- **Axes principaux** $\boldsymbol{a}_k\,; k = 1, \dots, r$

On appelle ***axes principaux d'inertie*** les *vecteurs propres* $\boldsymbol{a}_k$ de $\mathbf{SQ} = \mathbf{S}_I$, $\mathbf{Q}$-normés à 1 ; ils sont au nombre de $r \le p$ si $p < n$.

Propriété : L'inertie expliquée par un axe $\boldsymbol{a}_k$ est égale à la valeur propre λ_k associée à cet axe principal

- **Facteurs principaux** $\boldsymbol{u}_k;\ k = 1, \dots, r$

Le facteur principal $\boldsymbol{u}_k$ associé à l'axe principal $\boldsymbol{a}_k$ est la fonction linéaire sur $\Re^p$ définie par la relation : $\boldsymbol{u}_k = \mathbf{Q}\boldsymbol{a}_k$.

<u>En anglais :</u> les valeurs des composantes des vecteurs $\boldsymbol{u}$ sont les *loadings*.

Propriétés:

1. $\mathbf{QS}\boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k$
2. Les $\boldsymbol{u}_k$ sont les vecteurs propres $\mathbf{Q}^{-1}$-normés de $\mathbf{QS}$; on a :
$$\forall k = 1, \dots, r,;\ \boldsymbol{u}_k{}^{\mathrm{T}}\mathbf{Q}^{-1}\boldsymbol{u}_k = 1$$

2. Eléments principaux de l'*ACP*

- ***Composantes principales (CP)*** $F^k \in \mathbb{R}^n; k = 1, \ldots, r \leq min(n,p)$

ie Coord. des individus

Définition : Les CP sont les variables $\boldsymbol{F}^k$ elts de $\mathbb{R}^n$ définies par les **facteurs principaux** par la relation :

$$\boldsymbol{F}^k = \mathbf{X}\boldsymbol{u}_k$$

$\boldsymbol{F}^k$ contient les coordonnées des projections $\mathbf{D}$-orthogonales des observations sur l'axe de rang k défini par $\boldsymbol{a}_k$.

<u>En anglais :</u> les coordonnées des observations sur les composantes principales sont les *scores*.

Propriétés :

1. $var(\boldsymbol{F}^k) = \lambda_k = \|\boldsymbol{F}^k\|_D^2$

Les $\boldsymbol{F}^k$ sont les combinaisons linéaires de $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p$ de variance maximale sous la contrainte : $\boldsymbol{u}_k^{\mathrm{T}}\mathbf{Q}^{-1}\boldsymbol{u}_k = 1$.

2. On a : $\mathbf{WD}\,\mathbf{F}^k = \lambda_k \boldsymbol{F}^k$ où $\mathbf{W} = \mathbf{XQX}^{\mathrm{T}}$ est la matrice de t.g. le produit scalaire entre individus.

Les composantes principales $\boldsymbol{F}^k$ sont $\mathbf{D}$-orthog. (i.e. sont non corrélées entre elles).

Remarque : $\boldsymbol{F}^k / \sqrt{\lambda_k}$ est appelée *CP* normée

2. Eléments principaux de l'*ACP*

- **Coord. des variables** $\boldsymbol{G}^k \in \mathbb{R}^p \,; k = 1, \ldots, r \leq min(n,p)$
  - contient les coord. des projections $\mathbf{Q}$-orthogonales des obs. sur l'axe k
  - $\boldsymbol{G}^k = \sqrt{\lambda_k}\boldsymbol{a}_k$; $\|\boldsymbol{G}^k\|_Q^2 = \lambda_k$ et $cor(\boldsymbol{F}^k, \boldsymbol{x}^j) = \frac{G^k(j)}{\|x^j\|_D}$
- **Relations de transition** (sous entendu d'un espace à l'autre)

  $\boldsymbol{F}^k(i) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^p x_i^j \boldsymbol{G}^k(j)$ *coordonnée de l'individu i sur l'axe de rang k;*

  $\boldsymbol{G}^k(j) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^n x_i^j \boldsymbol{F}^k(i)$ *coordonnée de la variable j sur l'axe de rang k*

- **Formule de reconstitution**

  A partir des facteurs principaux et des composantes principales, il est possible de reconstituer le tableau $\mathbf{X}$ initial (cf. DVS avant) :

$$\mathbf{X} = \sum_{k=1}^{r} \boldsymbol{F}^k (\boldsymbol{u}_k)^{\mathrm{T}} \, \mathbf{Q}^{-1} = \sum_{k=1}^{r} \sqrt{\lambda_k} \boldsymbol{v}_k (\boldsymbol{u}_k)^{\mathrm{T}}$$

2. Eléments principaux de l'*ACP*

- ***Résumé des propriétés des éléments principaux***

Pour $k = 1, \ldots, r$, on a :

Elts principaux	Définition	Propriété	Relation
$\boldsymbol{a}_k$: axes principaux $\in \mathbb{R}^p$	$\mathbf{SQ}\boldsymbol{a}_k = \lambda_k \boldsymbol{a}_k$	$\mathbf{Q}$-orthonormés	
$\boldsymbol{u}_k$: facteurs principaux $\in \mathbb{R}^{p*}$	$\mathbf{QS}\boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k$	$\mathbf{Q}^{-1}$-orthonormés	$\boldsymbol{u}_k = \mathbf{Q}\boldsymbol{a}_k$
$\boldsymbol{F}^k$: *CP* $\in \mathbb{R}^n$ *coord des ind sur l'axe k*	$\mathbf{WD}\,\boldsymbol{F}^k = \lambda_k \boldsymbol{F}^k$	$\mathbf{D}$-orthogonales $\mathbf{D}$-normées à λ_k	$\boldsymbol{F}^k = \boldsymbol{X}\boldsymbol{u}_k$
$\boldsymbol{G}^k \in \mathbb{R}^p$ *Coord des var sur l'axe k $\in \mathbb{R}^p$*		$\mathbf{Q}$-orthogonales $\mathbf{Q}$-normées à λ_k	$\boldsymbol{G}^k = \sqrt{\lambda_k}\boldsymbol{a}_k$ **et** $\boldsymbol{G}^k(j) = \|\boldsymbol{x}^j\|_{\mathbf{D}}\, cor(\boldsymbol{F}^k; \boldsymbol{x}^j)$

Voir résumé 2 pages

2. Eléments principaux de l'*ACP*

- ***Remarques***
  - Le sens d'un axe est arbitraire, les résultats se retrouvent par symétrie
  - En présence de valeurs propres multiples, la *DVS* n'est plus unique. En pratique, ce cas n'arrive jamais !
- ***Exemples complets : à faire !***

  ACP de $(\mathbf{X}_C, \mathbf{I}_3, (1/4)\mathbf{I}_4)$ **feuille Td-TP ch 2** page 5

où

$$\mathbf{X} = \begin{bmatrix} -2 & 3 & -1 \\ -1 & 1 & 0 \\ 2 & -1 & -1 \\ 1 & -3 & 2 \end{bmatrix}$$

et ACP feuille Td-TP page 8

2. Eléments principaux de l'ACP

- **L'ACP Normée : Travailler avec**

$\mathbf{Q} = diag(\mathbf{S})^{-1}$ sur le tableau centré $\mathbf{X}_C$

⇔

$\mathbf{Q} = \mathbf{I}_p$ sur le tableau centré réduit $\mathbf{X}_{CR}$

C'est l'ACP sur les **données centrées réduites**, appelée *ACPN*.

Dans ce cas, $\mathbf{S}_l = \mathbf{SQ} = \mathbf{R}$, c'est la matrice des corrélations entre variables.

La diagonalisation de $\mathbf{R}$ fournira **facteurs principaux** $\boldsymbol{u} = \boldsymbol{a}$ et CP $\boldsymbol{F} = \mathbf{X}_{CR}\boldsymbol{a}$.

Propriété: $\boldsymbol{F}^1$, la 1ère CP est celle qui possède ***la variance maximale*** λ_1. Elle se définit par $\boldsymbol{F}^1 = \mathbf{X}_{CR}\boldsymbol{u}_1$ et est tq :

$\sum_{j=1}^{p} r^2(\boldsymbol{F}^1; \boldsymbol{x}^j)$ est maximale.

2. Eléments principaux de l'ACP

- **Remarque :** Transformation des données
  - Exemple : âge de l'exploitant, revenu et surface relevés sur 3 exploitations agricoles.
  - Rappel : recherche des axes -> maximisation de l'inertie ou conservation des distances entre points.

	Age	Revenu	Surface
1	30	290 000	20
2	50	300 000	30
3	52	320 000	28

Exploitant jeune + faible surface

Se ressemblent : structure + âge de l'exploitant

2. Eléments principaux de l'ACP

- **Remarque :** Transformation des données

Processus de standardisation des données

- *Centrage* : on retire la moyenne
  $x_{ik} - \bar{x}_k$
- *Réduction* : on divise par l'écart-type
  $(x_{ik} - \bar{x}_k)/S_k$

- Les variables sont par défaut centrées et généralement, elles sont aussi réduites (choix possible dans les logiciels).

⚠ Si le centrage est neutre pour l'analyse, la réduction ne l'est pas !
  - **Avantage** = permettre une comparaison entre variables mesurées dans des unités très différentes.
  - **Inconvénient** = donne une importance identique à chaque variable.

- Selon la problématique de l'analyse, ce peut être un bien ou un mal. En effet, si toutes les variables sont mesurées dans la même unité, il peut être préférable de conserver leurs variances respectives. On parle alors d'ACP non normée.

2. Eléments principaux de l'ACP

- **Remarque :** Transformation des données
  - Distance utilisée : distance euclidienne

d	1	2	3
1	0	10 000	30 000
2	10 000	0	20 000
3	30 000	20000	0

Exemple : $d(1;2) = \sqrt{(50-30)^2 + (290000-300000)^2 + (20-30)^2}$

- **Exploitation 2 apparaît plus proche de 1 que de 3** : importance écrasante prise par la variable revenu -> Une variation de 10% du revenu n'a pas la même incidence qu'une variation équivalente des deux autres variables.

2. Eléments principaux de l'*ACP*

- **Remarque :** Transformation des données

  - Changement d'échelle : revenu non plus en francs mais en dizaine de milliers de francs.

	Age	Revenu	Surface
1	30	29	20
2	50	30	30
3	52	32	28

→

d	1	2	3
1	0	22.4	23.6
2	22.4	0	3.46
3	23.6	3.46	0

Exemple 22.4 entre exp.1 et exp. 2 : $d = \sqrt{(30-50)^2 + (29-30)^2 + (20-30)^2}$

- Désormais 2 apparaît plus proche de 3 que de 1.
- Double inconvénient :
  - Hétérogénéité des variables.
  - Choix arbitraire de l'échelle de mesure.

2. Eléments principaux de l'*ACP*

- **Remarque :** Transformation des données

  - Pour y remédier : Données centrées réduites.
    $\Rightarrow$ Cette transformation accorde un poids identique aux variables dans le calcul des distances.

  - Age : moyenne = 44 ; écart-type = 9,9
    Revenu : moyenne = 303 333 ; écart-type = 12 472,2
    Surface : moyenne = 26 ; écart-type = 4,3

Données centrées réduites

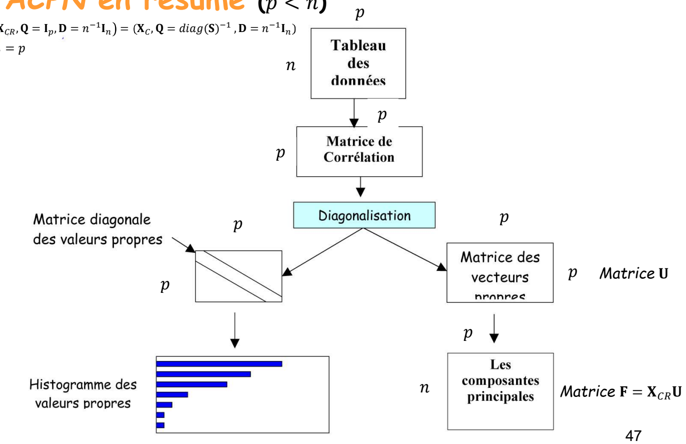
	Age	Revenu	Surface
1	-1,4	-1	-1,4
2	0,6	0	0,9
3	0,8	1	0,5

Distances

d	1	2	3
1	0	3,2	3,5
2	3,2	0	1,1
3	3,5	1,1	0

2. Eléments principaux de l'*ACP*

- ***ACPN en résumé*** $(p < n)$
- $(\mathbf{X}_{CR}, \mathbf{Q} = \mathbf{I}_p, \mathbf{D} = n^{-1}\mathbf{I}_n) = (\mathbf{X}_C, \mathbf{Q} = diag(\mathbf{S})^{-1}, \mathbf{D} = n^{-1}\mathbf{I}_n)$
- $I_g = p$



3. Interprétation et qualité des résultats d'une *ACP*

Construction des nuages de points projetés

- Chaque nuage de points (variables et individus) est construit en projection sur les **plans factoriels** : un plan factoriel est un repère du plan défini par deux des r axes factoriels retenus.
  - Si l'on retient 3 axes, on tracera 3 graphiques pour chaque nuage : le nuage projeté sur le plan (axe1, axe2), celui projeté sur le plan (axe1, axe3) et celui projeté sur le plan (axe2,axe3).
  - **Projection des individus** dans différents plans factoriels : Graphe de $(F^k; F^{k'})$: plan principal $k - k'$
- L'examen des plans factoriels permettra de visualiser :
  - les corrélations entre les variables et
  - d'identifier les groupes d'individus ayant pris des valeurs proches sur certaines variables.

  Mais avant de lire directement les graphiques : il faut interpréter les axes et s'assurer que la projection est fidèle à la réalité !

3. Interprétation et qualité des résultats d'une *ACP*

Indicateurs numériques

- **Pourcentage d'inertie associé à un axe**
  - Rapport de l'inertie projetée associée à l'axe k à l'inertie totale

$$\frac{\lambda_k}{\sum_{\alpha=1}^{r} \lambda_\alpha}$$

- Si ACPN : $\sum_{\alpha=1}^{r} \lambda_\alpha = p$;
- multiplié par 100, on obtient le **% d'inertie exprimé par l'axe k** ;

Interprétation :
- ✓ Mesure de la qualité de représentation des données ;
- ✓ Mesure de l'importance relative de chaque axe.

  - Inertie cumulée

Les axes étant orthogonaux, les % d'inertie s'additionnent pour plusieurs axes

$$\frac{\lambda_1 + \cdots + \lambda_k}{\sum_{\alpha=1}^{r} \lambda_\alpha}$$

Interprétation :
- ✓ Mesure de la qualité de représentation des données dans un espace de dim k.

3. Interprétation et qualité des résultats d'une *ACP*

Interprétation « interne » : individus

- **Interprétation de la position des individus** :
  - Contribution de l'individu i à l'axe k :

aussi appelée **contribution absolue** (ctr)

➢ mesurée par la part d'inertie expliquée par l'individu i sur l'axe k

$$ctr_k(i) = \frac{p_i\left(F^k(i)\right)^2}{\lambda_k}$$

↪ Somme des contributions des individus$=\sum_{i=1}^{n} ctr_k(i) = 100\%$

➢ permet d'identifier les individus les + influents pour un axe donné i.e. qui contribuent bcp à la construction d'un axe factoriel.

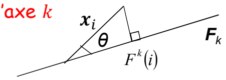
➢ En pratique : On retient pour l'interprétation les individus dont la contribution est > à la contribution moyenne (>1/n), le sens de la contribution dépend du signe de $F^k(i)$.

3. Interprétation et qualité des résultats d'une *ACP*

Interprétation « interne » : individus

- **Interprétation de la position des individus** :
  - Qualité de la représentation d'un individu i sur l'axe k

aussi appelée **contribution relative** (qlt)



$$qlt_k(i) = cos^2(\theta_i^k) = \frac{\left(F^k(i)\right)^2}{\|x_i\|^2} \text{ ; où } \|x_i\|^2 = \sum_{\alpha=1}^{r}\left(F^\alpha(i)\right)^2$$

avec $\sum_{\alpha=1}^{r} qlt_\alpha(i) = 1$

distance entre le pt dans l'espace et sa projection sur l'axe (resp. le plan principal considéré) ; mesurée grâce au cosinus (au carré) de l'angle entre l'axe (resp. le plan principal) et le point $x_i \in \mathbb{R}^p$. (cf. ch1)

- ✓ Lorsque l'angle est proche de 0, c'est-à-dire que le cosinus est proche de 1, l'individu est bien représenté. Dans le cas inverse, l'angle est proche de 90° et le cosinus est proche de 0.
- ✓ Mise en évidence des individus mal représentés.
- ✓ Dans le plan principal $k - k'$: $qlt_{k,k'}(i) = qlt_k(i) + qlt_{k'}(i)$

3. Interprétation et qualité des résultats

Interprétation « interne » : variables

- **Coordonnées des variables sur les axes** :
  - $G^k(j) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} x_i^j F^k(i) = \sqrt{\lambda_k}\, a_k(j)$ *coordonnée de la variable j sur l'axe de rang k*

On va chercher à représenter au mieux les corrélations entre les variables pour interpréter les axes :

- ***Corrélations entre CP et variables initiales***

On a la relation suivante :

$$r\left(F^k; x^j\right) = \left[\frac{\lambda_k}{s_{kk}}\right]^{1/2} a_k(j) = \frac{G^k(j)}{\|x^j\|_D}$$

où $a_k(j)$ est la j ème composante de l'axe principal a_k.

  - Si toutes les variables initiales sont centrées réduites :

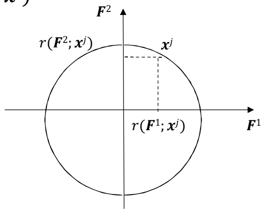
$$r\left(F^k; x^j\right) = \sqrt{\lambda_k}\, a_k(j) = G^k(j)$$

  - Pour un couple de CP F^1 et F^2, on représente ces corrélations linéaires sur une figure appelée **cercle des corrélations**.

3. Interprétation et qualité des résultats

Interprétation « interne » : variables

- **Cercle des corrélations**
  - Chaque variable x^j est repérée par un pt d'abscisse $r(F^1, x^j)$ et d'ordonnée $r(F^2, x^j)$



En ACP normée : $\sum_{k=1}^{r} r^2(F^k; x^j) = \sum_{k=1}^{r} \left(G^k(j)\right)^2 = 1 \;; j \text{ fixé.}$

↪ Les variables qui sont proches du bord du cercle sont celles qui contribuent le plus !

3. Interprétation et qualité des résultats

Interprétation « interne » : variables

- **Interprétation de la position des variables**
  - Qualité de la représentation de la variable x^j sur l'axe k

  aussi appelée **contribution relative** (qlt)

$$qlt_k(j) = cos^2(\theta_j^k) = \frac{\left(G^k(j)\right)^2}{\|x^j\|^2} \text{ où } \|x^j\|^2 = \sum_{\alpha=1}^{r} \left(G^\alpha(j)\right)^2 \text{ avec } \sum_{\alpha=1}^{r} qlt_\alpha(j) = 1$$

➢ Distance entre le point x^j dans l'espace et sa projection sur l'axe (resp. dans le plan principal $k - k'$: $qlt_{k,k'}(j) = qlt_k(j) + qlt_{k'}(j)$).

⟹ Mise en évidence des variables mal représentées.

➢ En ACP normée, $qlt_k(j) = r\left(F^k; x^j\right)^2$:

  - ➢ une variable est d'autant mieux représentée sur un axe qu'elle est proche du bord du cercle des corrélations et de l'axe, d'autant plus mal représentée qu'elle est proche de l'origine.
  - ➢ Les variables qui contribuent le plus à l'axe sont aussi celles qui sont le mieux représentées et inversement, donc pas besoin d'étude spécifique de la représentativité.

3. Interprétation et qualité des résultats

Interprétation « interne » : variables

- **Interprétation de la position des variables**
  - Contribution de la variable j à l'axe k :

  aussi appelée **contribution absolue** (ctr)

  ➢ mesurée par la part d'inertie expliquée par la variable j sur l'axe k

$$ctr_k(j) = \frac{\left(G^k(j)\right)^2 / s_{jj}}{\lambda_k} \text{ pour } \mathbf{Q} = diag(\mathbf{S})^{-1} \text{ et } ctr_k(j) = \frac{\left(G^k(j)\right)^2}{\lambda_k} \text{ si } \mathbf{Q} = \mathbf{I}_p$$

avec $\sum_{j=1}^{p} ctr_k(j) = 100\%$ pour k fixé.

➢ permet d'identifier les variables les + influentes pour un axe donné i.e. qui contribuent bcp à la construction d'un axe factoriel.

➢ <u>En pratique:</u> On retient pour l'interprétation les variables dont la contribution est > à la contribution moyenne (>1/p), le sens de la contribution dépend du signe de $G^k(j)$.

3. Interprétation et qualité des résultats

Interprétation « externe » : variables et individus supplémentaires

On distingue les élts **actifs** des élts **supplémentaires** ou **illustratifs**

Définition :

- ***Eléments actifs :*** variables (ou indiv) contribuant à la construction des axes factoriels contrairement aux
- ***Eléments supplémentaires :*** variables (ou indiv) ne participant pas à la construction des axes ; mais représentées sur les plans principaux.

- **Indiv. supp. :** $F_S^\alpha = (X^S)^T u^\alpha$
  - On peut avoir 2 jeux de données et vouloir étudier l'évolution ou la ≠
  - Certains indiv. sont aberrants ou atypiques ou de nature ≠
  - Certains individus sont artificiels : individu moyen
- **Variables supp :** calcul des proj. sur le cercle des corrélations avec les corrélations $r(x^v; F^k)$
  - Variables que l'on veut lier aux variables actives mais pas lier entre elles
  - Variables que l'on veut expliquer par les variables actives
  - Variables que l'on veut utiliser pour conforter l'interprétation des axes sans faire appel à des variables ayant servi à les déterminer

3. Interprétation et qualité des résultats

- Transformation des données
  - Si données sont hétérogènes, avec ordres de grandeur différents, diagonaliser la matrice des corrélations $\mathbf{R}$
    - ➢ sinon, diagonaliser la matrice des covariances $\mathbf{S}$
- Nombre d'axes à retenir : Pas de règles entièrement satisfaisantes
  - **Qualité globale du nuage** : Calculer les valeurs cumulées successives $\lambda_1/\sum_\alpha \lambda_\alpha$, $(\lambda_1+\lambda_2)/\sum_\alpha \lambda_\alpha$, …pour voir quelle proportion de la somme des variances $\sum_\alpha \lambda_\alpha$ (i.e. quelle part de l'inertie totale) est restituée par les k premiers axes principaux
    - ➢ plus les variables sont nombreuses et moins elles sont corrélées, plus est faible la part d'inertie restituée par les k premiers axes
  - **Critère de Kaiser** : avec la matrice des corrélations (ie sur des données centrées réduites), conserver les axes correspondant aux valeurs propres > 1
  - **Règle du coude** : le diagramme des valeurs propres λ_α montre souvent des cassures dans la baisse des λ_α : on retient les axes avant la cassure
  - **Interprétation** : conserver les axes « interprétables »

3. Interprétation et qualité des résultats

Etapes d'une analyse

- 1ère étape : **Valeurs propres**
  - (% de variation expliquée par chaque composante principale)
  - ⇒ **Déterminer le nombre d'axes à retenir**
- 2ème étape : **Variables**
  - **Qualité de la représentation**
    - ➢ Cosinus carrés
    - ➢ Proximités du cercle de corrélation
  - **Structure des variables**
    - ➢ Corrélations entre variables
    - ➢ Corrélations avec les axes
  - ⇒ **Interpréter les axes**

3. Interprétation et qualité des résultats

Etapes d'une analyse

- 3ème étape : **Individus**
  - **Qualité de la représentation**
    - ➢ Cosinus carrés qlt
    - ➢ Proximités des axes ctr
  - **Répartition des individus**
    - ➢ **Identifier les individus « extrêmes »**
  - ⇒ **Identifier la présence ou non de groupes**
- 4ème étape : **Interprétation conjointe**
  - ⇒ **Identifier des liens variables - individus**

3. Interprétation et qualité des résultats

Interpréter une analyse

- Biplot introduit par Gabriel (1971) : Cf Gower (1996) et Legendre (1998)
  - **Représentation simultanée** à la fois des observations et des variables.
  - Terme réservé aux représentations simultanées qui respectent le fait que la projection des observations sur les vecteurs variables doit être représentative des données d'entrée pour ces mêmes variables.
    - ➢ i.e. les points projetés sur le vecteur variable, doivent respecter l'ordre et les distances relatives des données de départ correspondant à la même variable.
  - La représentation simultanée des observations et des variables :
    - ne peut être faite directement en prenant les coordonnées des variables et des observations dans l'espace des facteurs.
    - Une *transformation* est nécessaire afin de rendre l'interprétation exacte.
  - **3 méthodes** sont proposées

3. Interprétation et qualité des résultats

Interpréter une analyse : Biplot

3 méthodes sont proposées en fonction du type d'interprétation que l'on souhaite faire à partir de la représentation graphique :

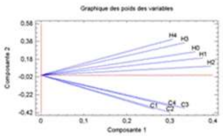
- ***biplot de corrélation*** (*correlation biplot*) : **permet d'interpréter les angles entre les variables** car directement liés aux corrélations entre les variables. La position de deux obs projetées sur un vecteur variable permet de conclure quant à leur niveau relatif sur cette même variable. La distance entre deux observations est une approximation de la distance de Mahalanobis dans l'espace des k facteurs..
- ***biplot de distance*** (*distance biplot*) : **permet d'interpréter les distances entre les observations** car elles sont une approximation de leur distance euclidienne dans l'espace des p variables. La position de deux observations projetées sur un vecteur variable permet de conclure quant à leur niveau relatif sur cette même variable.
- ***biplot symétrique*** (*symmetric biplot*) : proposé par Jobson (1992), **intermédiaire entre les 2 précédents**. Si ni les angles, ni les distances ne peuvent être interprétés, bon compromis.

Sous R : `biplot(stats)`

3. Interprétation et qualité des résultats

Interpréter une analyse : Effet « taille »

- Si l'*ACP* sur les données restitue un surprenant cercle des corrélations où tous les points sont regroupés du même côté d'un axe factoriel et sont caractérisés par des valeurs élevées. Vous avez constaté un *facteur taille* ou *effet taille*.



- **Que signifie ceci ?** Que toutes les variables sont corrélées positivement entre elles et avec le 1er axe. La 1ère composante est la variable latente qui prédit au mieux les autres, c'est aussi celle qui est le mieux prédit par toutes les autres.
- Peut-être jugé comme « polluant » l'*ACP*.

3. Interprétation et qualité des résultats

Interpréter une analyse : Effet « taille »

- **Que faire pour l'éviter ?** si son élimination est nécessaire …
  - soit en délaissant le premier axe factoriel (solution de facilité)
  - soit en utilisant un autre type d'*ACP*, notamment
    - Varimax (Rotation des axes : solution de difficulté, mais les conclusions seront plus sûres) ou
    - une *ACP sur les rangs*,
  - soit en retraitant les données. Une idée possible est par ex. de se servir des pourcentages.

4. Tableau de distances (*PCoA*)

Dans certaines applications, on ne connaît pas les valeurs prises par les variables.

On ne connaît seult que les ***distances*** *entre les individus. Les données de base sont un tableau de distances* $\mathrm{D}_{n \times n}$ *entre les* n *individus du tableau* $\mathrm{X}_{n \times p}$.

- *Cas classique* : Analyse d'un tableau de *distances euclidiennes*
  $\Rightarrow$ **analyse en coordonnées principales** (ou **principal coordinate analysis**) notée ***PCoA*** *(Principal coordinate analysis)*

  Ex. : étude de marché
  On recueille auprès de consommateurs des données de proximités subjectives entre différentes marques concurrentes.

- D'une manière plus générale : ***positionnement multidimensionnel***
  noté ***MDS*** *(Multidimensional Scaling)*
  Pb : représenter graphiquement ses proximités

4. Tableau de distances

- **L'ACP permet de cartographier un ensemble d'indiv. en essayant de déformer le moins possible leurs distances respectives.**
- On a vu que les *CP* $\mathbf{F} = \mathbf{XU}$ sont les vecteurs propres de la matrice $\mathbf{XQX}^{\mathrm{T}}\mathbf{D} = \mathbf{WD}$:

Les facteurs principaux $\boldsymbol{u}$ sont les vecteurs propres de $\mathbf{QS} = \mathbf{QX}^{\mathrm{T}}\mathbf{DX}$ d'où $\mathbf{QX}^{\mathrm{T}}\mathbf{DX}\boldsymbol{u} = \lambda \boldsymbol{u}$ d'où en multipliant à gauche par $\mathbf{X}$, $\mathbf{XQX}^{\mathrm{T}}\mathbf{DF} = \lambda \boldsymbol{F}$

- La matrice $\mathbf{XQX}^{\mathrm{T}}\mathbf{D}$ peut se calculer en connaissant uniqt les distances entre individus. On pourra alors :
  - représenter les individus sur un plan ou un espace de dim q : calcul des vecteurs propres ;
  - mesurer la qualité : calcul du % d'inertie expliquée.

4. Tableau de distances

- *Cas euclidien*
  - La **matrice du carré des distances** entre les pts est :

$$\boxed{\mathbf{D} = \left[d_{ij}^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right]}$$

➢ *Elle est symétrique et sa diagonale est nulle*

  - La matrice $\mathbf{XQX}^{\mathrm{T}} = \mathbf{W}$ est la **matrice des produits scalaires** entre les vecteurs observations

$$\boxed{\mathbf{W} = \left[w_{ij} = \langle \boldsymbol{x}_i ; \boldsymbol{x}_j \rangle_{\mathbf{Q}}\right]}$$

et $w_{ii} = \|\boldsymbol{x}_i\|_{\mathbf{Q}}^{2}$

  - En appliquant la relation triangulaire :

$$\boxed{d_{ij}^2 = \|\boldsymbol{x}_i\|_{\mathbf{Q}}^{2} + \|\boldsymbol{x}_j\|_{\mathbf{Q}}^{2} - 2w_{ij}}$$

4. Tableau de distances

- *Cas euclidien*

➢ Passage de l'une à l'autre, en posant :

$$d_{i\cdot}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 \;;\; d_{\cdot j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 \;;\, d_{\cdot\cdot}^{2} = \frac{1}{n}\sum_{i=1}^{n} d_{i\cdot}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{\cdot j}^2 = 2I_g$$

➢ On déduit la formule de Torgerson (1958) :

$$\boxed{w_{ij} = \frac{1}{2}\left(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^{2}\right)}$$

$\mathbf{W}$ est doublement centrée en lignes et en colonnes.

<u>On peut mq</u> :

La matrice des distances $\mathbf{D}$ est euclidienne
$\Leftrightarrow$ les valeurs propres de $\mathbf{W}$ sont positives ou nulles.

- L'application de l'*ACP* à ce type de données porte le nom **d'analyse factorielle d'un tableau de distances ou Analyse en Coordonnées Principales.**

4. Tableau de distances

- *Cas euclidien*
  - **Matriciellement**

Opérateur de centrage : $\mathbf{A}_{n\times n} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}/n$
Double centrage en lignes et en colonnes, obtention de $\mathbf{W}$ à partir de $\mathbf{D}$:

$$\mathbf{W} = -\tfrac{1}{2}\mathbf{ADA}$$

➢ les vecteurs propres de $\mathbf{W}$ sont les composantes principales $\boldsymbol{F}$ des n points,
➢ le nb de valeurs propres non nulles donne la dimension de l'espace (au plus égale à $n-1$) ; c'est la dimension de l' espace de représentation,
➢ si la distance est euclidienne, aucune valeur propre n'est négative.

<u>Sous R</u> : `dudi.pco(ade4)`

4. Tableau de distances

- La procédure précédente n'est valable que si la distance d^2 est **euclidienne** ; or, on peut avoir d'autres types de distances.

  - ***Peut-on faire la même analyse ?***
    Pas tout à fait …

  $\Rightarrow$ *on appelle les techniques de ce type le* ***positionnement multidimensionnel (MDS)***

4. Tableau de distances

- **Si d n'est pas euclidienne**

certaines des valeurs propres de $\mathbf{W}$ sont négatives

- **Méthode de la constante additive** : en ajoutant c^2 à tous les carrés de distance, on peut la rendre euclidienne

$$\delta_{ij}^2 = d_{ij}^2 + c^2 \quad \text{et} \quad \delta_{ii} = 0$$

$$\mathbf{W}_\delta = \mathbf{W}_d + \mathbf{W}_c$$

$$\mathbf{W}_c = -\frac{1}{2}\mathbf{A}\begin{pmatrix} 0 & c^2 & c^2 & c^2 \\ c^2 & 0 & c^2 & c^2 \\ c^2 & c^2 & 0 & c^2 \\ c^2 & c^2 & c^2 & 0 \end{pmatrix}\mathbf{A} = -\frac{c^2}{2}\mathbf{A}\left(\mathbf{11'} - \mathbf{I}\right)\mathbf{A}$$

$$\textit{puisque} \quad \mathbf{11'} = n(\mathbf{I} - \mathbf{A})$$

$$\mathbf{W}_c = -\frac{c^2}{2}\mathbf{A}\left((n-1)\mathbf{I} - n\mathbf{A}\right)\mathbf{A} = \frac{c^2}{2}\mathbf{A}$$

4. Tableau de distances

- **Si d n'est pas euclidienne**
  - **La méthode de la constante additive**

    - Les vecteurs propres de $\mathbf{W}_\delta = \mathbf{W} + \mathbf{W}_c$ sont les mêmes que ceux de $\mathbf{W}$ car ils sont centrés.
    - Leurs valeurs propres sont augmentées de $c^2/2$
    - Il suffit alors de prendre comme constante :

$$c^2 = 2|\lambda_{min}|$$

    où λ_{min} est la plus petite des valeurs propres négatives de $\mathbf{W}$

    - Transforme directement une dissimilarité (pas d'inégalité triangulaire) en une distance euclidienne !

  $\Rightarrow$ ***il ne faut pas que la constante ajoutée soit trop grande***

  Sous R : `cmdscale(stats)` ou `caillez(ade4)`

5. *ACP* sous R : `Olympic`

1. Avant propos

Les fonctions utilisées

- Il existe un grand nombre de fonctions différentes (**`PCA`**, **`prcomp`**, **`princomp`**, **`dudi.pca`**, **`cca`** ...) se trouvant dans des packages différents (**`vegan`**, **`ade4`**, **`stats`**, **`FactoMineR`**)
- Les fonctions utilisées ici sont disponibles dans les bibliothèques standard de R et dans la bibliothèque **`ade4`**.
  - Pour aider à la compréhension, l'écriture des «programmes» sera détaillée. Par la suite, vous pourrez condenser cette écriture. Mais n'oubliez pas de les commenter abondamment !

5. *ACP* sous R : `Olympic`

2. Rappels

- L'*ACP* est une méthode descriptive.
- Son objectif est de représenter sous forme graphique l'essentiel de l'information contenue dans un tableau de données quantitatif.
- Dans un tableau de données à p variables, les individus se trouvent dans un espace à p dimensions.
- Lorsqu'on projette ces données sur un plan, on obtient un graphique déformé de la réalité.
- Le rôle de l'*ACP* est de trouver des espaces de dimensions plus petites minimisant ces déformations.
- On utilise un espace à 2 dimensions. Ce plan est appelé le ***plan principal***.

73

5. *ACP* sous R : `Olympic`

3. Les données d'exemple

Le tableau des données d'exemple: `olympic` dans `ade4`

Le fichier `olympic` présente les performances au **décathlon** de 33 athlètes lors des jeux olympiques de 1988.

Les variables sont :

dossard : numéro du dossard
m100 : course 100 mètres
long : saut en longueur
poid : lancer du poids
haut : saut en hauteur
m400 : course 400 mètres
m110 : course 110 mètres
disq : lancer du disque
perc : saut à la perche
jave : lancer du javelot
m1500 : course de 1500 mètres

74

5. *ACP* sous R : `Olympic`

4.Préparation du tableau des données :

- Les données seront dans le tableau `olympic` :

```
> library(ade4)
> data(olympic)
> head(olympic$tab)
    100 long  poid haut   400   110  disq perc  jave   1500
1 11.25 7.43 15.48 2.27 48.90 15.13 49.28  4.7 61.32 268.95
2 10.87 7.45 14.97 1.97 47.71 14.46 44.36  5.1 61.76 273.02
3 11.18 7.44 14.20 1.97 48.29 14.81 43.66  5.2 64.16 263.20
4 10.62 7.38 15.02 2.03 49.06 14.72 44.80  4.9 64.04 285.11
5 11.02 7.43 12.92 1.97 47.44 14.40 41.20  5.2 57.46 256.64
6 10.83 7.72 13.58 2.12 48.34 14.18 43.06  4.9 52.18 274.07
```

- On élimine les individus ayant des valeurs manquantes (ici il n'y en a pas …)

```
> olympic.na<-na.omit(olympic$tab)
```

- L'identifiant des individus est `row.names(olympic$tab)` Il contient le numéro de l'athlète.

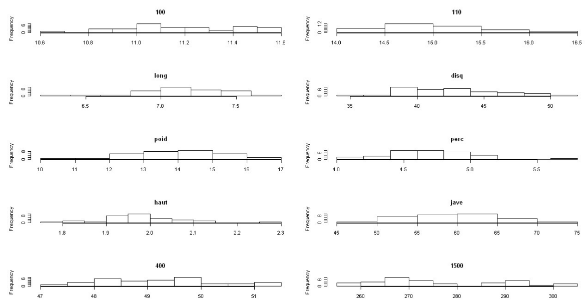
75

5. *ACP* sous R : `Olympic`

5.Description des variables, les histogrammes :

Les histogrammes de toutes les variables

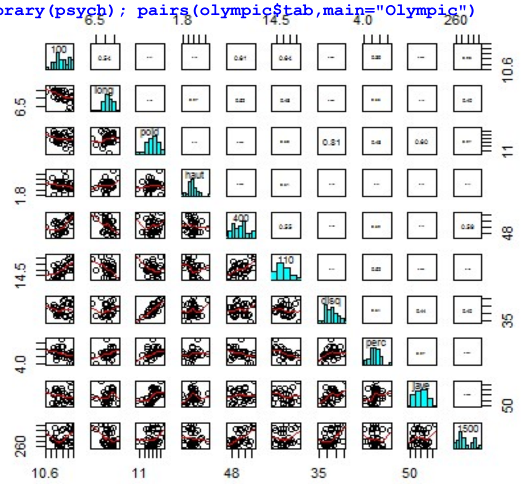
```
> layout(matrix(c(1:10),5,2))
> for(i in 1:10)
  {hist(olympic$tab[,i],main=names(olympic$tab)[i],xlab=
  "")}
> layout(1)
```



5. *ACP* sous R : Olympic

6.Contrôle de la linéarité des relations

```
> library(psych) ; pairs(olympic$tab,main="Olympic")
```



77

5. *ACP* sous R : Olympic

7.L'A.C.P. - fonction dudi.pca (ade4)

- On lance l'*ACP*

Les résultats de l'analyse sont stockés dans la variable z

```
> library(ade4)
> z<- dudi.pca(olympic$tab, center = T, scale = T, scannf = F)
```

- Choix du type d'analyse :

Les options `center` et `scale` de la fonction `dudi.pca` sont utilisées pour centrer et réduire les variables.

78

5. *ACP* sous R : Olympic

7.L'A.C.P. - fonction dudi.pca (ade4)

- Objectifs de cette *ACP*

↪ Description du jeu de données pour le résumer et réduire la dimension du problème :

- *Etude des individus* (*i.e.* des athlètes) : variabilité entre individus
  - 2 athlètes sont proches s'ils ont des résultats similaires.
  - Y a-t-il des similarités entre les individus pour toutes les variables ?
  - Peut-on établir des profils d'athlètes ? Peut-on opposer un groupe d'individus à un autre ?
- *Etude des variables* (*i.e.* des performances) : liaisons linéaires entre les variables.
  - peut-on résumer les performances des athlètes par un petit nombre de variables ?
- *Lien entre les deux études* :
  - peut-on caractériser des groupes d'individus par des variables ?

79

5. *ACP* sous R : Olympic

8. Les valeurs propres et choix du nb d'axes :

- Impression des valeurs propres (variances de chaque composante):

```
> round(z$eig,2) # utiliser round() !!!!!
[1] 3.42 2.61 0.94 0.88 0.56 0.49 0.43 0.31 0.27 0.10
```

- Les variances cumulées (Σ des variances = 10, données centrées réduites) :

```
> round(cumsum(z$eig),2)
[1]  3.42  6.02  6.97  7.85  8.40  8.89  9.32  9.63  9.90 10.00
```

- Les variances en pourcentages et pourcentages cumulés :

```
> round(z$eig/sum(z$eig)*100,2)
[1] 34.18 26.06  9.43  8.78  5.57  4.91  4.31  3.07  2.67  1.02
> round(cumsum(z$eig/sum(z$eig)*100),2)
[1]  34.18  60.25  69.68  78.46  84.03  88.94  93.24  96.31  98.98 100.00
```

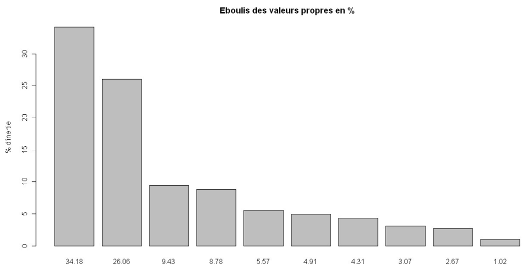
80

5. *ACP* sous R : Olympic

9. L'éboulis des valeurs propres

Une représentation en % de variance expliquée:

```
> inertie<-z$eig/sum(z$eig)*100
> barplot(inertie,ylab="%
  d'inertie",names.arg=round(inertie,2))
> title("Eboulis des valeurs propres en %")
```



81

5. *ACP* sous R : Olympic

10. Interprétation des axes : les contributions absolues et relatives

```
> inertierow<-inertia.dudi(z,row.inertia=TRUE)
> inertiecol<-inertia.dudi(z,col.inertia=TRUE)
> inertierow$TOT # idem résultats p48 : vp inertie

# Analyse des variables
# Coordonnees des variables : G
> round(z$co,2)
# qualite de representation des variables qlt
# contributions relatives : contribution de chaque axe qui explique la
# position d.une variable ctr
> round(inertiecol$col.rel,2)/100 # qlt
# contributions absolues : contributions des variables (sites) aux axes
> inertiecol$col.abs/100 # ctr

# Analyse des individus
# coordonnees des individus : F
> round(z$li,2)
# contrib relative : qualite de representation des indiv : qlt
> round(inertierow$row.rel,2)/100 #qlt
# ctr absolue des indiv : ctr
> inertierow$row.abs/100 #ctr
```

82

5. *ACP* sous R : Olympic

10. Interprétation des axes : les contributions absolues et relatives

Contributions des variables à la construction des axes:

```
> inertia.dudi(z,col.inertia =
T)$col.abs #ctr
```

	Comp1	Comp2
X100	1730	221
long	1553	231
poid	724	2338
haut	451	8
X400	1266	1240
X110	1879	48
disq	309	2533
perc	1475	224
jave	324	1384
X1500	289	1772
Somme	10000	10000

```
> inertia.dudi(z,col.inertia =
T)$col.rel #qlt
```

	Comp1	Comp2
X100	5912	577
long	-5308	-603
poid	-2475	6094
haut	-1540	20
X400	4328	3232
X110	6423	126
disq	-1056	6603
perc	-5043	583
jave	-1107	3606
X1500	990	4619

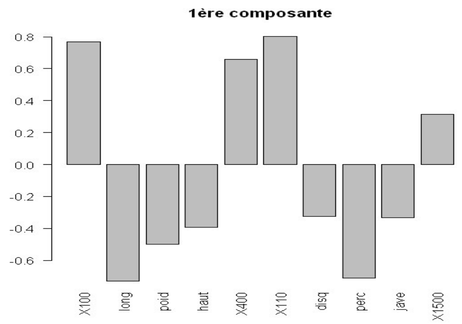
83

5. *ACP* sous R : Olympic

10. Interprétation des axes : graphique des coordonnées

- Profil des coordonnées des variables sur l'axe 1

```
> barplot(cc1, names.arg=row.names(z$co), las=2)
> title(main="1ère composante")
> abline(h=0)
```

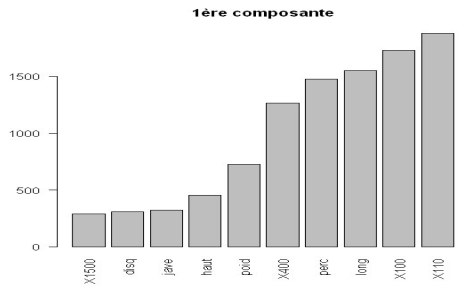


5. *ACP* sous R : Olympic

10. Interprétation des axes : graphique des contributions

- Profil des contributions des variables sur l'axe 1

```
> ctr<-inertia.dudi(z,col.inertia = T)$col.abs
> ctr<-ctr[order(ctr[,1]),] #trier par ordre croissant
> barplot(ctr[,1], names.arg=row.names(z$ctr), las=2)
> title(main="1ère composante")
```



5. *ACP* sous R : Olympic

10. Interprétation des axes : les contributions absolues et relatives

 # Analyse des variables : tableau des ctr

Axe 1		Axe 2	
+	-	+	-
long ()	X110 ()	disq ()	
perc ()	X100 ()	poid ()	

 # Analyse des individus : tableau des ctr

Axe 1		Axe 2	
+	-	+	-

5. *ACP* sous R : Olympic

11. Présentation des résultats – le plan principal

- Le résultat de l'*ACP* ont été stockés dans la variable `z`.
  - Les coordonnées des lignes et des colonnes se trouvent respectivement dans `z$li` et `z$co`

➢ La **première composante** sera :

```
cl1<-z$li[,1] # pour les individus
cc1<-z$co[,1] # pour les variables
```

➢ La **deuxième composante** sera :

```
cl2<-z$li[,2] # pour les individus
cc2<-z$co[,2] # pour les variables
```


5. *ACP* sous R : Olympic

11. Présentation des résultats - le plan des variables

- La représentation graphique du plan des variables :

```
> plot(cc1,cc2,type="n",
  main="Les variables",
  xlim=c(-1,1), ylim=c(-1,1),
  asp=1, #rapport entre "Echelle X" et "Echelle Y"
  ylab= "Comp2 26.1%",
  xlab= "Comp1 34.2%")
> abline(h=0,v=0)
> text(cc1,cc2,row.names(z$co))
```

- Le cercle des corrélations :

```
> symbols(0,0,circles=1,inches=FALSE,add=TRUE)
```

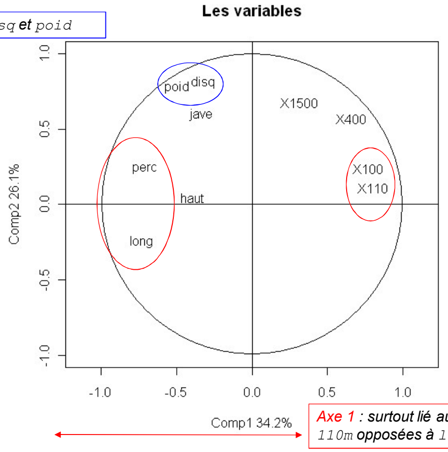
Ou

```
> s.corcircle(z$co)
```

5. *ACP* sous R : Olympic

11. *Présentation des résultats* - plan des variables

Axe 2 : lié à `disq` et `poid`



***Attention** sur axe 1 interprétation particulière de corrélation négative entre les variables `100m`, `110m` opposées à `longueur` et `perche`.*

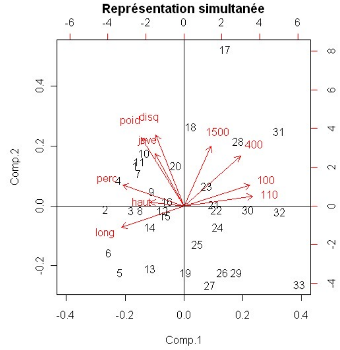
« Un athlète qui court vite réalise un tps faible et … ici réalise souvent de bonnes performances en longueur et perche »

Axe 1 : surtout lié aux variables `100m`, `110m` opposées à `longueur` et `perche`.

5. *ACP* sous R : Olympic

12. Représentation simultanée : "le biplot"

```
> olympinc.cd<-as.matrix(scale(olympic$tab))
# centrées-réduites
> biplot(princomp(olympinc.cd),main="Représentation
  simultanée",xlim=c(-0.4,+0.4))
> abline(v=0,h=0)
```



5. *ACP* sous R : Olympic

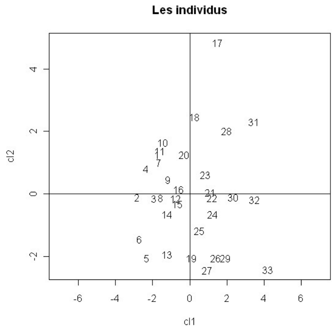
11. Présentation des résultats - le plan des individus

Représentation du graphique du plan des individus :

```
> plot(cl1,cl2,type="n",main="Les individus",xlim=c(-7,7))
> abline(h=0,v=0)
> text(cl1,cl2,row.names(z$li))
```

Axe 1 oppose des profils de perfomance élevés en `100m`, `110m` ;`longueur` et `perche` à des profils relativement faibles dans ces disciplines

Axe 2 met en évidence des profils de performance élevés en `disq` et `poid`



5. *ACP* sous R : Olympic

13. Les données supplémentaires :

- La bibliothèque **ade4** propose les fonctions **supcol()** et **suprow()** pour calculer les coordonnées des variables et individus supplémentaires.
  - *Ces fonctions s'utilisent après le calcul de l'ACP.*

92

5. *ACP* sous R : Olympic

13. Les données supplémentaires : individus supplémentaires

- Les coordonnées des individus supplémentaires sont calculées en tenant compte des options utilisées dans l'*ACP* sur les individus actifs. Ici, ils seront **centré-réduits**.

```
> z<- dudi.pca(olympic$tab, center = T, scale = T, scannf = F)
> ligsup<-suprow(z,olympic$tab[1:3,])
```

- Les coordonnées des individus supplémentaires se trouve dans **ligsup$lisup** :

```
$tabsup
        X100      long      poid       haut       X400       X110      disq      perc      jave      X1500
1  0.22385237 0.9899004 1.1463669  3.1040119 -0.3575966  0.1627406 1.8911534 0.1196240 0.3475947 -0.5270805
2 -1.36208476 1.0566353 0.7575448 -0.1375195 -1.4873485 -1.1798697 0.5477527 1.0950199 0.4288942 -0.2244464
3 -0.06829394 1.0232678 0.1704996 -0.1375195 -0.9367131 -0.4785061 0.3566184  1.3986809 0.8723462 -0.9546350

$lisup
      Axis1      Axis2
1 -1.759830  1.2504617
2 -2.830456 -0.1022408
3 -1.908359 -0.1390149
```

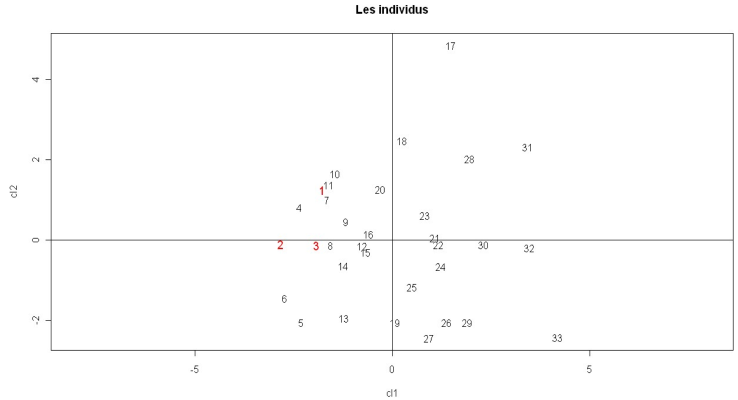

5. *ACP* sous R : Olympic

13. Les données supplémentaires : individus supplémentaires

Représentation simultanée des individus actifs et supplémentaires

```
#coordonnées des individus actifs
> cl1<-z$li[,1]
> cl2<-z$li[,2]
#coordonnées des individus supplémentaires
> csup1<-ligsup$lisup[,1]
> csup2<-ligsup$lisup[,2]
#le graphique "vide"
> plot(cl1,cl2,type="n",main="Les individus",xlim=c(-8,8))
> abline(h=0,v=0)
#on ajoute les individus actifs
> text(cl1,cl2,row.names(z$li),)
#on ajoute les individus supplémentaires
> text(csup1,csup2,row.names(ligsup$lisup),col="red",cex=1.2)
```


13. Les données supplémentaires : individus supplémentaires

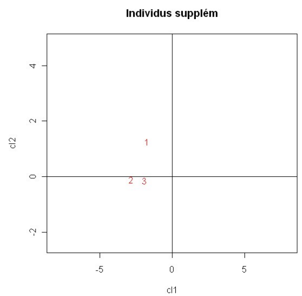


5. *ACP* sous R : Olympic

13. Les données supplémentaires : individus supplémentaires

Pour plus de lisibilité, on peut aussi **représenter séparément** les individus supplémentaires

```
> plot(cl1,cl2,type="n",main="Individus supplém",xlim=c(-8,8))
> abline(h=0,v=0)
> text(csup1,csup2,row.names(ligsup$lisup),col="red",cex=1.1)
```



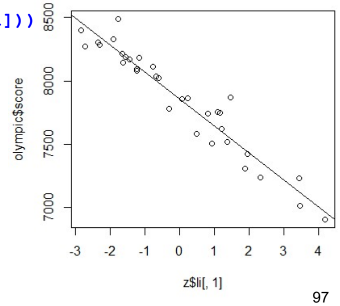
5. ACP sous R : Olympic

14. L'ACP et régression …

Dans la table `olympic` , on dispose aussi d'un vecteur `score` représentant les scores finaux obtenus par chaque compétiteur à la compétition, il est alors possible de prévoir le score final de chaque athlète en fonction de la 1ère CP, au lieu d'utiliser les 10 variables et … ça marche bien !

```
> plot(z$li[,1],olympic$score)
> abline(lm(olympic$score~z$li[,1]))
```

Prévision du score en fonction de la 1ère composante principale par lm



97

5. ACP sous R : Olympic

15. L'ACP avec FactoMineR …

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

Autre Package ***R*** :

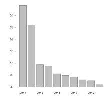
- dédié à l'analyse exploratoire multidimensionnelle de données « à la Française »
- permet de réaliser :
  - des analyses classiques telles que ACP, Analyse Factorielles des Correspondances (AFC) et Analyse Factorielles des Correspondances Multiples (AFCM) ainsi que
  - des analyses plus avancées.
- développé et maintenu par F. Husson, J. Josse, S. Lê (Agrocampus Rennes), et J. Mazet.

98

5. ACP sous R : Olympic

15. L'ACP avec FactoMineR … PCA( )

```
> library(FactoMineR)
# ACPN
> res.pca = PCA(olympic$tab, scale.unit=TRUE, ncp=2, graph=T)
# choisir le nb d'axes
> barplot(res.pca$eig[,2],
          names=paste("Dim",1:nrow(res.pca$eig)))
# % d'inertie expliquée par chaque axe
> round(res.pca$eig,2)
```



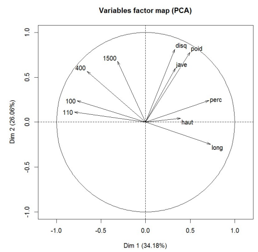
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.42	34.18	34.18
comp 2	2.61	26.06	60.25
comp 3	0.94	9.43	69.68
comp 4	0.88	8.78	78.46
comp 5	0.56	5.57	84.03
comp 6	0.49	4.91	88.94
comp 7	0.43	4.31	93.24
comp 8	0.31	3.07	96.31
comp 9	0.27	2.67	98.98
comp 10	0.10	1.02	100.00

99

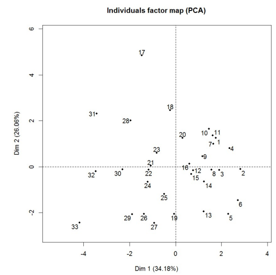
5. ACP sous R : Olympic

15. L'ACP avec FactoMineR … PCA( )

```
# 2 axes : analyse graphiques des résultats
```



ACP Olympic Plan 1-2 : cercle des corrélations



ACP Olympic Plan 1-2 : représentation des individus

100

5. ACP sous R : Olympic

15. L'ACP avec FactoMineR … PCA( )

```
# Précisions sur les interprétations
# Pour les individus : res.pca$ind
# tableau  : coord, qlt de la projection et ctr des individus pour chaque axe
> round(cbind(res.pca$ind$coord[,1:2],res.pca$ind$cos2[,1:2],res.pca$ind$contrib[,1:2]),digits=2)
  Dim.1 Dim.2 Dim.1 Dim.2 Dim.1 Dim.2
1  1.76  1.25  0.19  0.10  2.75  1.82
2  2.83 -0.10  0.90  0.00  7.10  0.01
3  1.91 -0.14  0.61  0.00  3.23  0.02
4  2.35  0.81  0.57  0.07  4.89  0.76
5  2.30 -2.05  0.47  0.38  4.67  4.89
6   2.72 -1.45  0.51  0.15  6.54  2.46
7   ...
# Pour les variables : res.pca$var
> round(cbind(res.pca$var$coord[,1:2],res.pca$var$cos2[,1:2],res.pca$var$contrib[,1:2]),digits=2)
     Dim.1 Dim.2 Dim.1 Dim.2 Dim.1 Dim.2
100  -0.77  0.24  0.59  0.06 17.30  2.21
long  0.73 -0.25  0.53  0.06 15.53  2.31
poid  0.50  0.78  0.25  0.61  7.24 23.38
haut  0.39  0.05  0.15  0.00  4.51  0.08
400  -0.66  0.57  0.43  0.32 12.66 12.40
110  -0.80  0.11  0.64  0.01 18.79  0.48
disq  0.33  0.81  0.11  0.66  3.09 25.33
perc  0.71  0.24  0.50  0.06 14.75  2.24
jave  0.33  0.60  0.11  0.36  3.24 13.84
1500 -0.31  0.68  0.10  0.46  2.89 17.72
```

```
summary(res.pca,ncp=2,nbelements=3)
```


5. ACP sous R : Olympic

15. L'ACP avec Factoshiny

Factoshiny permet d'utiliser une interface graphique interactive et conviviale pour réaliser les analyses précédentes via la fonction **PCAshiny** !

```
library(Factoshiny)
PCAshiny(olympic$tab)
```


5. ACP sous R : Olympic

15. L'ACP avec FactoMineR … PCA( )

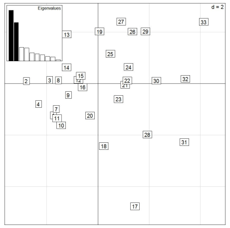
```
# description automatique des principales dimensions de variablité : dimdesc( )
# Tri des quantitavives en fonction de cor(X,F), seuls les + signif sont conservés
> dimdesc(PCA(olympic$tab, scale.unit=TRUE, graph=F))
 $Dim.1
$Dim.1$quanti
     correlation      p.value
long   0.7285412 1.533797e-06
perc   0.7101094 3.679521e-06
poid   0.4975355 3.218939e-03
haut   0.3924767 2.387193e-02
400   -0.6579077 3.170485e-05
100   -0.7689031 1.723515e-07
110   -0.8014415 2.095129e-08
$Dim.2
$Dim.2$quanti
     correlation      p.value
disq   0.8126000 9.288574e-09
poid   0.7806386 8.397179e-08
1500   0.6796201 1.364084e-05
jave   0.6004996 2.201837e-04
400    0.5685383 5.565015e-04
$Dim.3
$Dim.3$quanti
     correlation      p.value
haut   0.8303924 2.257166e-09
```

Pour aller plus loin, voir par exemple :
http://factominer.free.fr/classical-methods/analyse-en-composantes-principales.html

6. Analyse d'un Tableau de distances sous R

1. Distances euclidiennes : Olympic

```
> library(ade4) ; data(olympic)
> D.Olympic<- dist(scale(olympic$tab))
# matrice de distances données centrées réduites !!!
> dim(as.matrix(D.Olympic))# Dimension de la matrice de distances 33*33
# Analyse en coordonnées principales de la matrice de distances
> Olympic.pco<-dudi.pco(D.Olympic,nf=2)
> scatter(Olympic.pco)
# coordonnées dans le plan 1-2 des 33 individus
```



```
# Autre solution cmdcsale( )
# pour  Classical (Metric) Multidimensional Scaling
# ou  principal coordinates analysis (Gower, 1966)
> cmdscale(D.Olympic,k=2,eig=T)
# Analyse factorielle de la matrice de distance
> Olympic.mds$eig # valeurs propres = (n-1)*valeurs propres de l'ACP
> colMeans(Olympic.mds$points) # la matrice est centrée
>  round(diag(var(Olympic.mds$points)),3)
# Variances des colonnes = valeurs propres de l'ACPN
```

6. Analyse d'un Tableau de distances sous R

2. Distances non euclidiennes : capitales(ade4)

```
> library(ade4) ; data(capitales)
# cmds(stats) ou caillez(ade4) fournissent une approximation de la cte c²
# positionnement multidimensionnel
> dim(as.matrix(capitales$dist))
[1] 15 15
> attr(capitales$dist, "Labels")
 [1] "Madrid"     "Paris"      "London"     "Dublin"     "Rome"       "Brussels"
"Amsterdam"  "Berlin"     "Copenhagen" "Stockholm"  "Luxembourg" "Helsinki"
"Vienna"     "Athens"  "Lisbon"
> d0 <- as.dist(capitales$dist)
> is.euclid(d0)
[1] FALSE
> d1 <- cailliez(d0, TRUE) # Transformation to make Euclidean a distance matrix
Cailliez constant = 2429.87867
> is.euclid(d1)
[1] TRUE
> plot(d0, d1) ; abline(lm(unclass(d1)~unclass(d0)))
> print(coefficients(lm(unclass(d1)~unclass(d0))), dig = 8) # d1 = d + Cte
> is.euclid(d0 + 2428) # FALSE
> is.euclid(d0 + 2430) # TRUE the smallest constant
```

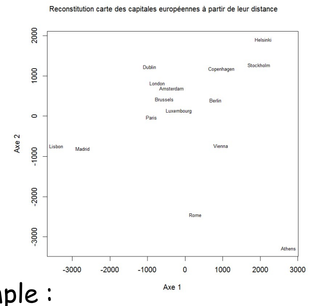
105

6. Analyse d'un Tableau de distances sous R

2. Distances non euclidiennes : capitales(ade4)

```
# d1 <- cailliez(d0, TRUE) # Transformation to make Euclidean a distance matrix
Cailliez constant = 2429.87867
> capitales.mds<-
cmdscale(d1,k=2,add=T,eig=T)
> plot(capitales.mds$points[,1:2],
   xlab="Axe 1",ylab="Axe 2",type="n")
> text(capitales.mds$points[,1:2],
 label= attr(capitales$dist, "Labels"),cex=0.7)

> mtext(outer=T,"Reconstitution carte des
capitales européennes à partir de leur
distance",side="3",line=-2,cex=1.0)
```



Pour aller plus loin , voir par exemple :

- la fonction **wcmdscale** *(Weighted Classical Multidimensional Scaling)* dans la library **vegan** ;
- la fonction **smacofSym** *(Multidimensional scaling (stress minimization: SMACOF) on symmetric dissimilarity matrix)* dans la library **SMACOF** ;
- http://www.r-bloggers.com/7-functions-to-do-metric-multidimensional-scaling-in-r/

106

Ce qu'il faut retenir

- *Il existe plusieurs méthodes d'analyse de données. on retiendra la plus appropriée en fonction :*
  - *du type de données dont on dispose, et*
  - *de l'objectif recherché.*
- *L'ACP en est une. Elle permet de :*
  - *décrire et de synthétiser sous forme de cartes l'information contenue dans un tableau de données quantitatives en essayant de déformer le moins possible les distances entre les points ;*
  - *simplifier et schématiser les liaisons entre variables ; détecter des liaisons entre variables ;*
  - *localiser les regroupements d'observations ou de variables ;*
  - *détecter des observations exceptionnelles ou aberrantes, d'éventuels groupes isolés d'observations ;*
  - *construire des variables synthétiques non-corrélées (régression sur composantes principales, …).*
- *PCoA et, de manière plus générale, MDS permettent quant à elles de traiter directement d'une matrice de distances ou de dissemblances.*

107

Exercices Td/TP ch II: *ACP « à la main »*, fonction sous R

Données supplémentaires

Données	Description
Eaux1	Corpus 20 eaux minérales décrites par 7 variables
Eaux2	Corpus données supplémentaires
Projet M1 AD 1920.csv	Corpus 1000 gènes pour 40 individus
Diabete	Corpus de 46 patients et 5 variables
Glucose	Résultats d'analyse du glucose dans le sang à trois occasions pour 52 femmes à jeun et une heure après avoir consommé du sucre
Capitales	capitales(ade4)

Travail sur **Projet M1 AD 1920.csv** *à rendre par binôme*
au plus tard semaine 43 (IS) ou 45 (Info) lors du dernier TP :
1 version papier + envoi mail (version pdf, données et script R)

108

Références logiciels

- **Logiciel** :
  URL http://www.R-project.org.
  - **library ade4 :**
    - Jean Thioulouse, Anne-Beatrice Dufour and Daniel Chessel (2004). ade4: Analysis of Environmental Data : Exploratory and Euclidean methods in Environmental sciences. R package version 1.3-3. http://pbil.univ-lyon1.fr/ADE-4
  - **library** FactoMineR
- ***Autres logiciels*** **:**
  - SAS (proc princomp, Macros INSEE)
  - STATA (commande pca)
  - SPAD

109

Références bibliographiques

- L. Bellanger, R. Tomassone, *Exploration de données et méthodes statistiques : Data analysis & Data mining avec R. Collection Références Sciences*, Editions Ellipses, Paris, 2014.
- A. Bouchier, Documents et supports de cours disponibles sur le site : http://rstat.ouvaton.org/
- J.-M. Bouroche & G. Saporta, *L'analyse des données*. Presses Universitaires de France : Que sais-je ? 85, Paris, 1992.
- J.-F. Durand, *support intitulé « Elts de Calcul matriciel et d'Analyse Factorielle de Données » disponible sur le site:* www.math.univ-montp2.fr/~durand
- F. Husson, S. Lê & J. Pagès, *Analyse de données avec R*. PUR, Rennes, 2009.
- L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2006.
- G. Saporta, *Probabilités, Analyse des données*. Editions Technip, Paris, 2006.
- Statistics with R : http://zoonek2.free.fr/UNIX/48_R/all.html
- Une belle galerie de graphiques effectués avec R : http://addictedtor.free.fr/graphiques/

110