Cours d'Analyse des Données

Cours n°= 2 : Analyse en Composantes Principales (ACP)

Présenté par Mr. *Hamel Elhadj*

2021/2022

département de mathématiques
université de chlef

INTRODUCTION

Données :

n individus observés sur p variables quantitatives.

L'A.C.P. permet d'explorer les liaisons entre variables et les ressemblances entre individus.

Résultats :

⇨ **Visualisation des individus**

(Notion de distances entre individus)

⇨ **Visualisation des variables**

(en fonction de leurs corrélations)

PRINCIPE DE L'A.C.P.

On cherche une représentation des n individus , dans un sous-espace F_k de R^p de dimension k (k petit 2, 3 …; par exemple un plan)

Autrement dit, on cherche à définir **k nouvelles variables combinaisons linéaires des p variables initiales** qui feront perdre le moins ***d'information*** possible.

Ces variables seront appelées «***composantes principales*** »,

les axes qu'elles déterminent : « ***axes principaux*** »

les formes linéaires associées : « ***facteurs principaux*** »

« Perdre le moins d'information possible »

① F_k devra être « ajusté » le mieux possible au nuage des individus: la somme des carrés des distances des individus à F_k doit être minimale.

② F_k est le sous-espace tel que le nuage projeté ait une **inertie** (dispersion) maximale.

① et ② sont basées sur les notions de :

distance

projection orthogonale

Quel type de données ?

L'ACP s'intéresse à des tableaux de données rectangulaires avec des individus en lignes et des variables quantitatives en colonnes

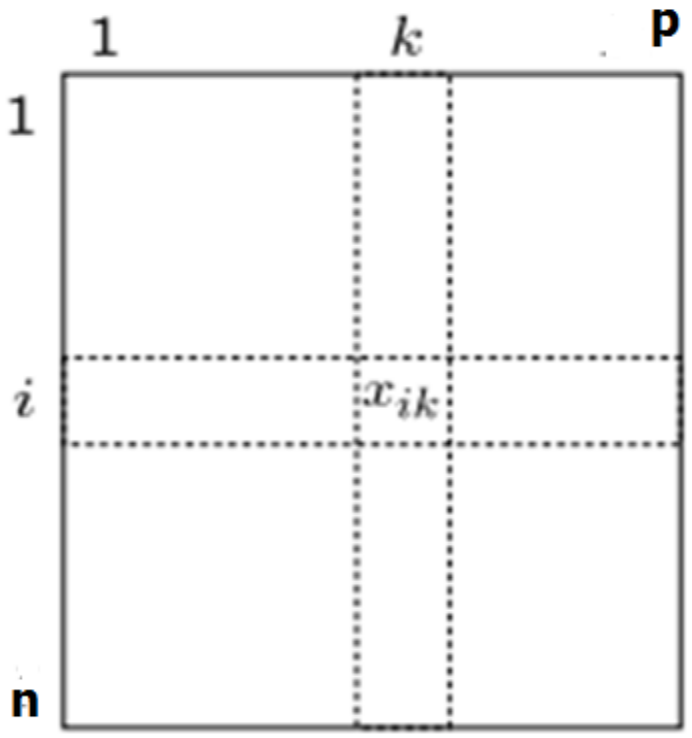


FIGURE : Tableau de données en ACP

Pour la variable k, on note :

$$\text{la moyenne : } \bar{x}_k = \frac{1}{n}\sum_{i=1}^{I} x_{ik}$$

$$\text{l'écart-type : } \sigma_k = \sqrt{\frac{1}{n}\sum_{i=1}(x_{ik} - \bar{x}_k)^2}$$

Exemples

- Analyse sensorielle : note du descripteur k pour le produit i
- Ecologie : concentration du polluant k dans la rivière i
- Economie : valeur de l'indicateur k pour l'année i
- Génétique : expression du gène k pour le patient i
- Biologie : mesure k pour l'animal i
- Marketing : valeur d'indice de satisfaction k pour la marque i
- Sociologie : temps passé à l'activité k par les individus de la CSP i
- etc.

$\Rightarrow$ Il existe de très nombreux tableaux comme cela

Description de données quantitatives

Définition On appelle variable un vecteur $\mathbf{x}$ de taille n. Chaque coordonnée x_i correspond à un individu. On s'intéresse ici à des valeurs numériques.

Poids Chaque individu a éventuellement un poids p_i, tel que $p_1 + \cdots + p_{n=1}$. On a souvent $p = 1/n$.

Représentation histogramme en découpant les valeurs de la variable en classes ; ou alors « boîte à moustache ».

Résumés on dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, étendue, 1er quartile, 3ème quartile, ... Ces indicateurs mesurent principalement la tendance centrale et la dispersion.

On utilisera principalement la moyenne, la variance et l'écart type.

Moyenne arithmétique

Définition On note

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

ou pour des données pondérés

$$\bar{x} = \sum_{i=1}^{n} n_i x_i.$$

Propriétés la moyenne arithmétique est une mesure de *tendance centrale* qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée à cause de ses bonnes propriétés mathématiques.

Variance et écart-type

Définition la *variance* de $\mathbf{x}$ est définie par

$$s_{\mathbf{x}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ ou } s_{\mathbf{x}}^2 = \sum_{i=1}^{n} n_i(x_i - \bar{x})^2$$

L'*écart-type* $s_{\mathbf{x}}$ est la racine carrée de la variance.

Propriétés La variance satisfait la formule suivante

$$s_{\mathbf{x}}^2 = \sum_{i=1}^{n} n_i x_i^2 - (\bar{x})^2$$

La variance est « la moyenne des carrés moins le carré de la moyenne ». L'écart-type, qui a la même unité que $\mathbf{x}$, est une mesure de *dispersion*.

Mesure de liaison entre deux variables

Définitions la covariance observée entre deux variables $\mathbf{x}$ et $\mathbf{y}$ est

cov(x,y)= $$s_{\mathbf{xy}} = \sum_{i=1}^{n} n_i(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{p} n_i x_i y_i - \bar{\mathbf{x}}\bar{\mathbf{y}}.$$

et le *coefficient de r de Bravais-Pearson* ou coefficient de corrélation est donné par

$$r_{\mathbf{xy}} = \frac{s_{\mathbf{xy}}}{s_{\mathbf{x}} s_{\mathbf{y}}} = \frac{\sum_{i=1}^{n} n_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} n_i(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} n_i(y_i - \bar{y})^2}}.$$

$$-1 \leq r_{\mathbf{xy}} \leq 1.$$

Notions de base

On considère un tableau de données numériques où n individus sont décrits sur p variables.

	1 ...	j	... p
1			
$\vdots$		$\vdots$	
i	...	x_{ij}	...
$\vdots$		$\vdots$	
n			

On notera :

$\mathbf{X} = (x_{ij})_{n \times p}$ la matrice des données brutes où $x_{ij} \in \mathbb{R}$ est la valeur du $i^{\text{ème}}$ individu sur la $j^{\text{ème}}$ variable.

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p$$

la description du $i^{\text{ème}}$ individu (ligne de $\mathbf{X}$)

$$\mathbf{x}^j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \in \mathbb{R}^n$$

la description de la $j^{\text{ème}}$ variable (colonne de $\mathbf{X}$).

Les données

On observe n individus, et q variables (quantitatives, sur $\mathbb{R}$). Le nuage de points peut se décomposer de deux manières,

- l'espace des individus, i.e. $\mathbb{R}^p$
- l'espace des variables, i.e. $\mathbb{R}^n$

On note x_{ij} l'observation de la jème variable sur le ième individu. On a donc le tableau de données

		variables				
		1	⋯	j	⋯	p
individus	1	x_{11}	⋯	x_{1j}	⋯	x_{1p}
⋮	⋮	⋮		⋮		⋮
	i	x_{i1}	⋯	x_{ij}	⋯	x_{ip}
⋮	⋮	⋮		⋮		⋮
	n	x_{n1}	⋯	x_{nj}	⋯	x_{np}

Le tableau de données $X = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ est une matrice rectangulaire de taille $n \times p$.

Notations

	1 ...	j	... p
1			
i	...	x_{ij}	...
n			
$\bar{x}$	...	$\bar{x}^j$	...
s	...	s_j	...

Matrice **X** des données brutes

	1 ...	j	... p
1			
i	...	y_{ij}	...
n			
$\bar{y}$	...	0	...

Matrice **Y** des données centrées

$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}^j)^2$ est la variance de la variable j (colonne j de **X**),

Les colonnes de la matrice centrée **Y** sont de moyenne nulle : $\bar{y}^j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = 0.$

Notations

	1 . . .	j	. . . p
1			
.		.	
.		.	
.		.	
i	. . .	z_{ij}	. . .
.		.	
.		.	
.		.	
n			
$\bar{z}$	. . .	0	. . .
s	. . .	1	. . .

$z_{ij} = \frac{x_{ij} - \bar{x}^j}{s_i}$ est le terme général de la matrice $\mathbf{Z}$ des données centrées-réduites.

$$Z = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sigma_1} & \frac{x_{12}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{1p}-\bar{x}_p}{\sigma} \\ \frac{x_{21}-\bar{x}_1}{\sigma_1} & \frac{x_{22}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{2p}-\bar{x}_p}{\sigma} \\ \vdots & & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sigma_1} & \frac{x_{n2}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{np}-\bar{x}_p}{\sigma_p} \end{pmatrix}$$

Matrice $\mathbf{Z}$ des données centrées-réduites

Les colonnes de la matrice centrées-réduites $\mathbf{Z}$ sont de moyenne 0 et de variance 1 :

$$\bar{z}^j = \frac{1}{n}\sum_{i=1}^{n} z_{ij} = 0, \; var(z^j) = \frac{1}{n}\sum_{i=1}^{n}(z_{ij} - \bar{z}^j)^2 = 1.$$

Notations

$$N = \begin{pmatrix} \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_n \end{pmatrix}$$

Dans le nuage des individus :

- chaque individu i est un point $\mathbf{x}_i$ de $\mathbb{R}^p$ (une ligne de $\mathbf{X}$),
- chaque individu i est pondéré avec un poids w_i. En pratique :
  - $w_i = \frac{1}{n}$ pour des tirage aléatoire par exemple.
  - $w_i \neq \frac{1}{n}$ pour des échantillons redressés, des données regroupées,etc.

Les données sont souvent pré-traitées avant d'être analysées. On pourra :

- centrer les données pour avoir des colonnes (variables) de moyenne nulle,
- réduire les données pour avoir des colonnes (variables) de variance égale à 1.

Exemple :

mesure de la tension artérielle diastolique, systolique et du taux de cholestérol de 6 patients.

ind\var	diast	syst	chol
1	90	140	60
2	60	85	5,9
3	75	135	6,1
4	70	145	5,8
5	85	130	5,4
6	70	145	5,0

$\Rightarrow$ Deux nuages de points.

Le premier nuage de points est le nuage des individus.

Le second nuage de points associé à une matrice de données quantitatives est le nuage des variables.

Exemple :

Matrice X des données brutes

```
##       diast syst chol
## ind1     90  140  6.0
## ind2     60   85  5.9
## ind3     75  135  6.1
## ind4     70  145  5.8
## ind5     85  130  5.4
## ind6     70  145  5.0
```

Moyennes et écart-types des colonnes de X

```
##             diast  syst  chol
## moyenne        75 130.0 5.700
## écart-type     10  20.8 0.383
```

Matrice Y des données centrées

```
      diast syst chol
        15   10  0.3
       -15  -45  0.2
         0    5  0.4
        -5   15  0.1
        10    0 -0.3
        -5   15 -0.7
```

Moyennes des colonnes de Y

```
## diast  syst  chol
##     0     0     0
```

Matrice Z des données centrées-réduites

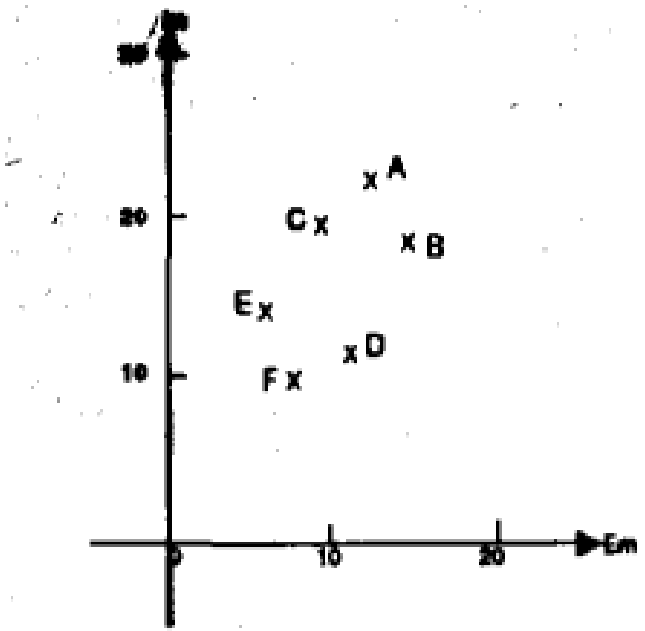
```
  diast  syst  chol
   1.5  0.48  0.78
  -1.5 -2.16  0.52
   0.0  0.24  1.04
  -0.5  0.72  0.26
   1.0  0.00 -0.78
  -0.5  0.72 -1.83
```

Moyennes et écart-types des colonnes de Z

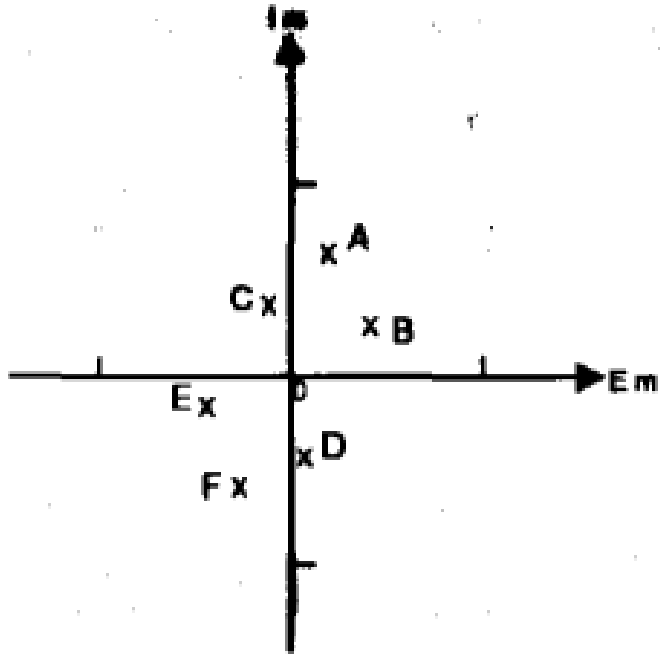
```
## diast  syst  chol
##     0     0     0
## diast  syst  chol
##     1     1     1
```

En résumé, trois nuages de points-individus.

Centrer-réduire les données revient à faire une translation et une normalisation du nuage des individus.

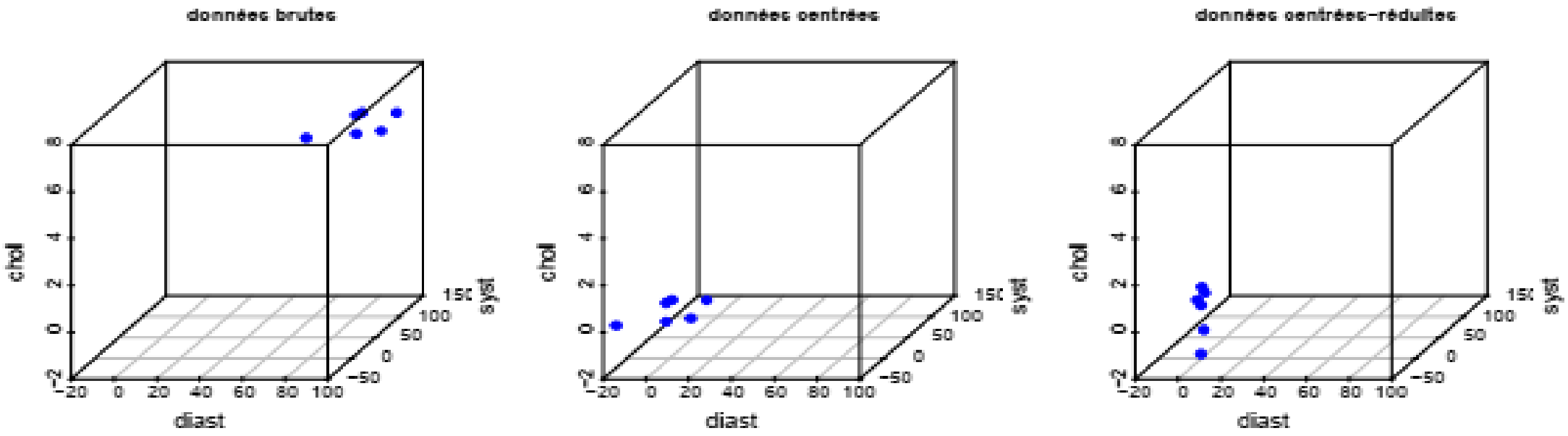


Le nuage de points brut.



Le nuage de points centré-

En résumé, trois nuages de points-individus.



- Centrer les données ne modifie pas les distances entre les individus :

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) = d^2(\mathbf{y}_i, \mathbf{y}_{i'}).$$

- Centrer-réduire les données modifie les distances entre les individus :

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) \neq d^2(\mathbf{z}_i, \mathbf{z}_{i'}).$$

La proximité entre deux individus

- La proximité entre deux individus se mesure avec la distance Euclidienne.
  Lorsque les données sont brutes (pas de pré-traitement) la distance Euclidienne entre deux individus *i* et *i'* (deux lignes de **X**) est :

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2.$$

- Lorsque les données centrées-réduites la distance Euclidienne entre deux individus *i* et i' (deux lignes de **Z**) est

$$d^2(\mathbf{z}_i, \mathbf{z}_{i'}) = \sum_{j=1}^{p} \frac{1}{s_j^2} (x_{ij} - x_{i'j})^2.$$

On en déduit que :

- si les variables (les colonnes de **X**) sont mesurées sur des échelles différentes, les variables de forte variance auront plus de poids dans le calcul de la distance Euclidienne que les variables de petite variance,
- centrer-réduire les données permet donc de donner le même poids à toutes les variables dans le calcul de la distance entre deux individus.

Exemple : distance entre ind1 et ind2

Données brutes ($\mathbf{X}$) :

diast	syst	chol
90	140	6.0
60	85	5.9
75	135	6.1
70	145	5.8
85	130	5.4
70	145	5.0

Données centrées-réduites ($\mathbf{Z}$)

diast	syst	chol
1.5	0.48	0.78
-1.5	-2.16	0.52
0.0	0.24	1.04
-0.5	0.72	0.26
1.0	0.00	-0.78
-0.5	0.72	-1.83

Moyennes et écart-types des colonnes :

	diast	syst	chol
moyenne	75	130.0	5.700
écart-type	10	20.8	0.383

Distance Euclidienne entre les deux premières lignes de $\mathbf{X}$:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(90-60)^2 + (140-85)^2 + (6-5.9)^2}$$
$$= \sqrt{30^2 + 55^2 + 0.1^2}$$

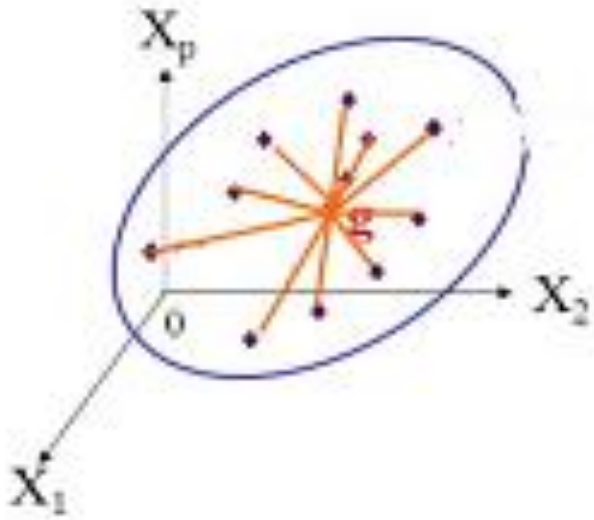
Distance euclidienne entre les deux premières lignes de $\mathbf{Z}$:

$$d(\mathbf{z}_1, \mathbf{z}_2) = \sqrt{\frac{1}{10^2}(90-60)^2 + \frac{1}{20.8^2}(140-85)^2 + \frac{1}{0.383^2}(6-5.9)^2}$$
$$= \sqrt{(1.5+1.5)^2 + (0.48+2.16)^2 + (0.78-0.52)^2}$$
$$= \sqrt{3^2 + 2.7^2 + 0.26^2}$$

La dispersion du nuage des individus (l'inertie.)

Inertie totale du nuage de points

	X_1	...	X_p	
1				
⋮				
i	x_{1i}	...	x_{pi}	$\mathbf{x_i}$
⋮				
n				
	$\bar{x}_1$	...	$\bar{x}_p$	**g**

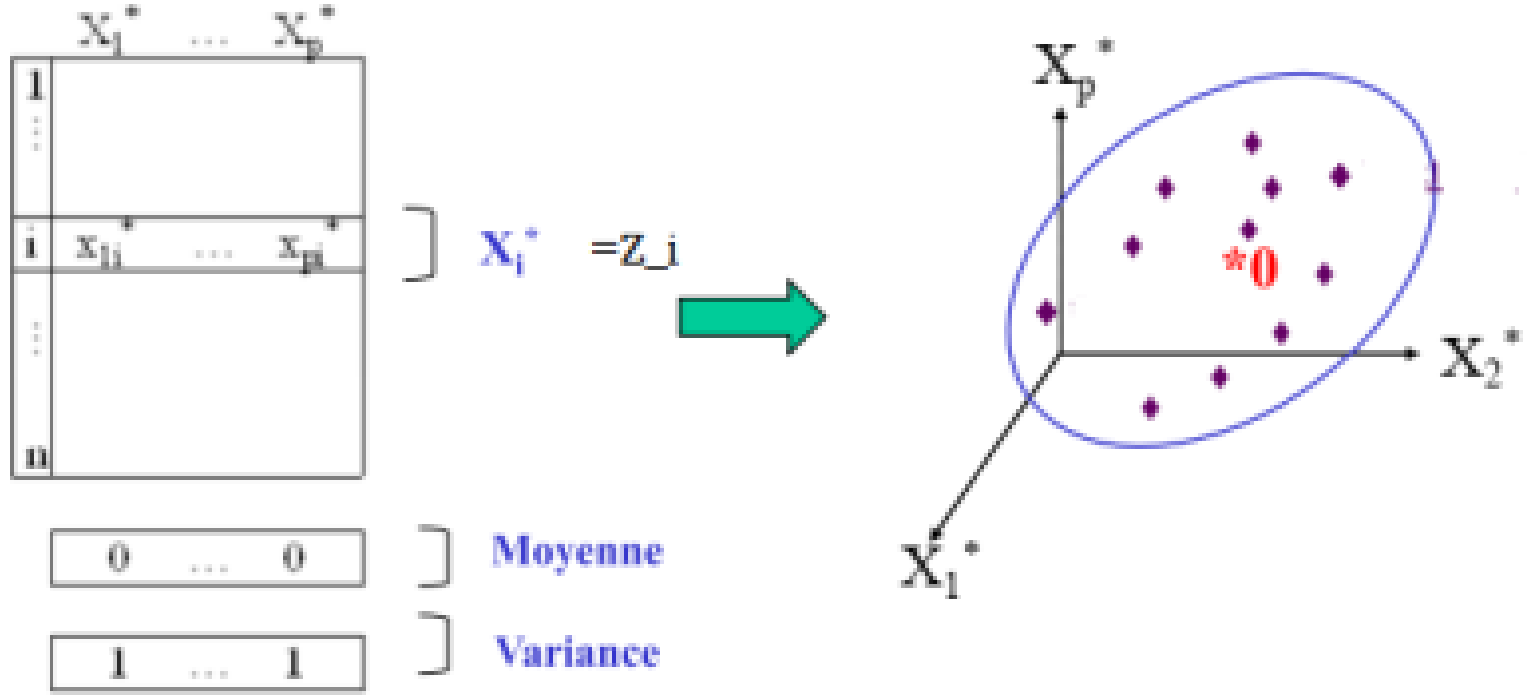


$$\text{Inertie totale} = I(N, g) = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g)$$

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

La dispersion du nuage des individus (l'inertie.)

Le nuage de points associé aux données réduites



$x^* = \{x_1^*, \ldots, x_i^*, \ldots, x_n^*\}$

Centre de gravité : $g^* = 0$, **Inertie totale :** $I(x^*, 0) = p$

La dispersion du nuage des individus (l'inertie.)

La dispersion du nuage des individus se mesure avec l'inertie.

- Lorsque les données sont brutes (pas de pré-traitement), l'inertie des individus (les n lignes de $\mathbf{X}$) est :

$$I(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} d^2(\mathbf{x}_i, \bar{\mathbf{x}}).$$

- L'inertie est donc une généralisation de la notion de variance au cadre multivarié la dispersion des données sur mesure sur p variables.

- On peut montrer que :

$$I(\mathbf{X}) = \sum_{j=1}^{p} var(\mathbf{x}^j).$$

On en déduit donc que :

- lorsque les variables sont centrées, $I(\mathbf{Y}) = \sum_{i=1}^{p} s_j^2$,

- lorsque les variables sont centrées-réduites, $I(\mathbf{Z}) = p$.

La dispersion du nuage des individus (l'inertie.)

$$\mathbb{I}_g = tr(\frac{1}{n}\sum_{i=1}^{n} y_i y_i^t M) = tr(\frac{1}{n} Y^t Y M) = tr(VM)$$

Remarque 1

- ***Si*** $M = I$ *inertie totale peut s'écrire :*

$$\mathbb{I}_g = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \overline{x}^j)^2 = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_i^j - \overline{x}^j) = \sum_{i=1}^{n}\sigma_j^2$$

L'inertie est donc égale à la somme de variance des variables étudées qui n'est autre que le trace de la matrice de variance $\mathcal{V}$.

- ***Si*** $M = D_{1/\sigma^2}$, *on a :*

$$\mathbb{I}_g = tr(M\mathcal{V}) = tr(D_{1/\sigma^2}\mathcal{V}) = tr(D_{1/\sigma}\mathcal{V}D_{1/\sigma})$$

Ce qui égale à

$$tr(R) = p$$

l'inertie est donc égale au nombre p de variables et ne dépand pas leurs valeurs .

La dispersion du nuage des individus (l'inertie.)

Exemple : Inertie du nuage des 6 patients

Données centrées ($\mathbf{Y}$) :

diast	syst	chol
15	10	0.3
-15	-45	0.2
0	5	0.4
-5	15	0.1
10	0	-0.3
-5	15	-0.7

Variance des colonnes :

diast	syst	chol
100.00	433.33	0.15

Données centrées-réduites ($\mathbf{Z}$)

diast	syst	chol
1.5	0.48	0.78
-1.5	-2.16	0.52
0.0	0.24	1.04
-0.5	0.72	0.26
1.0	0.00	-0.78
-0.5	0.72	-1.83

Variance des colonnes :

diast	syst	chol
1	1	1

Matrice des covariances :

	diast	syst	chol
diast	100.00	112.5	0.25
syst	112.50	433.3	-2.17
chol	0.25	-2.2	0.15

Matrice des corrélations

	diast	syst	chol
diast	1.000	0.54	0.065
syst	0.540	1.00	-0.272
chol	0.065	-0.27	1.000

- Inertie du nuage centré :

$$I(\mathbf{Y}) = 100 + 433.33 + 0.15$$

- Inertie du nuage centré-réduit :

$$I(\mathbf{Z}) = 1 + 1 + 1 = 3$$

nuage des variables.

Le second nuage de points associé à une matrice de données quantitatives est le nuage des variables.

Exemple : les variables tension arterielle diastolique, systolique et taux de cholestérol définissent un nuage de $p = 3$ points de $\mathbb{R}^6$.

```
##          ind1 ind2  ind3  ind4  ind5 ind6
## diast      90 60.0  75.0  70.0  85.0   70
## syst      140 85.0 135.0 145.0 130.0  145
## chol        6  5.9   6.1   5.8   5.4    5
```

Il n'est pas possible de visualiser ce nuage de points !

nuage des variables.

Dans le nuage des variables :

- chaque variable j est un point $\mathbf{x}^j$ de $\mathbb{R}^n$ (une colonne de $\mathbf{X}$),
- chaque variable j est pondérée par un poids m_j. En pratique :
  - $m_j = 1$ en ACP,
  - $m_j \neq 1$ en ACM (Analyse des Correspondances Multiples) par exemple.

Lorsque les données sont centrées :

- chaque variable j est un point $\mathbf{y}^j$de $\mathbb{R}^n$ (colonne de $\mathbf{Y}$),
- on parle du nuage des variables centrées.

Lorsque les données sont centrées-réduites :

- chaque variable j est un point $\mathbf{z}^j$ de $\mathbb{R}^n$ (colonne de $\mathbf{Z}$),
- on parle du nuage des variables centrées-réduites.

La liaison entre deux variables

La liaison entre deux variables se mesure avec la covariance ou la corrélation.

Pour définir la covariance et la corrélation, on munit $\mathbb{R}^n$ de la métrique :

$$\mathbf{N} = diag(\frac{1}{n}, \ldots, \frac{1}{n}).$$

- Le produit scalaire entre deux points $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^n$ est alors :

$$< \mathbf{x}, \mathbf{y} >_{\mathbf{N}} = \mathbf{x}^T \mathbf{N} \mathbf{y} = \frac{1}{n} \mathbf{x}^T \mathbf{y} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i.$$

- La norme d'un point $\mathbf{x}$ de $\mathbb{R}^n$ est :

$$\|\mathbf{x}\|_{\mathbf{N}} = \sqrt{< \mathbf{x}, \mathbf{x} >_{\mathbf{N}}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}.$$

La liaison entre deux variables

On en déduit que la variance s'écrit comme une norme (au carré) :

- $var(\mathbf{x}^j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}^j)^2 = \|\mathbf{y}^j\|_{\mathbf{N}}^2$,
- $var(\mathbf{z}^j) = \frac{1}{n}\sum_{i=1}^{n}(z_{ij} - \bar{z}^j)^2 = \|\mathbf{z}^j\|_{\mathbf{N}}^2$.

Le nuage des p variables centrées-réduites se trouve sur la boule unité de $\mathbb{R}^n$ avec $\|\mathbf{z}^j\|_{\mathbf{N}} = 1$.

On en déduit aussi que la covariance et la corrélation s'écrivent comme des produits scalaires :

- $c_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}^j)(x_{ij'} - \bar{x}^{j'}) = < \mathbf{y}^j, \mathbf{y}^{j'} >_{\mathbf{N}}$,
- $r_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_{ij} - \bar{x}^j}{s_j})(\frac{x_{ij'} - \bar{x}^{j'}}{s_{j'}}) = < \mathbf{z}^j, \mathbf{z}^{j'} >_{\mathbf{N}}$

la matrice S des covariances et de la matrice R des corrélations :

On en déduit une écriture matricielle de la matrice S des covariances et de la matrice R des corrélations :

- $S = Y^T N Y$,
- $R = Z^T N Z$.

Exemple :

Matrice des covariances :

```
##          diast  syst  chol
## diast   100.00 112.5  0.25
## syst    112.50 433.3 -2.17
## chol      0.25  -2.2  0.15
```

Matrice des corrélations

```
##        diast  syst   chol
## diast  1.000  0.54  0.065
## syst   0.540  1.00 -0.272
## chol   0.065 -0.27  1.000
```

Enfin on en déduit que la corrélation s'écrit comme un cosinus :

- $r_{jj'} = \frac{<\mathbf{y}^j, \mathbf{y}^{j'}>_N}{\|\mathbf{y}^j\|_N \|\mathbf{y}^{j'}\|_N} = \cos\theta_N(\mathbf{y}^j, \mathbf{y}^{j'})$,
- $r_{jj'} =< \mathbf{z}^j, \mathbf{z}^{j'} >_N = \cos\theta_N(\mathbf{z}^j, \mathbf{z}^{j'})$.

Cette propriété s'interprète de la manière suivante :

- un angle de 90 degré entre deux variables centrées-réduites correspond à une corrélation nulle entre les variables (cosinus égal à 0) et à l'absence de liaison linéaire,
- un angle de 0 degré entre deux variables centrées-réduites correspond à une corrélation de 1 entre les variables (cosinus égal à 1) et à l'existence d'une liaison linéaire positive,
- un angle de 180 degré entre deux variables centrées-réduites correspond à une corrélation de -1 entre les variables (cosinus égal à -1) à l'existence d'une liaison linéaire négative.

En ACP on peut analyser :

- la matrice des données centrées **Y**,
- la matrice des données centrées-réduites **Z**.

On distingue alors deux type d'ACP :

- l'ACP non normée (sur matrice des covariances) qui analyse **Y**,
- l'ACP normée (sur matrice des corrélations) qui analyse **Z**.

Dans la suite du cours, on se place dans le cadre de l'ACP normée.

Remarque *: On recommande **L'ACP centré** lorsque les variable sont homogène c'est-à-dire même signification , même unité de mesure , même ordre de grandeur. Dans le cas échéant on recommande **L'ACP centré réduit** lorsque les variable sont hétérogène.*

Résumé :

on muni l'espace

— $\mathbb{R}^p$ d'une metrique M de dimension $p \times p$.

— $\mathbb{R}^n$ d'une metrique N de dimension $n \times n$.

L'ACP va consister à analyser les *n points- individus (les lignes)* et les *p points-variables (les colonnes)* de la matrice des données centrés(Y) où centré réduite (Z) , avec les métriques

$$M_p = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = I_p sur \mathbb{R}^p$$

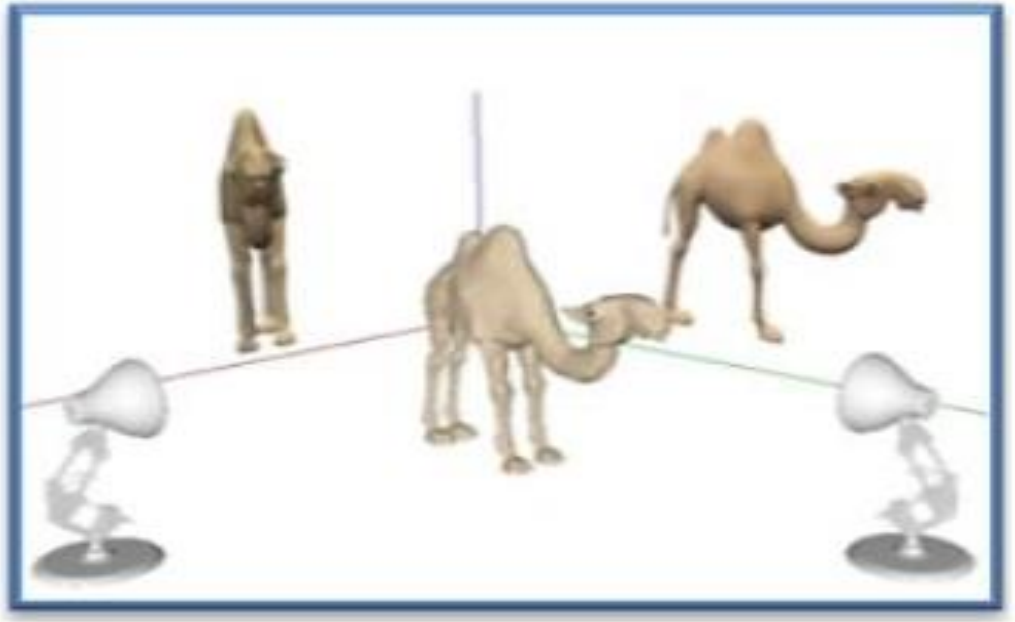
$$N = \begin{pmatrix} \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_n \end{pmatrix} \quad sur\mathbb{R}^p \quad \text{avec en général} \omega = \frac{1}{n} \Longrightarrow N = \frac{1}{n} \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = \frac{1}{n} I_n$$

Fait l'analyse avec la triplet $(\mathbf{Z}, \mathbf{I_p}, \mathbf{N}) \Longrightarrow \begin{cases} \text{ACP normé.,} \\ \text{ACP acec la matrice de corrélation.,} \end{cases}$

Fait l'analyse avec la triplet $(\mathbf{Y}, \mathbf{M_{1/\sigma}}, \mathbf{N}) \Longrightarrow \begin{cases} \text{ACP non normé.,} \\ \text{ACP avec la matrice de variance-covariance..,} \end{cases}$

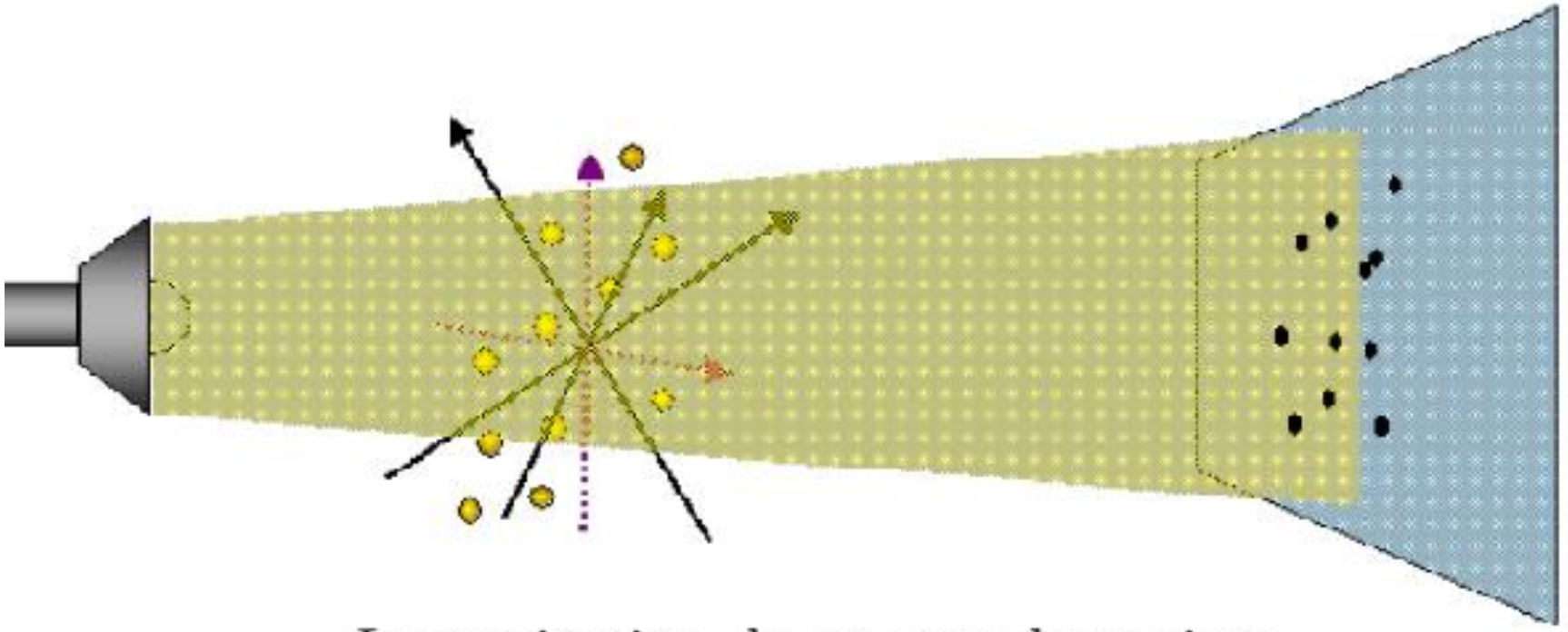
Analyse du nuage des individus

Trouver le sous-espace qui fournit la meilleure représentation des données.



- Meilleure approximation des données par projection.
- Meilleure représentation de la variabilité des données.

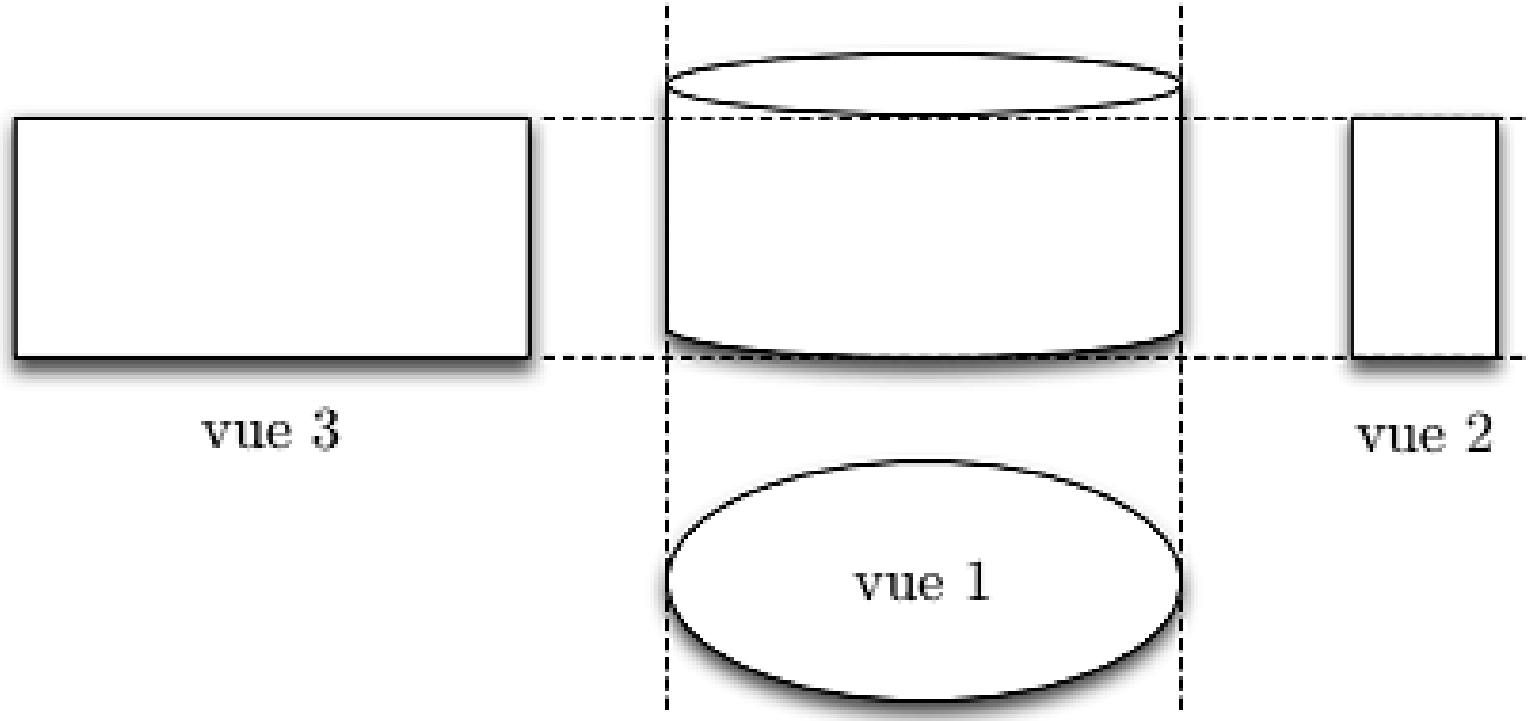
L'objectif est de trouver le plan de projection qui conserve le mieux possible les distances entre les individus (et donc la variabilité i.e. l'inertie du nuage de points).



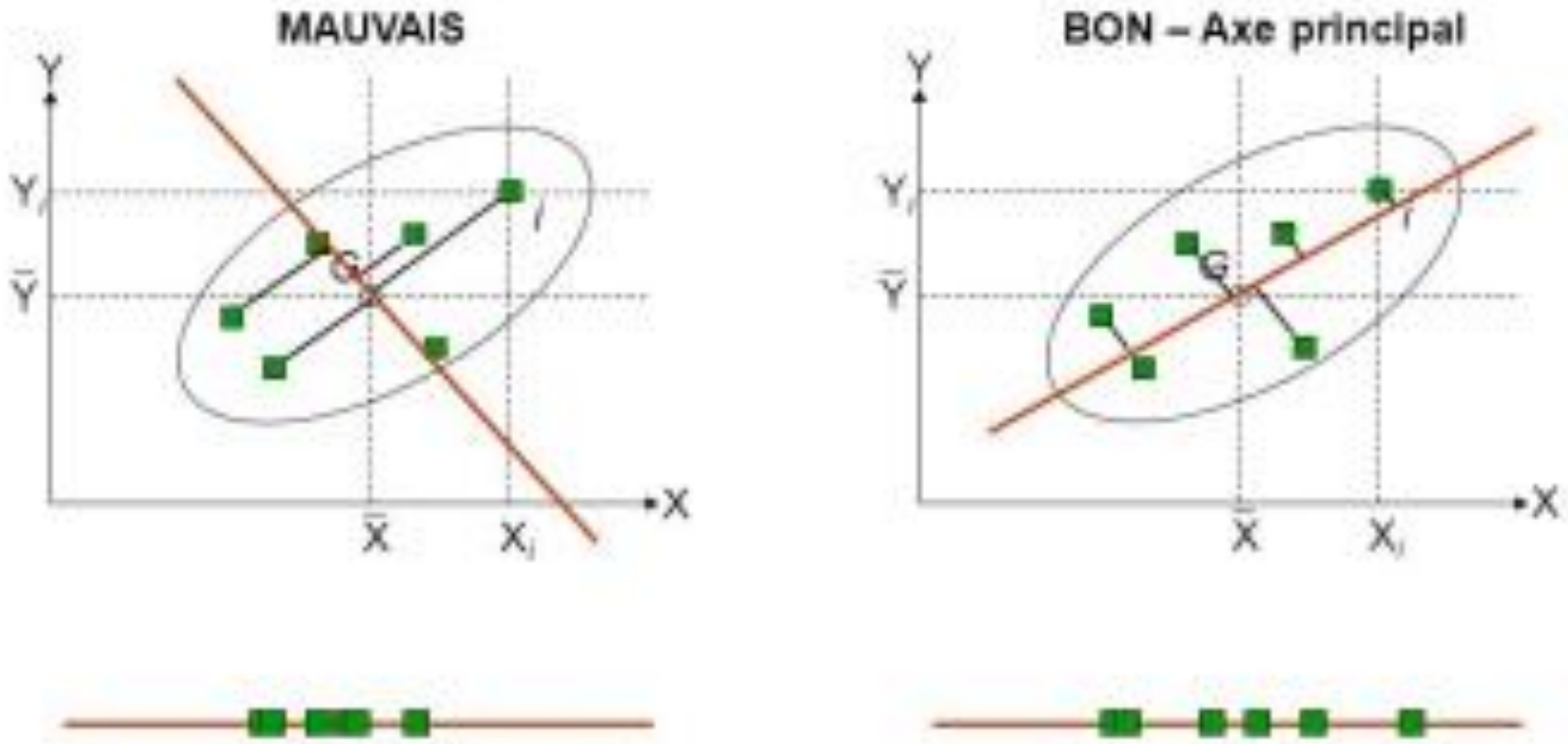
La projection du nuages des points

Analyse en composantes principales

Exemple



Projection d'un individu (un point de $\mathbb{R}^p$) sur un axe.



2 dimensions (nuage de points) → 1 dimension (1 axe)

Projection d'un individu (un point de $\mathbb{R}^p$) sur un axe.

La projection orthogonale d'un point $\mathbf{z}_i \in \mathbb{R}^p$ sur un axe Δ_α de vecteur directeur u_α ($u_\alpha^T u_\alpha = 1$) a pour coordonnée :

$$\Psi_{i\alpha} = < \mathbf{z}_i, u_\alpha > = \mathbf{z}_i^T u_\alpha,$$

et le vecteur des coordonnées de projections des n individus est :

$$\Psi^\alpha = \begin{pmatrix} \Psi_{1\alpha} \\ \vdots \\ \Psi_{n\alpha} \end{pmatrix} = \mathbf{Z} u_\alpha = \sum_{j=1}^{p} u_{j\alpha} \mathbf{z}^j.$$

- Ψ^α est une combinaison linéaire des colonnes de $\mathbf{Z}$.
- Ψ^α est centré si les colonnes de $\mathbf{Z}$ sont centrées.

En ACP les vecteurs directeurs u_1 et u_2 sont définis pour maximiser l'inertie du nuage des individus projeté et conserver ainsi au mieux les distances entre les individus.

Axes de projection des individus en ACP.

Δ_1 est l'axe de vecteur directeur $u_1 \in \mathbb{R}^p$ qui maximise la variance des n individus projetés :

$$u_1 = \arg\max_{\|u\|=1} Var(\mathbf{Z}\ ) \quad = \arg\max_{\|u\|=1} u^T \mathbf{R}\, u$$

où $\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ est la matrice des corrélations entre les p variables.

On peut montrer que :

- u_1 est le vecteur propre associé à la première valeur propre λ_1 de $\mathbf{R}$,
- La première composante principale $\Psi^1 = Zu_1$ est centrée :

$$\overline{\Psi^1} = 0,$$

- λ_1 est la variance la première composante principale :

$$Var(\Psi^1) = \lambda_1.$$

Δ_2 est l'axe de vecteur directeur $u_2 \perp u_1$ qui maximise la variance des n individus projetés :

$$u_2 = \arg \max_{\|u\|=1, u_2 \perp u_1} Var(\mathbf{Z}\, u\ ).$$

On peut montrer que :

- u_2 est le vecteur propre associé à la seconde valeur propre λ_2 de $\mathbf{R}$,
- La seconde composante principale $\Psi^2 = Z\, u_2$ est centrée :

$$\overline{\Psi^2} = 0,$$

- λ_2 est la variance la seconde composante principale :

$$Var(\ \Psi^2) = \lambda_2,$$

- Les composantes principales Ψ^1 et Ψ^2 ne sont pas corrélées.

On obtient ainsi $q \leq r$ (r est le rang de $\mathbf{Z}$) axes orthogonaux $\Delta_1, \ldots, \Delta_q$ sur lesquels on projette le nuage des individus.

Matrice des données centrées réduites

$$\underbrace{\left(\begin{array}{c|ccccc} & 1 & \cdots & j & \cdots p \\ \hline 1 & & & & \\ \vdots & & & \vdots & \\ \hline i & & \cdots & z_{ij} & \cdots \\ \hline \vdots & & & \vdots & \\ n & & & & \end{array}\right)}_{z_i^t \in \mathbb{R}} \Longrightarrow^{\text{projection sur} F_k} \Psi =$$

Matrice des coordonnées factorielles des individus

$$\underbrace{\left(\begin{array}{c|ccccc} & 1 & \cdots & \alpha & \cdots k & \\ \hline 1 & & & & & \\ \vdots & & & \vdots & & \\ \hline i & & \cdots & \psi_{i\alpha} & \cdots & \psi_i^t \\ \hline \vdots & & & \vdots & & \\ n & & & & & \end{array}\right)}_{\psi^\alpha \in \mathbb{R}}$$

En résumé :

1. On effectue la décomposition en valeurs propres de la matrice des corrélations $\mathbf{R}$ et on choisit q.
2. On calcule la matrice $= \mathbf{ZU}$ des q composante principales à partir de la matrice $\mathbf{V}$ des q premiers vecteurs propres de $\mathbf{R}$.
   - Les composantes principales $\Psi^{\alpha} = \mathbf{Z}\mathbf{u}_{\alpha}$ (colonnes de Ψ) sont centrées et variances λ_{α}.
   - Les éléments $\Psi_{i\alpha}$ sont appelés les coordonnées factorielles des individus ou encore les scores des individus sur les composantes principales.

$\Psi :=$

	$1 \ldots$	α	$\ldots q$
1			
$\vdots$		$\vdots$	
i	$\ldots$	$\Psi_{i\alpha}$	$\ldots$
$\vdots$		$\vdots$	
n			
moy	$\ldots$	0	$\ldots$
var	$\ldots$	λ_{α}	$\ldots$

Exemple

soit les matrices

$$X = \begin{pmatrix} 3 & 4 \\ 7 & 2 \\ 5 & 6 \\ 5 & 4 \end{pmatrix}, \qquad Z = \begin{pmatrix} -\sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \qquad R = \frac{1}{4}{}^t ZZ = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

1. *Les deux valeurs propres sont* $\lambda_1 = 3/2$ *et* $\lambda_2 = 1/2$.

2. *Les vecteurs propres unitaire associes à* λ_1 *et* λ_2 *sont* $u_1 = \begin{pmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix}$, $u_2 = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$

3. *Les nouvelles coordonnées des individus (les composantes principales) :*

$$\Psi = [\psi_1, \psi_2] = Z.U = \begin{pmatrix} -\sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 2 & 0 \\ -1 & 1 \\ 0 & 0 \end{pmatrix}$$

Exemple

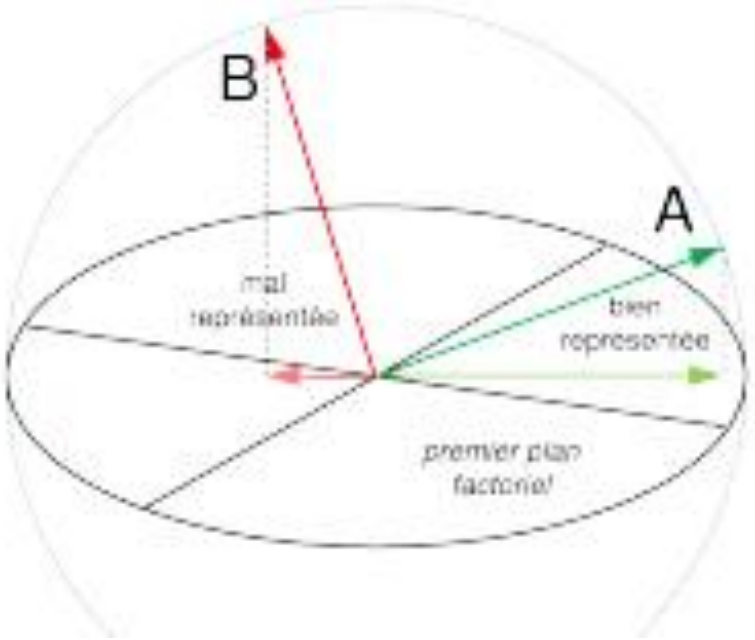
la matrice de covariance notée Σ ? et de de corrélation de X :

Donc $X \sim \mathcal{N}(0, \Sigma)$ *, avec* $\Sigma = \begin{pmatrix} \beta^2 + \sigma^2 & -\beta^2 \\ -\beta^2 & \beta^2 + \sigma^2 \end{pmatrix}$, $\mathcal{R} = \begin{pmatrix} 1 & \frac{-\beta^2}{\beta^2+\sigma^2} \\ \frac{-\beta^2}{\beta^2+\sigma^2} & 1 \end{pmatrix}$

1. *Calculer les valeurs propres* $\lambda_1 \geq \lambda_2$ *de la matrice de covariance de* X*. ?*

Analyse du nuage des variables

Trouver le sous-espace qui fournit la meilleure représentation des variables.

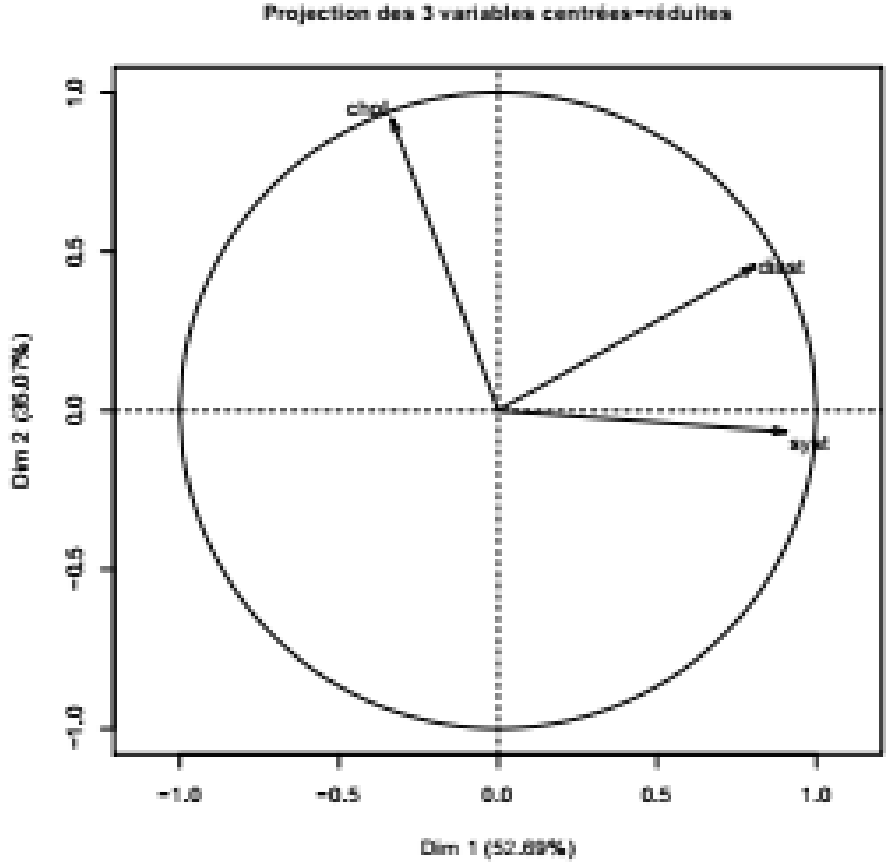


Analyse du nuage des variables

Exemple des 6 patients décrits sur 3 variables centées-réduites.

3 variables sur la boule unité de $\mathbb{R}^6$.

```
##        Brigitte Marie Vincent Alex Manue Fred
## diast       1.5  -1.5     0.0 -0.5   1.0 -0.5
## syst        0.5  -2.2     0.2  0.7   0.0  0.7
## chol        0.8   0.5     1.0  0.3  -0.8 -1.8
```



L'objectif est de trouver le plan de projection qui permet de représenter au mieux les variable et ainsi conserver le mieux possible les angles entre les variables (et donc leur corrélations).

Projection d'une variable (un point de $\mathbb{R}^n$) sur un axe.

La projection N-orthogonale d'une variable $\mathbf{z}^j \in \mathbb{R}^n$ sur un axe G_α de vecteur directeur $\mathbf{v}_\alpha$ ($\mathbf{v}_\alpha^T \mathbf{N} \mathbf{v}_\alpha = 1$) a pour coordonnée :

$$\Phi_{j\alpha} = < \mathbf{z}^j, \mathbf{v}_\alpha >_{\mathbf{N}} = (\mathbf{z}^j)^T \mathbf{N} \mathbf{v}_\alpha,$$

et le vecteur des coordonnées des projections des p variables est :

$$\Phi^\alpha = \begin{pmatrix} \Phi_{1\alpha} \\ \vdots \\ \Phi_{p\alpha} \end{pmatrix} = \mathbf{Z}^T \mathbf{N} \mathbf{v}_\alpha$$

Ici on a muni $\mathbb{R}^n$ d'une métrique $\mathbf{N}$:

- Dans le cadre général $\mathbf{N}$ une matrice $n \times n$ symétrique définie positive,
- Dans le cas particulier de l'ACP, $\mathbf{N}$ est la matrice diagonale des poids des individus :
$$\mathbf{N} = diag(w_1, \ldots, w_n).$$
- Dans le cas particulier où tous les individus sont pondérés par $\frac{1}{n}$
$$\mathbf{N} = \frac{1}{n} \mathbb{I}_n.$$

Exemple : les 3 variables (diast, syst, chol) sont les colonnes la matrice des données centrées-réduites

$$\mathbf{Z} = \begin{pmatrix} 1.50 & 0.48 & 0.78 \\ -1.50 & -2.16 & 0.52 \\ 0.00 & 0.24 & 1.04 \\ -0.50 & 0.72 & 0.26 \\ 1.00 & 0.00 & -0.78 \\ -0.50 & 0.72 & -1.83 \end{pmatrix}$$

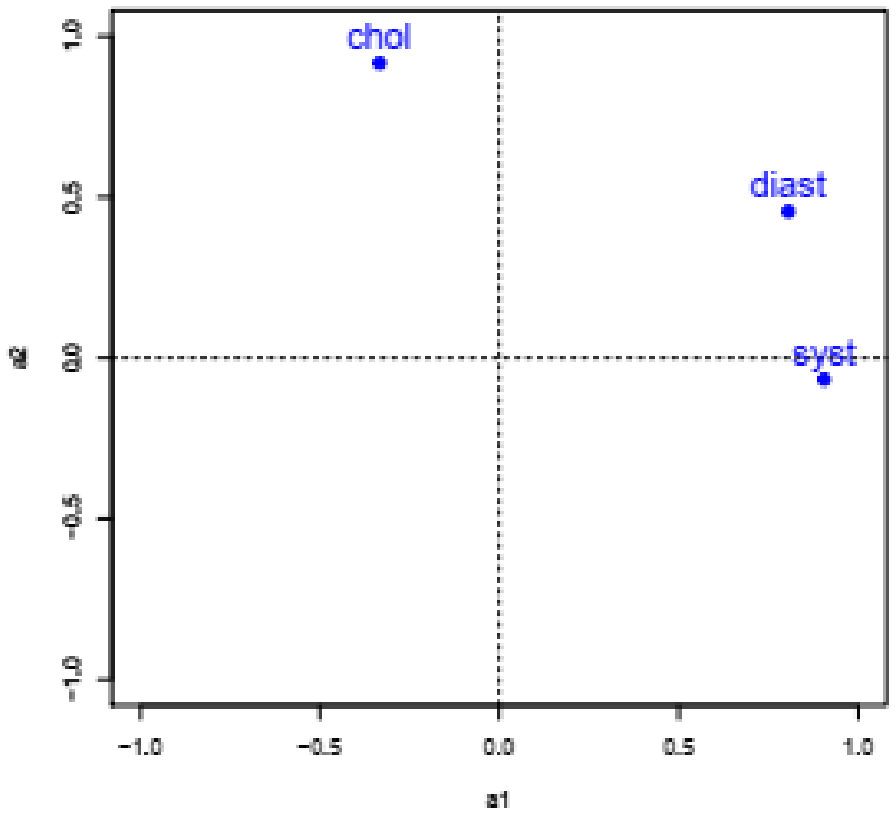
On veut projeter les 3 variabes sur deux axes N-orthogonaux G_1 et G_2 de vecteurs directeurs (ici $\mathbf{N} = \frac{1}{6}\mathbb{I}_6$) :

$$\mathbf{v}_1 = \begin{pmatrix} 0.87 \\ -2.11 \\ -0.08 \\ 0.10 \\ 0.67 \\ 0.54 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1.30 \\ -0.06 \\ 0.90 \\ -0.03 \\ -0.25 \\ -1.8 \end{pmatrix}.$$

Les vecteurs Φ^1 et Φ^2 des coordonnées des projections des 3 variables sur G_1 et G_2 sont :

$$\Phi^1 = \mathbf{Z}^T\mathbf{N}v_1 = \frac{0.87}{6}\begin{pmatrix}1.5\\0.48\\0.78\end{pmatrix}\frac{-2.11}{6}\begin{pmatrix}-1.5\\-2.16\\0.52\end{pmatrix} + \ldots\frac{+0.54}{6}\begin{pmatrix}-0.5\\0.72\\-1.83\end{pmatrix} = \begin{pmatrix}0.81\\0.91\\-0.33\end{pmatrix}$$

$$\Phi^2 = \mathbf{Z}^T\mathbf{N}v_2 = \frac{1.30}{6}\begin{pmatrix}1.5\\0.48\\0.78\end{pmatrix}\frac{-0.06}{6}\begin{pmatrix}-1.5\\-2.16\\0.52\end{pmatrix} + \ldots\frac{-1.80}{6}\begin{pmatrix}-0.5\\0.72\\-1.83\end{pmatrix} = \begin{pmatrix}0.45\\-0.07\\0.92\end{pmatrix}$$



En ACP les vecteurs directeurs $\mathbf{u}_1$ et $\mathbf{u}_2$ sont définis pour maximiser la somme des cosinus (au carré) des angles entre les variables et les axes de projection.

Axes de projection des variables en ACP.

G_1 est l'axe de vecteur directeur $\mathbf{v}_1 \in \mathbb{R}^n$ qui maximise la somme des carrés des cosinus des angles avec les variables.

$$\mathbf{v}_1 = \arg\max_{\|\mathbf{v}\|_\mathbf{N}=1} \sum_{j=1}^{p} \cos^2 \theta_\mathbf{N}(\mathbf{z}^j, \mathbf{v}) \quad = \arg\max_{\|\mathbf{v}\|_\mathbf{N}=1} \|\mathbf{Z}^T\mathbf{N}\mathbf{v}\|^2$$

On peut montrer qu'avec $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$:

- $\mathbf{v}_1$ est le vecteur propre associé à la plus grand valeur propre de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$,
- la plus grand valeur propre de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ est aussi la première valeur propre λ_1 de $\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$,
- λ_1 est la somme des carrés des cosinus entre les variables et $\mathbf{v}_1$:

$$\lambda_1 = \sum_{j=1}^{p} cos^2\theta_\mathbf{N}(\mathbf{z}^j, \mathbf{v}_1)$$

Axes de projection des variables en ACP.

G_2 est l'axe de vecteur directeur $\mathbf{v}_2 \perp_{\mathbf{N}} \mathbf{v}_1$ qui maximise la somme des carrés des cosinus des angles avec les variables :

$$\mathbf{v}_2 = \arg \max_{\|\mathbf{v}\|_{\mathbf{N}}=1, \mathbf{v}_2 \perp_{\mathbf{N}} \mathbf{v}_1} \sum_{j=1}^{p} \cos^2 \theta_{\mathbf{N}}(\mathbf{z}^j, \mathbf{v})$$

On peut montrer que :

- $\mathbf{v}_2$ est le vecteur propre associé à la seconde plus grand valeur propre de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$.
- la seconde plus grande valeur propre est aussi la seconde valeur propre λ_2 de $\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$,
- λ_2 est la somme des carrés des cosinus entre les variables et $\mathbf{v}_2$:

$$\lambda_2 = \sum_{j=1}^{p} \cos^2 \theta_{\mathbf{N}}(\mathbf{z}^j, \mathbf{v}_2)$$

On obtient ainsi $q \leq r$ (r est le rang de $\mathbf{Z}$) axes orthogonaux $G_1, \ldots, G_q$ sur lesquels on projette le nuage des variables.

En résumé :

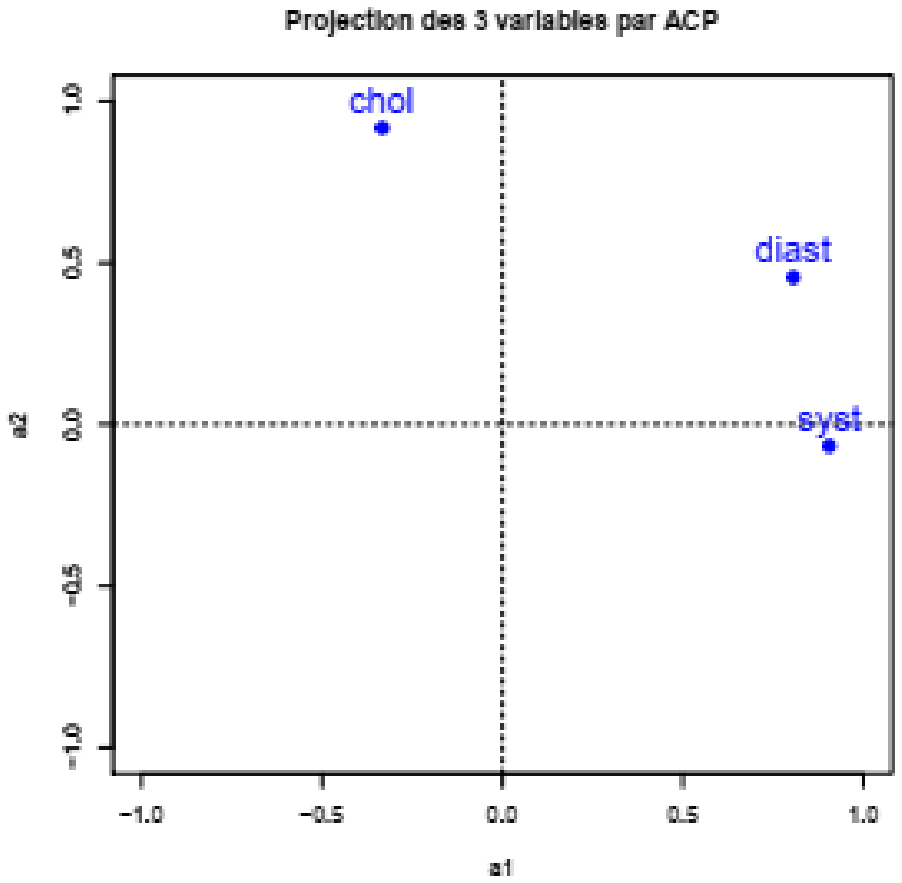
1. On effectue la décomposition en valeurs propres de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ et on choisit q.
2. On calcule la matrice $\Phi = \mathbf{Z}^T\mathbf{N}\mathbf{V}$ à partir de la matrice $\mathbf{V}$ des q premiers vecteurs propres de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$.
   - Les colonnes $\Phi^\alpha = \mathbf{Z}^T\mathbf{N}\mathbf{v}_\alpha$ de la matrice Φ contiennent les coordonnées des projections des variables sur l'axe G_α.
   - Les éléments $\Phi_{j\alpha}$ sont appelés les coordonnées factorielles des variables ou encore les loadings des variables.

$\Phi =$

	$1 \ldots$	α	$\ldots q$
1			
$\vdots$		$\vdots$	
i	$\ldots$	$\Phi_{i\alpha}$	$\ldots$
$\vdots$		$\vdots$	
p			
norme	$\ldots$	$\sqrt{\lambda_\alpha}$	$\ldots$

Exemple des 6 patients : matrice **A** pour $q = 2$.

```
##           a1     a2
## diast   0.81  0.455
## syst    0.91 -0.067
## chol   -0.33  0.917
```



Formules de passage.

On peut montrer que

- les composantes principales s'obtiennent aussi à partir de la décomposition en valeurs propres de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$:

$$\Psi^{\alpha} = \mathbf{Z}\,\mathbf{u}_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha\,,$$

- les loadings s'obtiennent aussi à partir à partir de la décomposition en valeurs propres de $\frac{1}{n}\mathbf{Z}^T\mathbf{Z}$:

$$\Phi^{\alpha} = \mathbf{Z}^T\mathbf{N}\,\mathbf{v}_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{u}_\alpha$$

On en déduit que :

$$\Psi = \mathbf{V}\Lambda$$
$$\Phi = \mathbf{U}\Lambda$$

où $\Lambda = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_q})$

On en déduit aussi que

- Les vecteurs propres $\mathbf{v}_\alpha$ de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ sont les composantes principales standardisées (divisées par leur écart-type) :

$$\mathbf{v}_\alpha = \frac{\Psi^\alpha}{\sqrt{\lambda_\alpha}},$$

- Les loadings sont les corrélations entre les variables et les composantes principales :

$$\Phi_{j\alpha} = cor(\mathbf{x}^j, \Psi^\alpha).$$

Cette relation est en rouge car elle est fondamentale pour l'interprétation des résultats en ACP.

Interprétation des résultats

- **Les composantes (Scores) des individus**
  Les scores des individus sont les valeurs des composantes principales sur les individus ,
- ***Contributions des individus***
  La contribution relative d'un individu i à la formation de la composante principale α est l'inertie relative de cet individu sur l 'axe factoriel k .
- ***Qualités de la représentation des individus***
  La qualité de la représentation d'un individu i par la composante principale α ,

Résultats relatifs aux variables

- ***Composantes (Saturation) des variables***
  Les saturations des variables sont les coordonnées factorielles des variables. Elles sont égales au coefficients de corrélation entre les variables (centrées réduites) de départ et les scores des individus :$\phi j\alpha = \rho(xj, \psi\alpha)$
- ***Contributions des variables*** Les contributions des variables à la formation des composantes principales sont définies de la même façon que celles des individus.
- ***Qualités de la représentation des variables***
  La qualité de la représentation d'une variable par une composante principale est définie de la même façon que pour les individus .

Interprétation des résultats

Variance des composantes principales.

Les composantes principales (colonnes de Φ) sont q nouvelles variables synthétiques non corrélées et de variance maximale avec

$$Var(\Phi^{\alpha}) = \lambda_{\alpha}$$

On en déduit que l'inertie des individus décrits par les q premières composantes principales vaut :

$$I(\Phi) = \lambda_1 + \ldots + \lambda_q.$$

Exemple des 6 patients :

Les $p = 3$ valeurs propres non nulles de la matrice des corrélations **R** sont :

```
##         eigenvalue
## lambda1       1.58
## lambda2       1.05
## lambda3       0.37
```

donc

$$\begin{aligned} Var(\Phi^1) &= 1.58 \\ Var(\Phi^2) &= 0.05 \end{aligned}$$

et l'inertie des individus décrits par $q = 2$ composantes principales est :

$$I(\Phi) = \lambda_1 + \lambda_2 = 1.58 + 1.05 = 2.63.$$

L'inertie

L'inertie est une notion fondamentale en ACP, puis qu'elle est une mesure de la dispersion du nuage des points autour de son centre de gravité g.

- ***Si*** $M = I$ *inertie totale peut s'écrire :*

$$\mathbb{I}_g = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \overline{x}^j)^2 = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_i^j - \overline{x}^j) = \sum_{i=1}^{n}\sigma_j^2$$

L'inertie est donc égale à la somme de variance des variables étudées qui n'est autre que le trace de la matrice de variance $\mathcal{V}$.

- ***Si*** $M = D_{1/\sigma^2}$, *on a :*

$$\mathbb{I}_g = tr(M\mathcal{V}) = tr(D_{1/\sigma^2}\mathcal{V}) = tr(D_{1/\sigma}\mathcal{V}D_{1/\sigma})$$

Ce qui égale à

$$tr(R) = p$$

l'inertie est donc égale au nombre p *de variables et ne dépand pas leurs valeurs .*

Inertie totale.

L'inertie totale en ACP normée est l'inertie des individus décrits par les p variables centrées-réduites (colonnes de $\mathbf{Z}$) :

$$I(\mathbf{Z}) = \sum_{j=1}^{p} Var(z^j) = p.$$

Lorsque $q = r$ l'inertie des individus décrits par toutes les composantes principales est égale à l'inertie totale :

$$I(\Psi) = \lambda_1 + \ldots + \lambda_r = I(\mathbf{Z}) = p$$

Exemple des 6 patients :

L'inertie des individus décrits par $q = 3$ (toutes) composantes principales est :

$$I(\Psi) = \lambda_1 + \lambda_2 + \lambda_3 = 1.58 + 1.05 + 0.37 = 3$$

Inertie totale. ACP non normée

L'inertie totale en ACP non normée est l'inertie des individus décrits par les p variables centrées (colonnes de Y) :

$$I(Y) = \sum_{j=1}^{p} Var(y^j) = \sum_{j=1}^{p} \sigma_j^2$$

L'inertie associées aux axes

1. ***Pourcentage de l'inertie du nuage de points initiale expliqué par l'axe Δ_α***

   (a) - *Le cas d'un* ***ACP normé***

   $$\Longrightarrow \frac{\textit{L'inertie (variance) du nuage des inds projeté sur } \Delta_\alpha = I(\psi^\alpha)}{\textit{L'inertie de nuage initiale}} \times 100$$

   $$= \frac{\lambda_\alpha}{\lambda_\alpha + ... + \lambda_p} = \frac{\lambda_\alpha}{p}$$

   (b) *Le cas d'un* ***ACP non normé***

   $$\Longrightarrow \frac{\textit{L'inertie (variance) du nuage des inds projeté sur } \Delta_\alpha = I(\psi^\alpha)}{\textit{L'inertie de nuage initiale}} \times 100 = \frac{\lambda_\alpha}{tr(VM)}$$

2. ***Pourcentage de l'inertie du nuage de points initiale expliqué par les k premiers axes***

   ce pourcentage vaut $\frac{\lambda_1 + ... + \lambda_\alpha}{\lambda_1 + ... + \lambda_p} \times 100$

L'inertie associées aux axes

par exemple le poucentage d'inertie des 03 variables.

	λ_α	% *d'inetie*	% *inertie cumulé*
*Axe*1	1.73	57.7	57.7
*Axe*2	1.06	35.4	93.3
*Axe*3	0.20	6.70	100

Choix de dimension

1. Part d'inertie

La "qualité globale" des représentations est mesurée par la part d'inertie expliquée (pourcentage expliqué par l'axe Δ_α :

$$I_{\Delta_\alpha} = \frac{\sum_{\alpha=1}^{k} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha} \times 100$$

La valeur de k est choisie de sorte que cette part d'inertie expliquée I_{Δ_α} soit supérieure à une valeur seuil fixée a priori par l'utilisateur (par exemple : choisir k axes pour avoir plus de 80% d'inertie expliquée). C'est souvent le seul critère employé. par exemple :

	λ_α	% d'inetie	% inertie cumulé
$Axe1$	1.73	57.7	57.7
$Axe2$	1.06	35.4	93.3
$Axe3$	0.20	6.70	100

On va choisit l'axe(1 et 2) pour la représentation (>80%).

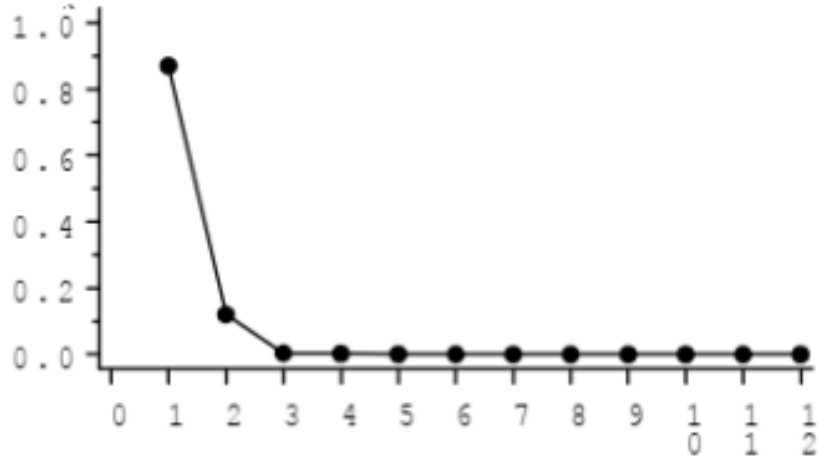
Choix de dimension

2. **Règle de Kaiser**

   le critère de Kaiser (1961), qui stipule de ne retenir que les valeurs propres supérieures à la moyenne des valeurs propres (c'est-à-dire à 1 dans le cas d'une analyse en composantes principales sur matrices de corrélation),

3. **Éboulis**

   C'est le graphique 1.9 présentant la décroissance des valeurs propres.
   Le principe consiste à rechercher, s'il existe, un **"coude"** (changement de signe dans la suite des différences d'ordre 2) dans le graphe et de ne conserver que les valeurs propres jusqu'à ce coude.



Exemple des 6 patients.

Données brutes ($p = 3$ et n=6)

	diast	syst	chol
##	90	140	6.0
##	60	85	5.9
##	75	135	6.1
##	70	145	5.8
##	85	130	5.4
##	70	145	5.0

Réduction aux deux premières CP

	ψ_1	ψ_2
##	1.10	1.334
##	-2.66	-0.057
##	-0.10	0.918
##	0.13	-0.035
##	0.85	-0.257
##	0.68	-1.903

Quelle est la qualité de cette réduction ?

```
res$eig
```

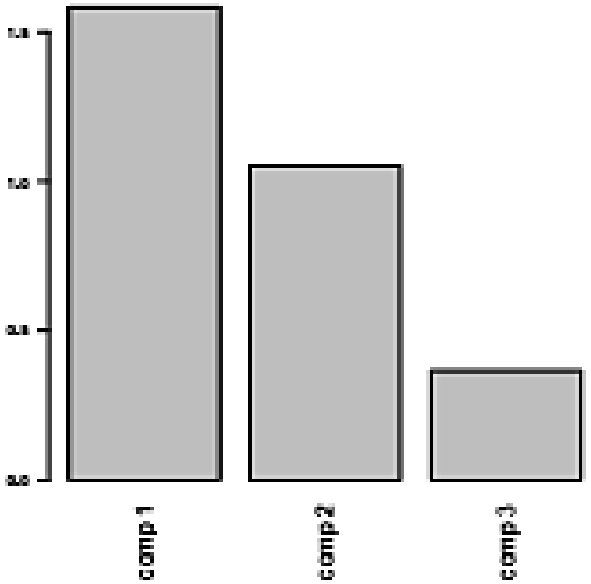
```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1       1.58                     53                                 53
## comp 2       1.05                     35                                 88
## comp 3       0.37                     12                                100
```

- $r = 3$ valeurs propres non nulles car $r = \min(n - 1, p) = 3$,
- La somme des valeurs propres vaut $p = 3$ (l'inertie totale),
- 53 % de l'inertie est expliquée par la première CP.
- 88 % de l'inertie est expliquée par les deux premières CP.
- 100 % de l'inertie est expliquée par toutes les CP.

Exemple des 6 patients.

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1       1.58                     53                                53
## comp 2       1.05                     35                                88
## comp 3       0.37                     12                               100
```

Ebouli des valeurs propres



- 88% d'inertie expliquée avec $q = 2$ composantes.
- Règle de Kaiser : deux valeurs propres plus grandes que 1.
- Règle du coude : "cassure" après 2 composantes.

On choisit de retenir $q = 2$ composantes principales pour résumer les données décrites sur $p = 3$ variables.

Ainsi on ne pert que 12% de l'information (l'inertie) de départ.

Interprétation des plans factoriels des individus.

Si deux individus sont bien projetés, alors leur distance en projection est proche de leur distance dans $\mathbb{R}^p$.

- On mesure la qualité de la projection d'un individu i sur l'axe Δ_α par le carré du cosinus de l'angle $\theta_{i\alpha}$ entre le vecteur $\mathbf{z}_i$ et l'axe Δ_α :

$$\cos^2(\theta_{i\alpha}) = \frac{\Psi_{i\alpha}^2}{\|\mathbf{z}_i\|^2}$$

- On mesure la qualité de la projection d'un individu i sur le plan $(\Delta_\alpha, \Delta_{\alpha'})$ par le carré du cosinus de l'angle $\theta_{i(\alpha,\alpha')}$ entre le vecteur $\mathbf{z}_i$ et le plan $(\Delta_\alpha, \Delta_{\alpha'})$:

$$\cos^2(\theta_{i(\alpha,\alpha')}) = \frac{\Psi_{i\alpha}^2 + \Psi_{i\alpha'}^2}{\|\mathbf{z}_i\|^2}$$

Plus la valeur du $\cos^2$ est proche de 1, meilleure est la qualité de la représentation de l'individu.

La contribution relative d'un individu

Les individus qui contribuent de manière excessive à l'inertie des individus projetés sont source d'instabilité.

- L'inertie (la variance) sur l'axe Δ_α est $\lambda_\alpha = \sum_{i=1}^{n} w_i \, \Psi_{i\alpha}^2$

  avec souvent $w_i = \frac{1}{n}$.

- La contribution relative d'un individu i à l'inertie de l'axe Δ_α est

$$Ctr(i,\alpha) = \frac{w_i \, \Psi_{i\alpha}^2}{\lambda_\alpha}.$$

- La contribution relative d'un individu i à l'inertie du plan $(\Delta_\alpha, \Delta'_\alpha)$ est

$$Ctr(i,(\alpha,\alpha')) = \frac{w_i \, \Psi_{i\alpha}^2 + w_i \, \Psi_{i\alpha'}^2}{\lambda_\alpha + \lambda_{\alpha'}}.$$

Si les poids w_i des individus sont tous identiques ($w_i = \frac{1}{n}$ par exemple), les individus excentrés sont ceux qui contribuent le plus.

Exemple

Interprétation des cercles de corrélations des variables.

Si deux variables sont bien projetées, alors leur angle en projection est proche de leur angle dans $\mathbb{R}^n$ et la la corrélation entre ces deux variables est proche du cosinus de l'angle entre leurs projections.

- On mesure la qualité de la projection d'une variable j sur l'axe G_α par le carré du cosinus de l'angle $\theta_{j\alpha}$ entre le vecteur $\mathbf{z}^j$ et l'axe G_α :

$$\cos^2(\theta_{j\alpha}) = \frac{\Phi_{j\alpha}^2}{\|\mathbf{z}^j\|^2} = \Phi_{j\alpha}^2$$

- On mesure la qualité de la projection d'une variable j sur le plan $(G_\alpha, G_{\alpha'})$ par le carré du cosinus de l'angle $\theta_{j(\alpha,\alpha')}$ entre le vecteur $\mathbf{z}^j$ et le plan $(G_\alpha, G_{\alpha'})$:

$$\cos^2(\theta_{j(\alpha,\alpha')}) = \Phi_{j\alpha}^2 + \Phi_{j\alpha'}^2.$$

$\sqrt{\cos^2(\theta_{j(\alpha,\alpha')})}$ est donc la "longueur de la flèche".

Plus la flèche est proche du cercle, meilleure est la qualité de la représentation de la variable.

cercles de corrélations des variables.

Exemple des 3 variables décrivant les patients.

Coordonnées factorielles des variables

```
##          Φ1     Φ2
## diast  0.81  0.455
## syst   0.91 -0.067
## chol  -0.33  0.917
```

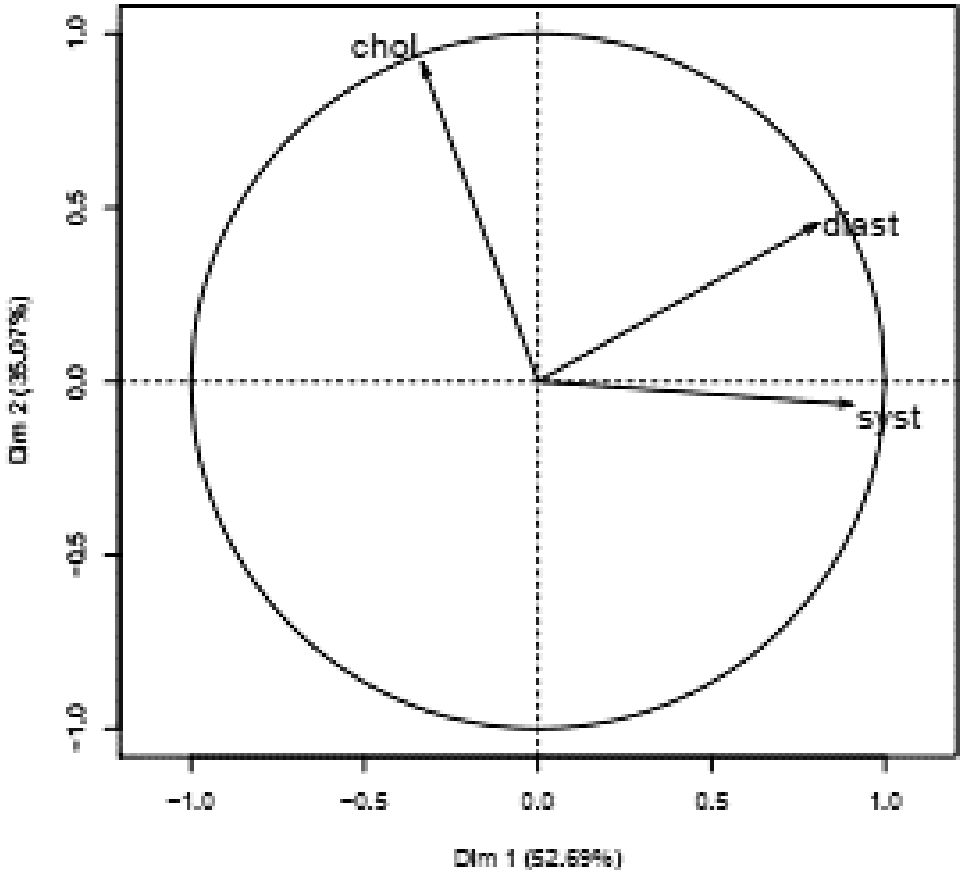
cos^2 des variables sur les axes

```
##           G1     G2
## diast 0.65 0.2068
## syst  0.82 0.0045
## chol  0.11 0.8410
```

Le cos^2 de la variable diast sur G_1 est 0.65 et le cos^2 de la variable diast sur (G_1, G_2) est $0.65 + 0.2068 = 0.8568$.

La variable diast est donc bien représentée sur ce plan et la longueur de sa flèche dans le cercle des corrélation sera donc de $\sqrt{0.8568} = 0.92563$.

cercles de corrélations des variables.



- Les variables sont-elles globalement bien projetées sur ce plan ?
- Interpréter ce cercle des corrélations.

Les contributions des variables

Les contributions des variables aux axes permettent de donner une interprétation aux axes.

- La qualité de l'axe G_α est $\lambda_\alpha = \sum_{j=1}^{p} \Phi_{j\alpha}^2 = \sum_{j=1}^{p} cos^2\theta_{j\alpha}$.
- La contribution relative d'une variable j à l'axe G_α est

$$Ctr(j,\alpha) = \frac{\Phi_{j\alpha}^2}{\lambda_\alpha}.$$

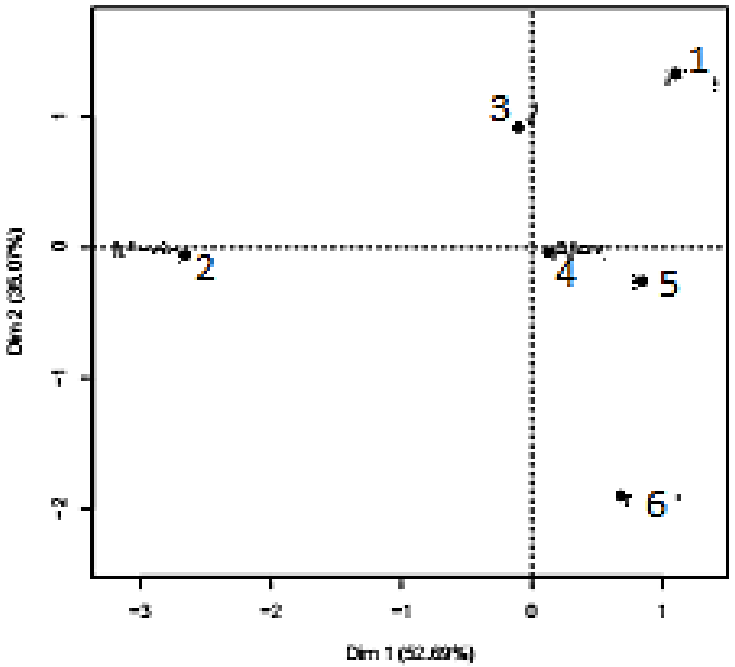
- La contribution relative d'une variable j au plan (G_α, G'_α) est

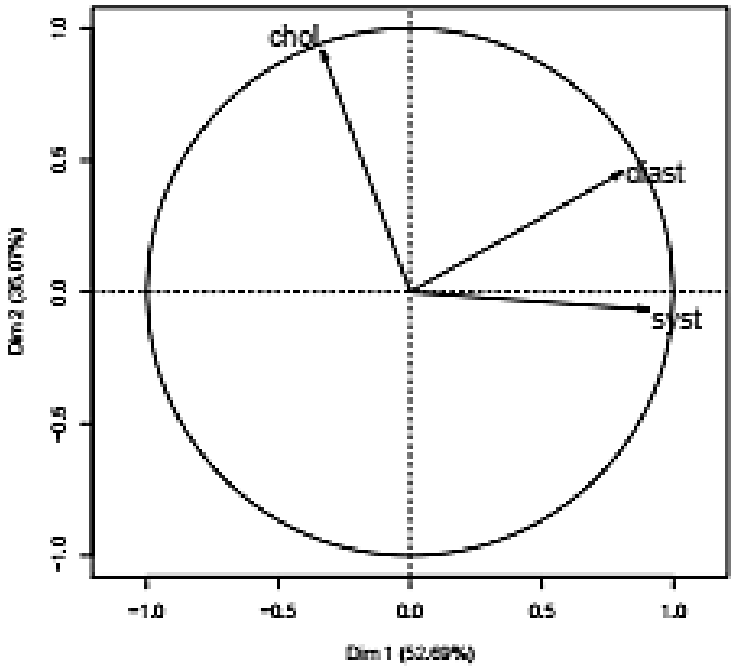
$$Ctr(j,(\alpha,\alpha')) = \frac{\Phi_{j\alpha}^2 + \Phi_{j\alpha'}^2}{\lambda_\alpha + \lambda_{\alpha'}}.$$

Interprétation du plan factoriel des individus à partir du cercle des corrélations.

$$\Phi_{j\alpha} = \mathbf{corr}(x^j, \Psi^\alpha)$$

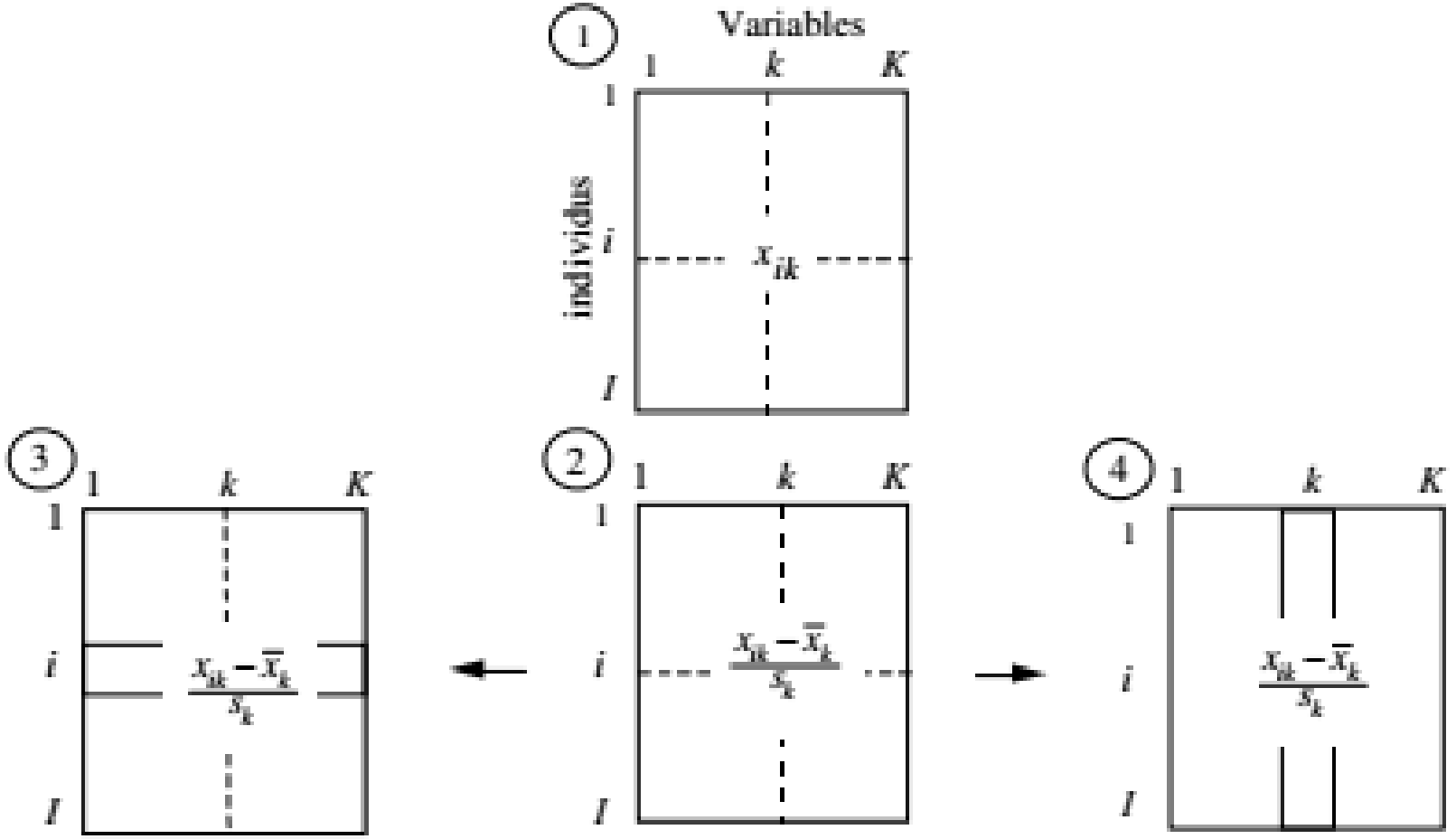
```
##        Dim.1  Dim.2
## diast   0.81  0.455
## syst    0.91 -0.067
## chol   -0.33  0.917
```





Interpreter la position des patients (gauche, doite, haut, bas) en fonction des variables.

Résumé



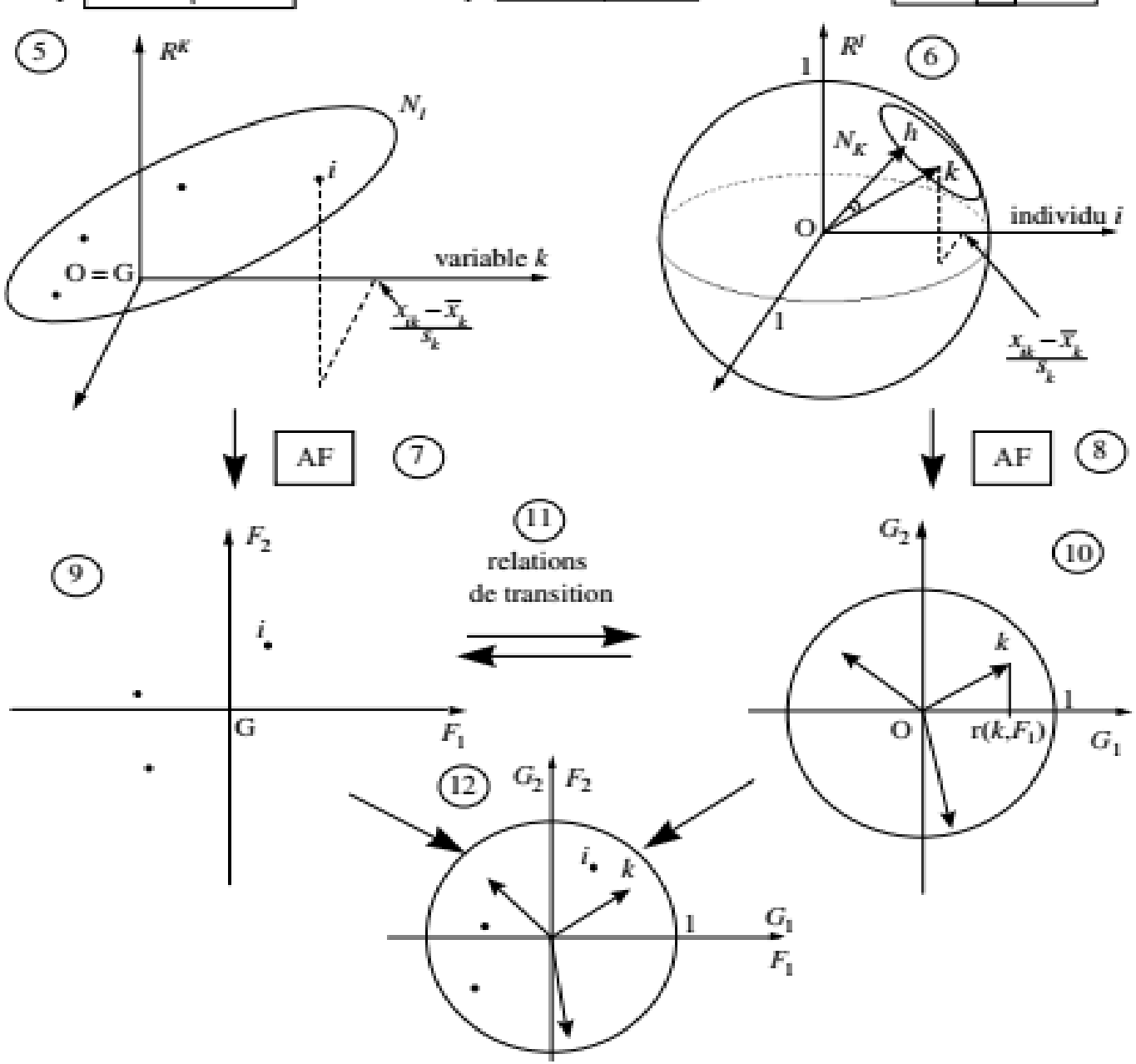


Figure 1.10 Schéma général de l'ACP.

Exemple

On considère le tableau de données X de type (3,2) suivant:

$$X = \begin{pmatrix} 2 & 3 \\ 4 & 5 \\ 6 & 1 \end{pmatrix}$$

1) Donner le tableau des données centrés et réduites (normées).
2) Déterminer la matrice des corrélations Γ.
3) Diagonaliser la matrice Γ. On note λ_1 et λ_2 ses valeurs propres avec $\lambda_1 > \lambda_2$.
4) Déterminer F_i les axes factoriels. Donner le vecteur unitaire u_i de chaque axe F_i.

   Vérifier que ces axes sont perpendiculaires.
5) Ecrire la matrice diagonale des valeurs propres Λ et calculer sa trace tr(Λ) et vérifier que tr(Λ)=tr(Γ).

$$Y = \begin{pmatrix} -2 & 0 \\ 0 & 2 \\ 2 & -2 \end{pmatrix} \quad Z = \begin{pmatrix} -\sqrt{\frac{3}{2}} & 0 \\ 0 & \sqrt{\frac{3}{2}} \\ \sqrt{\frac{3}{2}} & -\sqrt{\frac{3}{2}} \end{pmatrix} \quad V = \frac{1}{3} Y'Y = \begin{pmatrix} \frac{8}{3} & -\frac{4}{3} \\ -\frac{4}{3} & \frac{8}{3} \end{pmatrix} \quad \Gamma = \frac{1}{3} Z'Z = \begin{pmatrix} 1 & -1/2 \\ -\frac{1}{2} & 1 \end{pmatrix}$$

Diagonalisation $\Rightarrow$ $\lambda_2 = \frac{1}{2}$ et $\lambda_1 = \frac{3}{2}$ $\Rightarrow$ les valeurs propres de Γ

Les vecteurs et les sous-espaces propres : Pour $\lambda_2 = \frac{1}{2}$

$$\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}$$

$\lambda_1 = \frac{3}{2}$

$$u_1 = \left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right) \qquad u_2 = \left(+\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right)$$

$$\Lambda = \begin{pmatrix} \frac{3}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \Rightarrow tr(\Lambda) = 2 = tr(\Gamma)$$

Exemple théorique illustratif :

$$X=\begin{pmatrix}3 & 4\\ 7 & 2\\ 5 & 6\\ 5 & 4\end{pmatrix} \quad Y=\begin{pmatrix}-2 & 0\\ 2 & -2\\ 0 & 2\\ 0 & 0\end{pmatrix} \quad Z=\begin{pmatrix}-\sqrt{2} & 0\\ \sqrt{2} & -\sqrt{2}\\ 0 & \sqrt{2}\\ 0 & 0\end{pmatrix} \quad R=\frac{1}{4}Z^tZ=\begin{pmatrix}1 & -\frac{1}{2}\\ -\frac{1}{2} & 1\end{pmatrix}$$

$$\lambda_1=\frac{3}{2} \quad \lambda_2=\frac{1}{2} \quad s\begin{pmatrix}\frac{\sqrt{2}}{2}\\ -\frac{\sqrt{2}}{2}\end{pmatrix} et \begin{pmatrix}\frac{\sqrt{2}}{2}\\ \frac{\sqrt{2}}{2}\end{pmatrix}$$

Les nouvelles coordonnées des individus :

$$\Psi=[\psi_1,\psi_2]=Z.U=\begin{pmatrix}-\sqrt{2} & 0\\ \sqrt{2} & -\sqrt{2}\\ 0 & \sqrt{2}\\ 0 & 0\end{pmatrix}\begin{pmatrix}\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\end{pmatrix}=\begin{matrix}\text{les abscisses} & \text{les ordonnées}\\ \begin{pmatrix}-1 & -1\\ 2 & 0\\ -1 & 1\\ 0 & 0\end{pmatrix}\end{matrix}$$

Les nouvelles coordonnées des variables sont

$$\phi^1=\sqrt{\lambda_1}\,u=\sqrt{\frac{3}{2}}\begin{pmatrix}\frac{\sqrt{2}}{2}\\ -\frac{\sqrt{2}}{2}\end{pmatrix}=\begin{pmatrix}\frac{\sqrt{3}}{2}\\ -\frac{\sqrt{3}}{2}\end{pmatrix} \qquad \phi^2=\sqrt{\frac{1}{2}}\begin{pmatrix}\frac{\sqrt{2}}{2}\\ \frac{\sqrt{2}}{2}\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix}$$

Exemple théorique illustratif :

Qualité de représentation des individus :

$$\Psi = \begin{array}{c} \begin{matrix} \text{les abscisses} & \text{les ordonnées} \end{matrix} \\ \begin{pmatrix} -1 & -1 \\ 2 & 0 \\ -1 & 1 \\ 0 & 0 \end{pmatrix} \end{array}$$

$$qlt_{axe1}(individu\ \ 1) = \cos^2(\alpha_{11}) = \frac{c_{11}^2}{\sum_{j=1}^{2} c_{1j}^2} = \frac{(-1)^2}{(-1)^2 + (-1)^2} = \frac{1}{2}$$

$$qlt_{axe2}(individu\ \ 3) = \cos^2(\alpha_{32}) = \frac{c_{32}^2}{\sum_{j=1}^{2} c_{3j}^2} = \frac{1^2}{(-1)^2 + 1^2} = \frac{1}{2}$$

Contribution des individus à la formation des axes : $CTR_k(i) = \frac{c_{ik}^2}{n\lambda_k}$

$$\lambda_1 = \frac{3}{2}$$

$$\lambda_2 = \frac{1}{2}$$

$$CTR_1(1) = \frac{(-1)^2}{4 \times \frac{3}{2}} = \frac{1}{6} \qquad CTR_2(1) = \frac{(-1)^2}{4 \times \frac{1}{2}} = \frac{1}{2} \qquad CTR_1(2) = \frac{2^2}{4 \times \frac{3}{2}} = \frac{2}{3} \qquad CTR_2(2) = \frac{0^2}{4 \times \frac{3}{2}} = 0$$

$$CTR_1(3) = \frac{(-1)^2}{4 \times \frac{3}{2}} = \frac{1}{6} \qquad CTR_2(3) = \frac{1^2}{4 \times \frac{1}{2}} = \frac{1}{2} \qquad CTR_1(4) = \frac{0^2}{4 \times \frac{3}{2}} = 0 \qquad CTR_2(4) = \frac{0^2}{4 \times \frac{3}{2}} = 0$$

La somme des contributions vaut 1