Cours Analyse des données

(Chapitre1 Analyse en Composantes Principales(ACP))

Pr. A. GHAZDALI

a.ghazdali@usms.ma

2021

ENSA KHOURIBGA ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES

Université Sultan Moulay Slimane

Introduction

Il s'agit d'analyser un tableau de données quantitatives.

Exemple : données décrivant 8 eaux minérales sur 13 descripteurs sensoriels.

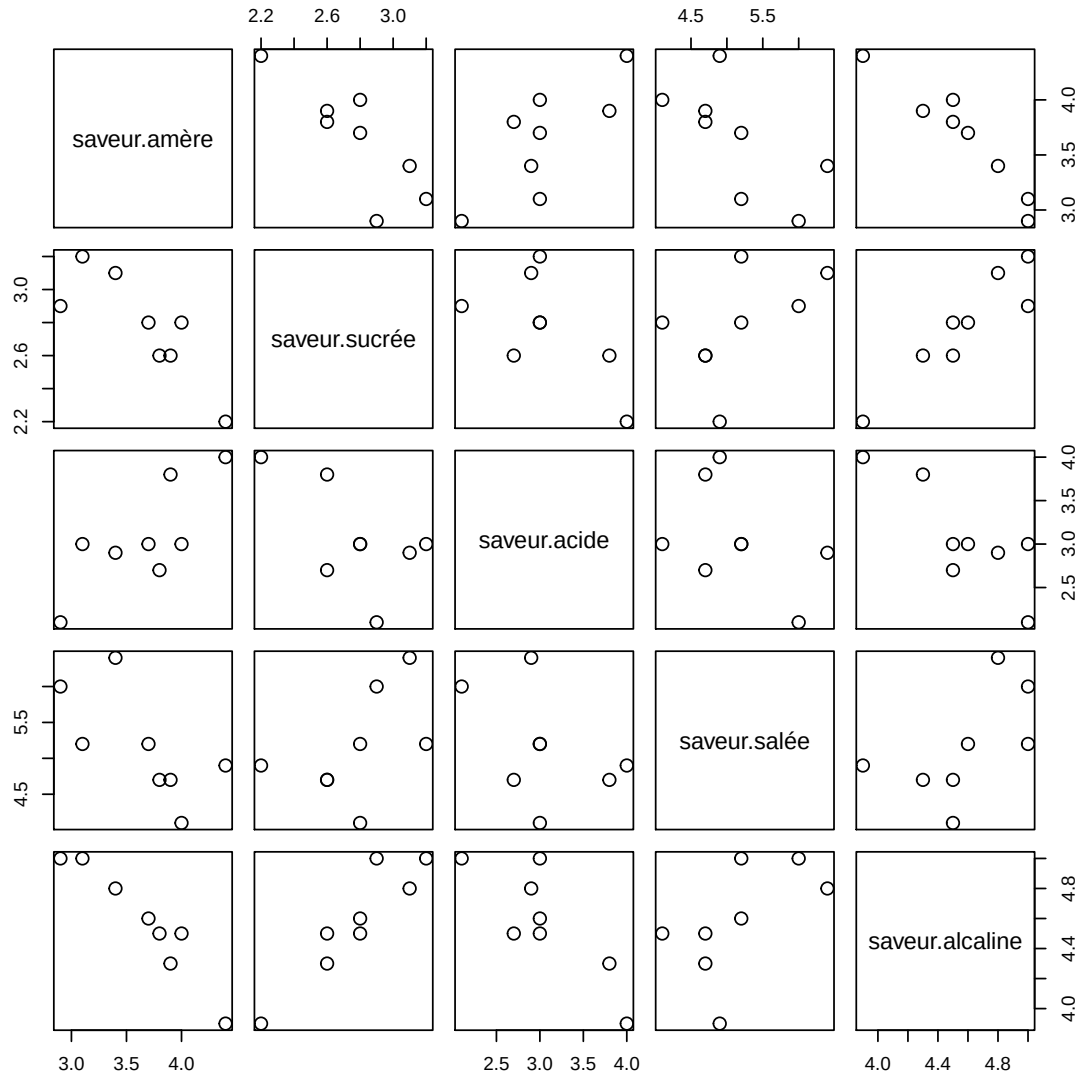
```
##             saveur.amère saveur.sucrée saveur.acide saveur.salée saveur.alcaline
## St Yorre             3.4           3.1          2.9          6.4             4.8
## Badoit               3.8           2.6          2.7          4.7             4.5
## Vichy                2.9           2.9          2.1          6.0             5.0
## Quézac               3.9           2.6          3.8          4.7             4.3
## Arvie                3.1           3.2          3.0          5.2             5.0
## Chateauneuf          3.7           2.8          3.0          5.2             4.6
## Salvetat             4.0           2.8          3.0          4.1             4.5
## Perrier              4.4           2.2          4.0          4.9             3.9
```

Les lignes correspondent à ce qu'on appelle des individus (ici des eaux minérales) et les colonnes à des variables (ici des descipteurs sensoriels).

L'objectif est alors de savoir :

- quels individus se ressemblent, (proximité entres les invidus)
- quelles variables sont liées.(les relations entre les variables)

Pour cela on peut faire de la statistique descriptive bivariée et visualiser les données par paires de variables :



Un graphique de la statistique descriptive bivariée ne fait que révéler des informations présentes dans le tableau de données par paires !

On pourrait croiser les variables 2 à 2, mais :

1. Très difficile de surveiller plusieurs cadrans en même temps.

2. Etiqueter les points rendrait le tout illisible.

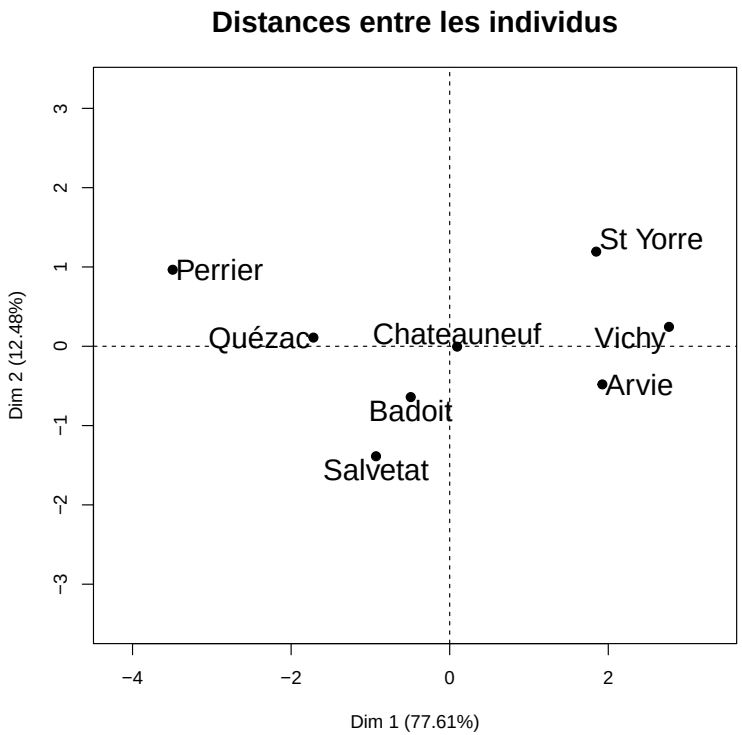
Ce type de représentation n'est utile que pour effectuer un diagnostic rapide et repérer les points atypiques.

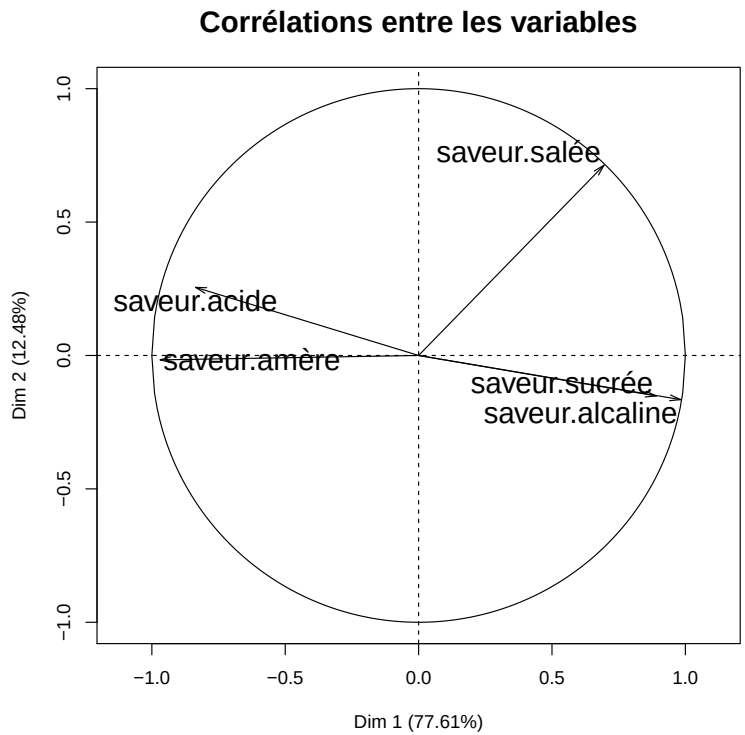
Impossible de créer un nuage à « p » dimensions.

Que faire si on veut prendre en compte (p >2) variables simultanément ?

Il existe aussi des méthodes de statistique descriptive multivariée comme l'ACP pour :

- visualiser sur des graphiques distances entre les individus et des corrélations entre les variables.





- Constuire des nouvelles variables "résumant" au mieux les variables initiales et ainsi réduire la dimension.

TABLE: Données initiales

	saveur.amère	saveur.sucrée	saveur.acide	saveur.salée	saveur.alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Badoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3
Arvie	3.1	3.2	3.0	5.2	5.0
Chateauneuf	3.7	2.8	3.0	5.2	4.6
Salvetat	4.0	2.8	3.0	4.1	4.5
Perrier	4.4	2.2	4.0	4.9	3.9

TABLE: Deux nouvelles variables résumant les variables initiales

	Dim.1	Dim.2
St Yorre	1.85	1.19
Badoit	-0.49	-0.64
Vichy	2.77	0.24
Quézac	-1.72	0.11
Arvie	1.93	-0.48
Chateauneuf	0.09	0.00
Salvetat	-0.93	-1.39
Perrier	-3.49	0.97

Principe : Construire un système de représentation de dimension réduite (q << p) qui préserve les distances entre les individus. On peut la voir comme une compression avec perte (contrôlée) de l'information.

Plan

Notions de base

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

	$1 \ldots$	j	$\ldots p$
1			
$\vdots$		$\vdots$	
i	$\ldots$	x_{ij}	$\ldots$
$\vdots$		$\vdots$	
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 $\mathbf{X}$).

Exemple : mesure de la tension arterielle diastolique, systolique et du taux de cholestérol de 6 patients.

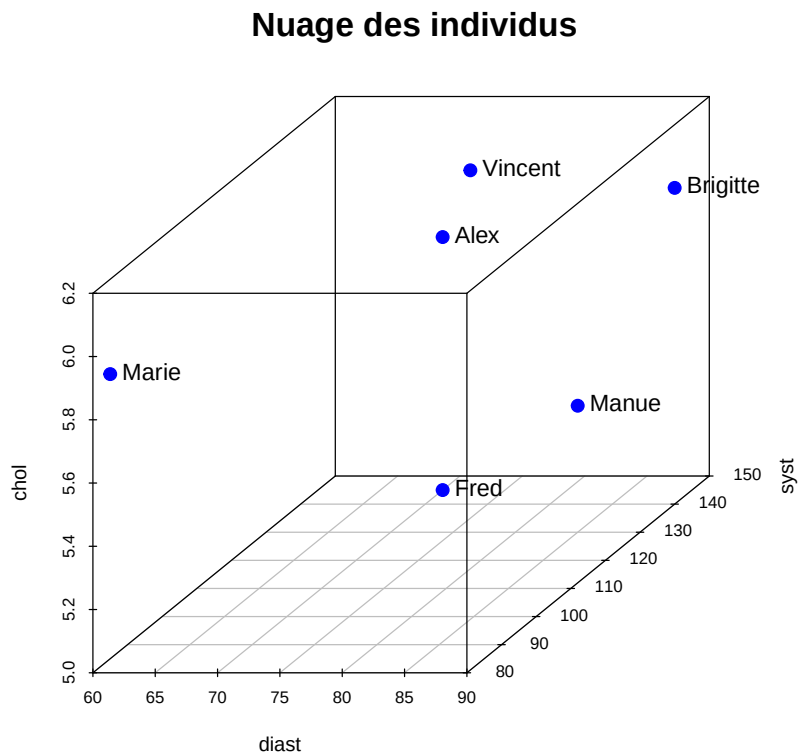
```
##          diast syst chol
## Brigitte    90  140  6.0
## Marie       60   85  5.9
## Vincent     75  135  6.1
## Alex        70  145  5.8
## Manue       85  130  5.4
## Fred        70  145  5.0
```

$n =$ $p =$ $\mathbf{X} =$ $\mathbf{x}_3 =$ $\mathbf{x}^2 =$

$\Rightarrow$ Deux nuages de points.

Le premier nuage de points est le nuage des individus.

Exemple : les 6 patients définissent un nuage de $n = 6$ points de $\mathbb{R}^3$.



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.

Matrice **X** des données brutes

	$1 \ldots$	j	$\ldots p$	
1				
$\vdots$		$\vdots$		
i	$\ldots$	x_{ij}	$\ldots$	
$\vdots$		$\vdots$		
n				
$\bar{\mathbf{x}}$	$\ldots$	$\bar{x}^j$	$\ldots$	

Matrice **Y** des données centrées

	$1 \ldots$	j	$\ldots p$	
1				
$\vdots$		$\vdots$		
i	$\ldots$	y_{ij}	$\ldots$	
$\vdots$		$\vdots$		
n				
$\bar{\mathbf{y}}$	$\ldots$	0	$\ldots$	

Ici :

- $\bar{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ est la moyenne de la variable j non centrée (colonne j de **X**),
- $y_{ij} = x_{ij} - \bar{x}^j$ est le terme général de la matrice **Y** des données centrées.

Les colonnes de la matrice centrée **Y** sont de moyenne nulle :

$$\bar{y}^j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = 0.$$

Exemple : le nuage des 6 patients.

Matrice **X** des données brutes

```
##          diast syst chol
## Brigitte    90  140  6.0
## Marie       60   85  5.9
## Vincent     75  135  6.1
## Alex        70  145  5.8
## Manue       85  130  5.4
## Fred        70  145  5.0
```

Moyennes des colonnes de **X**

```
## diast  syst  chol
##  75.0 130.0   5.7
```

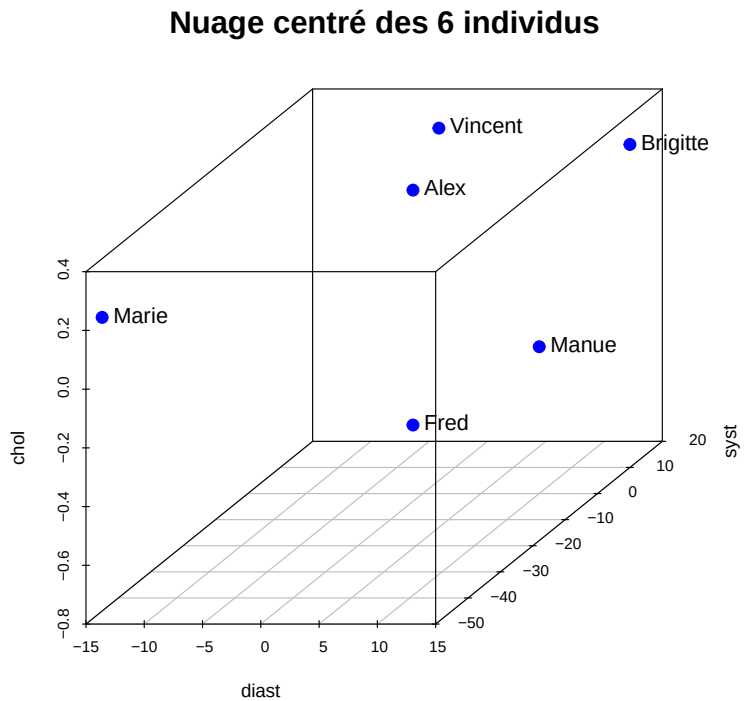
Matrice **Y** des données centrées

```
##          diast syst chol
## Brigitte    15   10  0.3
## Marie      -15  -45  0.2
## Vincent      0    5  0.4
## Alex        -5   15  0.1
## Manue       10    0 -0.3
## Fred        -5   15 -0.7
```

Moyennes des colonnes de **Y**

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

- Centrer les données revient à faire une translation du nuage de individus.



- Pour que barycentre G soit situé à l'origine [obligatoire]

Matrice **X** des données brutes

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

Matrice **Z** des données centrées-réduites

	$1 \ldots$	j	$\ldots p$
1			
$\vdots$		$\vdots$	
i	$\ldots$	z_{ij}	$\ldots$
$\vdots$		$\vdots$	
n			
$\bar{\mathbf{z}}$	$\ldots$	0	$\ldots$
s	$\ldots$	1	$\ldots$

Ici :

- $s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}^j)^2$ est la variance de la variable j (colonne j de **X**),
- $z_{ij} = \frac{x_{ij} - \bar{x}^j}{s_j}$ est le terme général de la matrice **Z** des données centrées-réduites.

Les colonnes de la matrice centrées-réduites **Z** sont de moyenne 0 et de variance 1 :

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

Exemple : le nuage des 6 patients.

Matrice **X** des données brutes

```
##          diast syst chol
## Brigitte    90  140  6.0
## Marie       60   85  5.9
## Vincent     75  135  6.1
## Alex        70  145  5.8
## Manue       85  130  5.4
## Fred        70  145  5.0
```

Matrice **Z** des données centrées-réduites

```
##          diast  syst  chol
## Brigitte   1.5  0.48  0.78
## Marie     -1.5 -2.16  0.52
## Vincent    0.0  0.24  1.04
## Alex      -0.5  0.72  0.26
## Manue      1.0  0.00 -0.78
## Fred      -0.5  0.72 -1.83
```

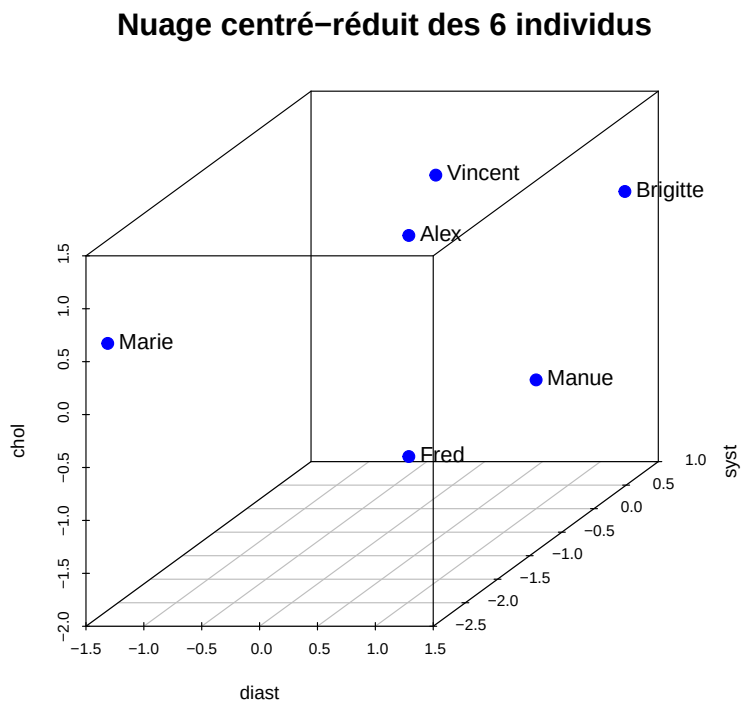
Moyennes et écart-types des colonnes de **X**

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

Moyennes et écart-types des colonnes de **Z**

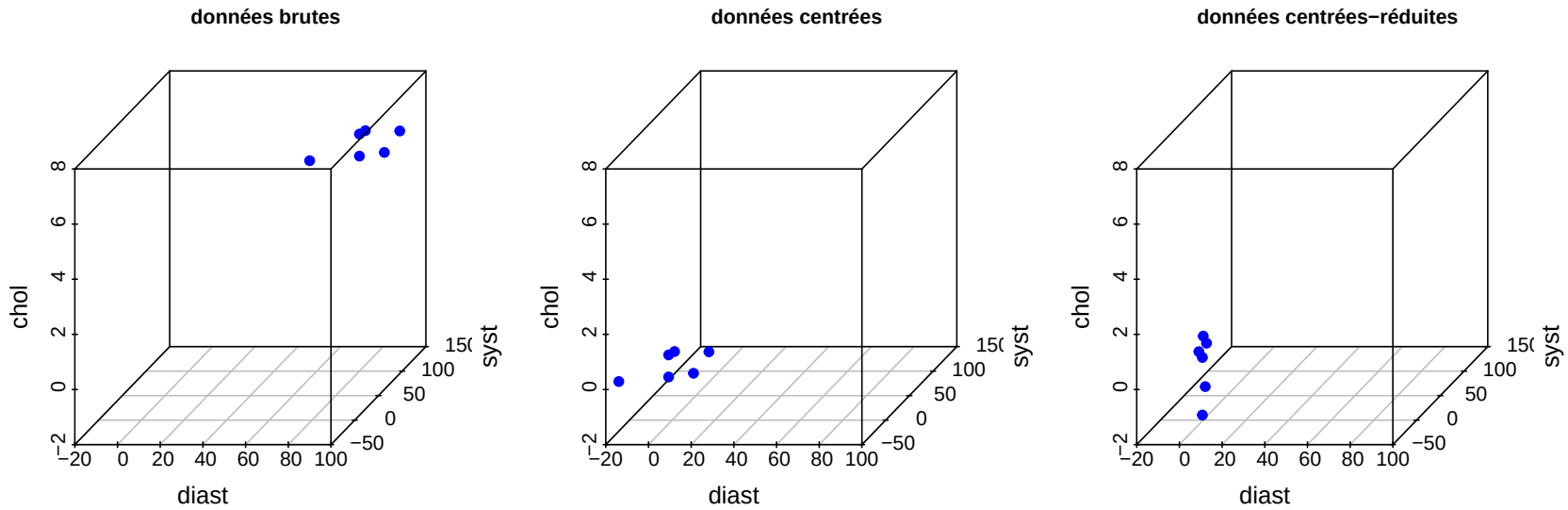
```
## diast  syst  chol
##     0     0     0
## diast  syst  chol
##     1     1     1
```

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



- Pour rendre comparables les variables exprimées sur des échelles (unités) différentes [non obligatoire]

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 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 $\mathbf{X}$) 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 $\mathbf{Z}$) 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 $\mathbf{X}$) 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 Brigitte et Marie

Données brutes ($\mathbf{X}$) :

```
##          diast syst chol
## Brigitte    90  140  6.0
## Marie       60   85  5.9
## Vincent     75  135  6.1
## Alex        70  145  5.8
## Manue       85  130  5.4
## Fred        70  145  5.0
```

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

```
##          diast  syst  chol
## Brigitte   1.5  0.48  0.78
## Marie     -1.5 -2.16  0.52
## Vincent    0.0  0.24  1.04
## Alex      -0.5  0.72  0.26
## Manue      1.0  0.00 -0.78
## Fred      -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 $\mathbf{X}$:

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

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

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

La dispersion du nuage des individus se mesure avec l'inertie.(quantité d'information)

- 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, $I(\mathbf{Z}) = p$.

L'inertie indique la dispersion autour du barycentre, c'est une variance multidimensionnelle (calculée sur p dimensions)

Exemple : Inertie du nuage des 6 patients

Données centrées ($\mathbf{Y}$) :

```
##          diast syst chol
## Brigitte    15   10  0.3
## Marie      -15  -45  0.2
## Vincent      0    5  0.4
## Alex        -5   15  0.1
## Manue       10    0 -0.3
## Fred        -5   15 -0.7
```

Variance des colonnes :

```
##  diast   syst   chol
## 100.00 433.33   0.15
```

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

```
##          diast  syst  chol
## Brigitte   1.5  0.48  0.78
## Marie     -1.5 -2.16  0.52
## Vincent    0.0  0.24  1.04
## Alex      -0.5  0.72  0.26
## Manue      1.0  0.00 -0.78
## Fred      -0.5  0.72 -1.83
```

Variance des colonnes :

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

- Inertie du nuage centré :

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

- Inertie du nuage centré-réduit :

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

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 $\mathbb{R}^6$.

```
##       Brigitte Marie Vincent  Alex Manue Fred
## 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 !

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 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 :

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

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 :

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

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

- $\mathbf{C} = \mathbf{Y}^T\mathbf{N}\mathbf{Y}$,
- $\mathbf{R} = \mathbf{Z}^T\mathbf{N}\mathbf{Z}$.

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 :

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

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 $\mathbf{Y}$,
- la matrice des données centrées-réduites $\mathbf{Z}$.

On distingue alors deux type d'ACP :

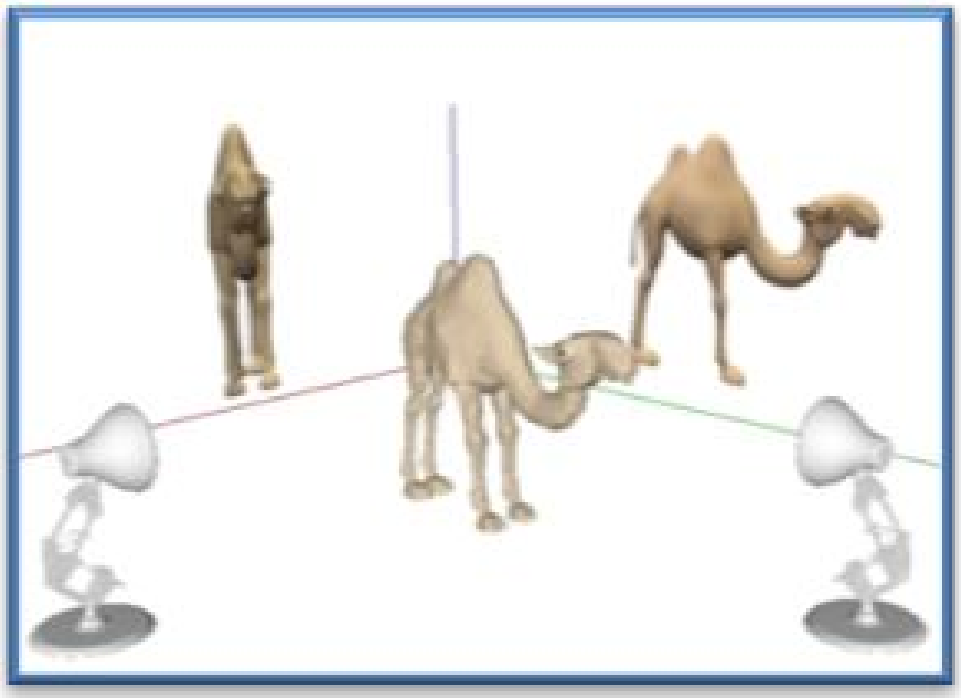
- l'ACP non normée (sur matrice des covariances) qui analyse $\mathbf{Y}$,
- l'ACP normée (sur matrice des corrélations) qui analyse $\mathbf{Z}$.

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

Plan

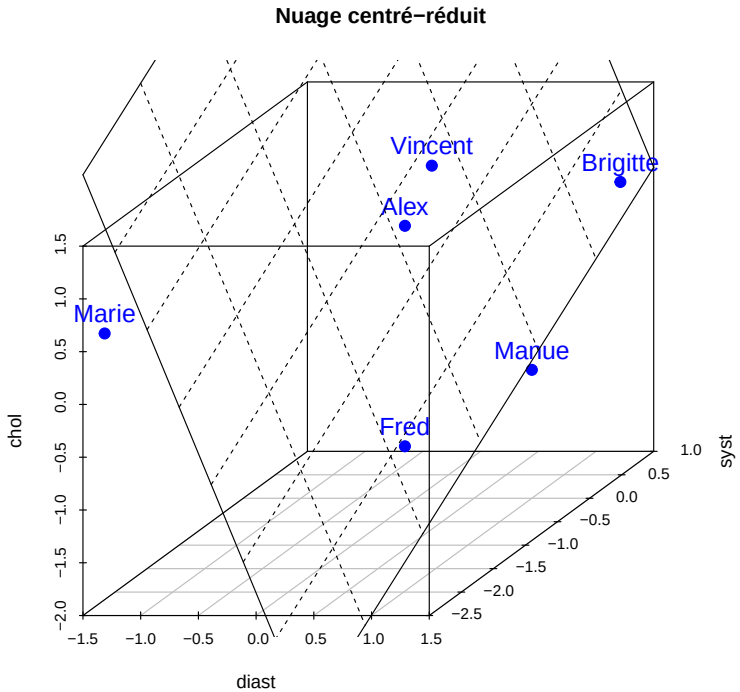
Analyse du nuage des individus

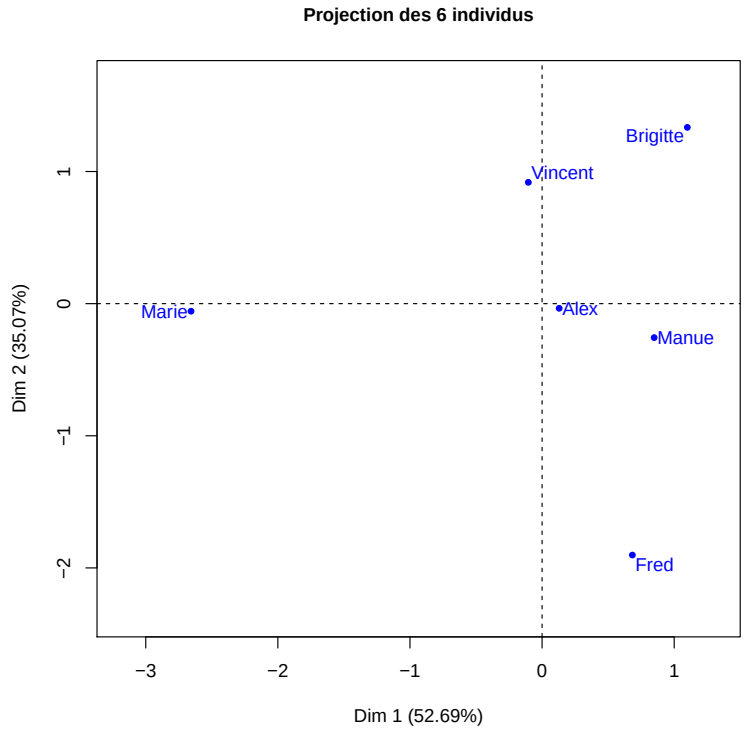
Trouver le sous-espace qui fournit la meilleure représentation des données.



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

Exemple des 6 patients décrits sur les 3 variables centrées-réduites.





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).

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 $\mathbf{v}_\alpha$ ($\mathbf{v}_\alpha^T \mathbf{v}_\alpha = 1$) a pour coordonnée :

$$f_{i\alpha} =< \mathbf{z}_i, \mathbf{v}_\alpha >= \mathbf{z}_i^T \mathbf{v}_\alpha,$$

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

$$\mathbf{f}^\alpha = \begin{pmatrix} f_{1\alpha} \\ \vdots \\ f_{n\alpha} \end{pmatrix} = \mathbf{Z}\mathbf{v}_\alpha = \sum_{j=1}^{p} v_{j\alpha}\mathbf{z}^j.$$

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

Exemple : les 6 individus (les patients) sont les lignes 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 6 individus du nuage centré-réduit sur deux axes orthogonaux Δ_1 et Δ_2 de vecteurs directeurs :

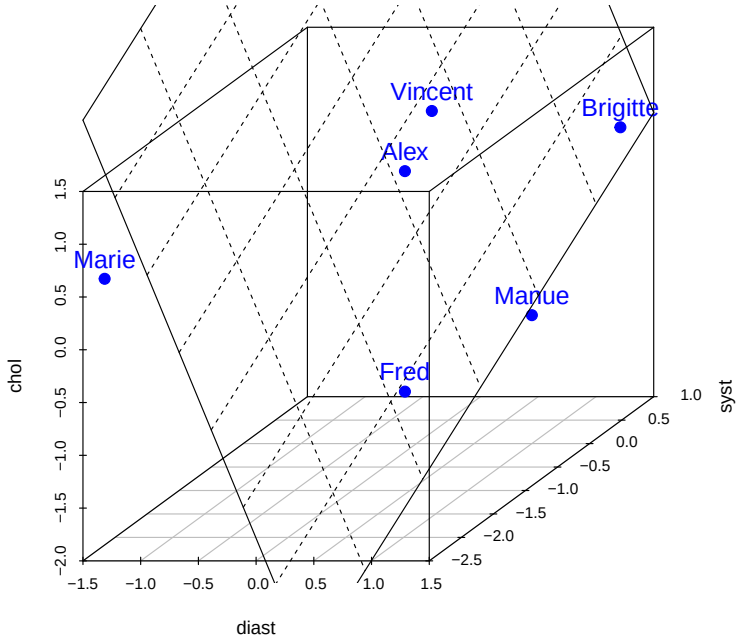
$$\mathbf{v}_1 = \begin{pmatrix} 0.641 \\ 0.72 \\ -0.265 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 0.4433 \\ -0.0652 \\ 0.894 \end{pmatrix}.$$

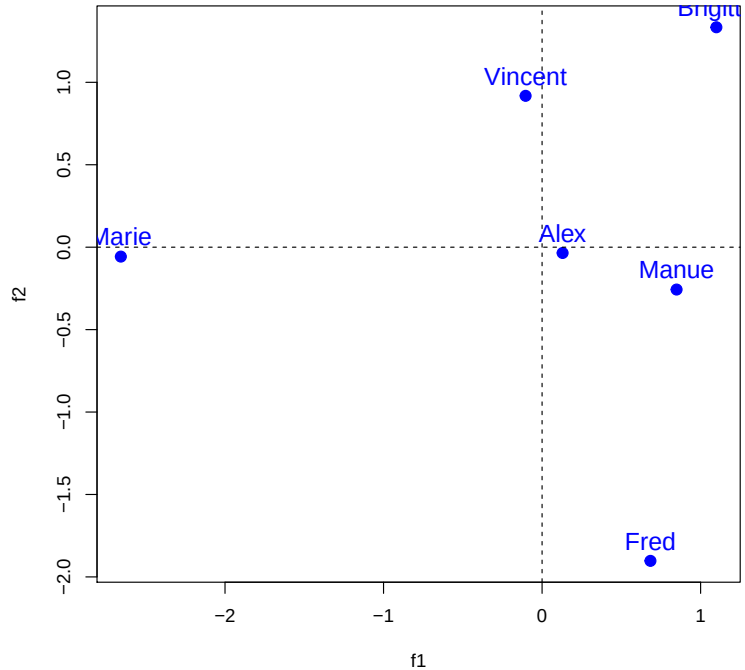
Les vecteurs $\mathbf{f}^1$ et $\mathbf{f}^2$ des coordonnées des projections des 6 individus sur Δ_1 et Δ_2 sont :

$$\mathbf{f}^1 = \mathbf{Z}\mathbf{v}_1 = 0.641 \begin{pmatrix} 1.5 \\ \vdots \\ -0.5 \end{pmatrix} + 0.72 \begin{pmatrix} 0.48 \\ \vdots \\ 0.72 \end{pmatrix} - 0.265 \begin{pmatrix} 0.78 \\ \vdots \\ -1.82 \end{pmatrix} = \begin{pmatrix} 1.09 \\ \vdots \\ 0.683 \end{pmatrix}$$

$$\mathbf{f}^2 = \mathbf{Z}\mathbf{v}_2 = 0.4433 \begin{pmatrix} 1.5 \\ \vdots \\ -0.5 \end{pmatrix} - 0.0652 \begin{pmatrix} 0.48 \\ \vdots \\ 0.72 \end{pmatrix} - 0.894 \begin{pmatrix} 0.78 \\ \vdots \\ -1.82 \end{pmatrix} = \begin{pmatrix} 1.333 \\ \vdots \\ -1.9 \end{pmatrix}$$

$\mathbf{f}^1$ et $\mathbf{f}^2$ sont deux nouvelles variables centrées.





En ACP les vecteurs directeurs $\mathbf{v}_1$ et $\mathbf{v}_2$ 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 $\mathbf{v}_1 \in \mathbb{R}^p$ qui maximise la variance des n individus projetés :

$$\begin{aligned}\mathbf{v}_1 &= \arg\max_{\|\mathbf{v}\|=1} Var(\mathbf{Z}\mathbf{v}) \\ &= \arg\max_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{R}\mathbf{v}\end{aligned}$$

où

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$$

est la matrice des corrélations entre les p variables.

On peut montrer que :

- $\mathbf{v}_1$ est le vecteur propre associé à la première valeur propre λ_1 de $\mathbf{R}$,
- La première composante principale $\mathbf{f}^1 = Z\mathbf{v}_1$ est centrée :

$$\bar{\mathbf{f}^1} = 0,$$

- λ_1 est la variance la première composante principale :

$$Var(\mathbf{f}^1) = \lambda_1.$$

Δ_2 est l'axe de vecteur directeur $\mathbf{v}_2 \perp \mathbf{v}_1$ qui maximise la variance des n individus projetés :

$$\mathbf{v}_2 = \arg \max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1} Var(\mathbf{Z}\mathbf{v}).$$

On peut montrer que :

- $\mathbf{v}_2$ est le vecteur propre associé à la seconde valeur propre λ_2 de $\mathbf{R}$,
- La seconde composante principale $\mathbf{f}^2 = Z\mathbf{v}_2$ est centrée :

$$\bar{\mathbf{f}^2} = 0,$$

- λ_2 est la variance la seconde composante principale :

$$Var(\mathbf{f}^2) = \lambda_2,$$

- Les composantes principales $\mathbf{f}^1$ et $\mathbf{f}^2$ 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.

En résumé :

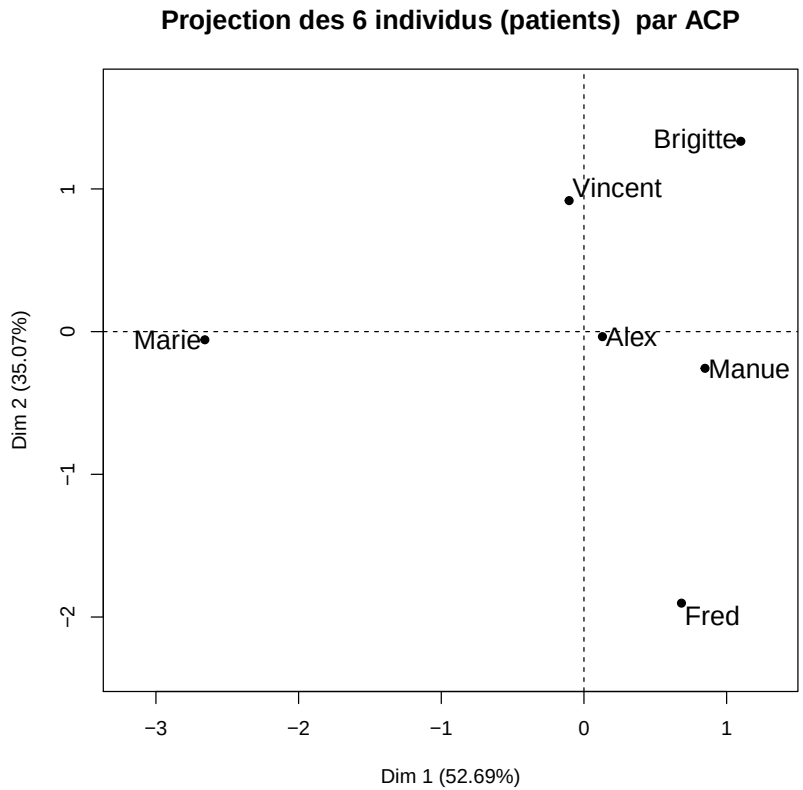
1. On effectue la décomposition en valeurs propres de la matrice des corrélations $\mathbf{R}$ et on choisit q.
2. On calcule la matrice $\mathbf{F} = \mathbf{ZV}$ des q composante principales à partir de la matrice $\mathbf{V}$ des q premiers vecteurs propres de $\mathbf{R}$.
   - Les composantes principales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

$\mathbf{F}=$

	$1 \ldots$	α	$\ldots q$	
1				
$\vdots$		$\vdots$		
i	$\ldots$	$f_{i\alpha}$	$\ldots$	
$\vdots$		$\vdots$		
n				
moy	$\ldots$	0	$\ldots$	
var	$\ldots$	λ_α	$\ldots$	

Exemple des 6 patients : matrice **F** des $q = 2$ premières CP

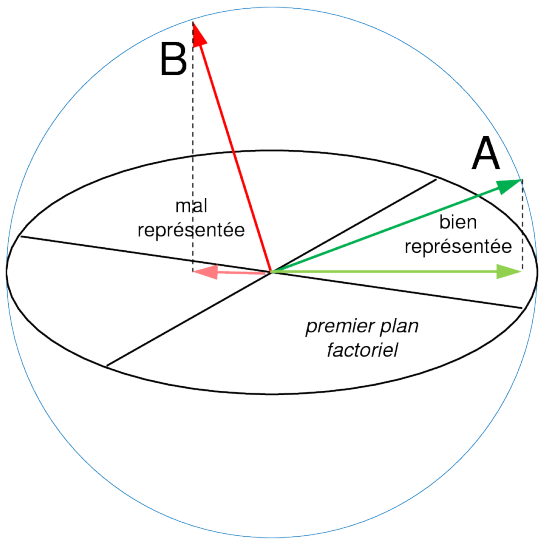
```
##             f1     f2
## Brigitte  1.10  1.334
## Marie    -2.66 -0.057
## Vincent  -0.10  0.918
## Alex      0.13 -0.035
## Manue     0.85 -0.257
## Fred      0.68 -1.903
```



Plan

Analyse du nuage des variables

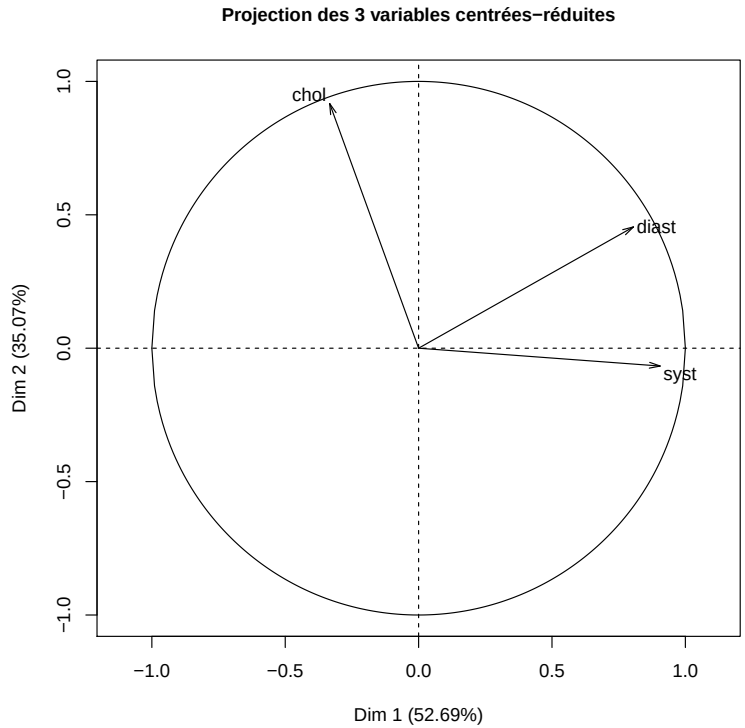
Trouver le sous-espace qui fournit la meilleure représentation des variables.



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 $\mathbf{u}_\alpha$ ($\mathbf{u}_\alpha^T \mathbf{N} \mathbf{u}_\alpha = 1$) a pour coordonnée :

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

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

$$\mathbf{a}^\alpha = \begin{pmatrix} a_{1\alpha} \\ \vdots \\ a_{p\alpha} \end{pmatrix} = \mathbf{Z}^T \mathbf{N} \mathbf{u}_\alpha$$

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

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

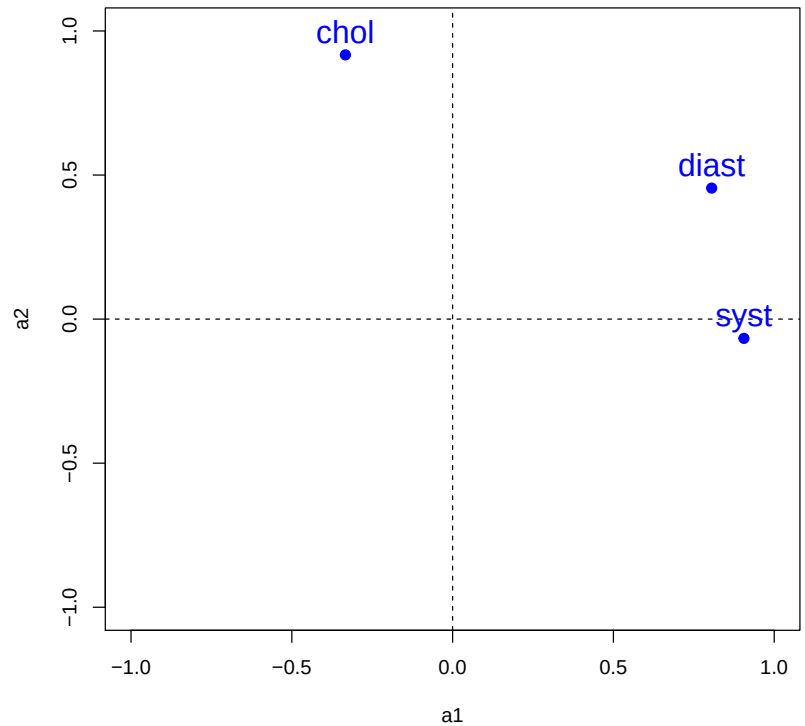
On veut projeter les 3 variabes sur deux axes N-orthogonaux G_1 et G_2 de vecteurs directeurs (ici $\mathbf{N} = \frac{1}{6}\mathbb{I}_6$) :

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

Les vecteurs $\mathbf{a}^1$ et $\mathbf{a}^2$ des coordonnées des projections des 3 variables sur G_1 et G_2 sont :

$$\mathbf{a}^1 = \mathbf{Z}^T\mathbf{N}\mathbf{u}_1 = \frac{0.87}{6}\begin{pmatrix}1.5\\0.48\\0.78\end{pmatrix}\frac{-2.11}{6}\begin{pmatrix}-1.5\\-2.16\\0.52\end{pmatrix} + \ldots \frac{+0.54}{6}\begin{pmatrix}-0.5\\0.72\\-1.83\end{pmatrix} = \begin{pmatrix}0.81\\0.91\\-0.33\end{pmatrix}$$

$$\mathbf{a}^2 = \mathbf{Z}^T\mathbf{N}\mathbf{u}_2 = \frac{1.30}{6}\begin{pmatrix}1.5\\0.48\\0.78\end{pmatrix}\frac{-0.06}{6}\begin{pmatrix}-1.5\\-2.16\\0.52\end{pmatrix} + \ldots \frac{-1.80}{6}\begin{pmatrix}-0.5\\0.72\\-1.83\end{pmatrix} = \begin{pmatrix}0.45\\-0.07\\0.92\end{pmatrix}$$



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 $\mathbf{u}_1 \in \mathbb{R}^n$ qui maximise la somme des carrés des cosinus des angles avec les variables.

$$\begin{aligned}\mathbf{u}_1 &= \arg \max_{\|\mathbf{u}\|_\mathbf{N}=1} \sum_{j=1}^{p} \cos^2 \theta_\mathbf{N}(\mathbf{z}^j, \mathbf{u}) \\ &= \arg \max_{\|\mathbf{u}\|_\mathbf{N}=1} \|\mathbf{Z}^T\mathbf{N}\mathbf{u}\|^2\end{aligned}$$

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

- $\mathbf{u}_1$ 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 $\mathbf{u}_1$:

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

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

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

On peut montrer que :

- $\mathbf{u}_2$ 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 $\mathbf{u}_2$:

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

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 :

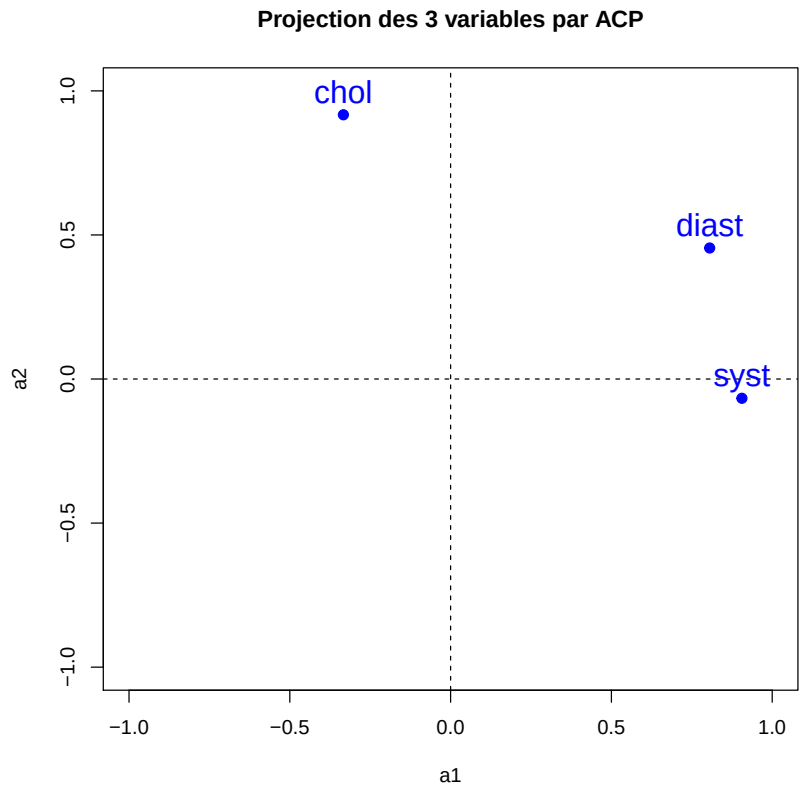
1. On effectue la décomposition en valeurs propres de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ et on choisit q.
2. On calcule la matrice $\mathbf{A} = \mathbf{Z}^T\mathbf{N}\mathbf{V}$ à partir de la matrice $\mathbf{U}$ des q premiers vecteurs propres de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$.
   - Les colonnes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

$\mathbf{A}=$

	$1 \ldots$	α	$\ldots q$	
1				
$\vdots$		$\vdots$		
i	$\ldots$	$a_{i\alpha}$	$\ldots$	
$\vdots$		$\vdots$		
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 $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$:

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

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

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

On en déduit que :

$$\mathbf{F} = \mathbf{U}\Lambda$$
$$\mathbf{A} = \mathbf{V}\Lambda$$

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

On en déduit aussi que

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

$$\mathbf{u}_\alpha = \frac{\mathbf{f}^\alpha}{\sqrt{\lambda_\alpha}},$$

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

$$a_{j\alpha} = cor(\mathbf{x}^j, \mathbf{f}^\alpha).$$

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

Plan

Interprétation des résultats

Variance des composantes principales.

Les composantes principales (colonnes de $\mathbf{F}$) sont q nouvelles variables synthétiques non corrélées et de variance maximale avec

$$Var(\mathbf{f}^{\alpha}) = \lambda_{\alpha}$$

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

$$I(\mathbf{F}) = \lambda_1 + \ldots + \lambda_q.$$

Exemple des 6 patients :

Les $p = 3$ valeurs propres non nulles de la matrice des corrélations $\mathbf{R}$ sont :

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

donc

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

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

$$I(\mathbf{F}) = \lambda_1 + \lambda_2 = 1.58 + 1.05 = 2.63.$$

Inertie totale.

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

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

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

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

Exemple des 6 patients :

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

$$I(\mathbf{F}) = \lambda_1 + \lambda_2 + \lambda_3 = 1.58 + 1.05 + 0.37 = 3$$

Qualité de la réduction de dimension.

- La proportion de l'inertie totale expliquée par la αème composante principale est :

$$\frac{Var(\mathbf{f}^\alpha)}{I(\mathbf{Z})} = \frac{\lambda_\alpha}{\lambda_1 + \ldots + \lambda_r}.$$

- La proportion de l'inertie totale expliquée par les q premières composantes principales est :

$$\frac{Var(\mathbf{F})}{I(\mathbf{Z})} = \frac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_r}.$$

Exemple des 6 patients.

Données brutes ($p = 3$ et n=6)

```
##          diast syst chol
## Brigitte    90  140  6.0
## Marie       60   85  5.9
## Vincent     75  135  6.1
## Alex        70  145  5.8
## Manue       85  130  5.4
## Fred        70  145  5.0
```

Réduction aux deux premières CP

```
##              f1     f2
## Brigitte  1.10  1.334
## Marie    -2.66 -0.057
## Vincent  -0.10  0.918
## Alex      0.13 -0.035
## Manue     0.85 -0.257
## Fred      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.

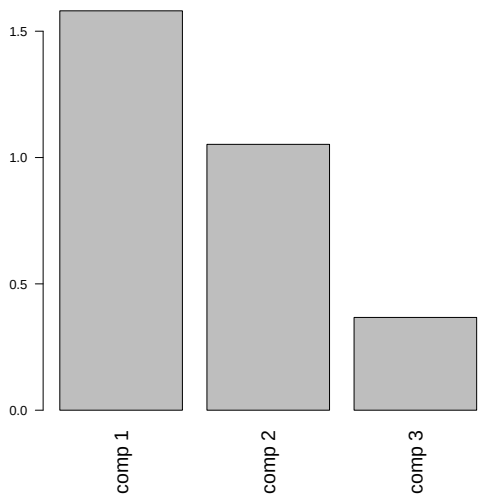
Combien de composantes retenir ?

- On peut choisir le nombre q de composantes à retenir en fonction d'un pourcentage d'inertie expliquée fixé a priori.
- On peut choisir de retenir les composantes apportant une inertie λ_α supérieure à l'inertie moyenne par variable. En ACP normée, l'inertie moyenne par variable vaut 1, et on choisit q tel que $\lambda_q > 1$ et $\lambda_{q+1} < 1$. C'est la règle de Kaiser.
- Visualiser l'histogramme des valeurs prores (qui n'est pas un histogramme) et chercher une "cassure". Pour quantifier cette cassure, on peut utiliser la règle du coude :
  i. calculer les différence premières : $\epsilon_1 = \lambda_1 - \lambda_2$, $\epsilon_2 = \lambda_2 - \lambda_2$, ...
  ii. calculer les différence secondes : $\delta_1 = \epsilon_1 - \epsilon_2$, $\delta_2 = \epsilon_2 - \epsilon_2$, ...
  iii. retenir le nombre q tel que $\delta_1, \ldots, \delta_{q-1}$ soient toutes positives et que δ_q soit négative.
- Choisir le nombre de composantes en fonction d'un critère de stabilité estimé par des approches bootstrap ou de validation croisée.

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 :

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

- 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 $(\Delta_\alpha, \Delta_{\alpha'})$:

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

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

Exemple des 6 patients.

Coordonnées factorielles des patients

```
##          Dim.1  Dim.2
## Brigitte  1.10  1.334
## Marie    -2.66 -0.057
## Vincent  -0.10  0.918
## Alex      0.13 -0.035
## Manue     0.85 -0.257
## Fred      0.68 -1.903
```

cos^2 des patients sur les axes

```
##          Delta1  Delta2
## Brigitte 0.3907 0.57504
## Marie    0.9812 0.00045
## Vincent  0.0094 0.73386
## Alex     0.0200 0.00148
## Manue    0.4462 0.04094
## Fred     0.1137 0.88080
```

La qualité de la représentation de Brigitte sur le premier axe factoriel est 0.3907.

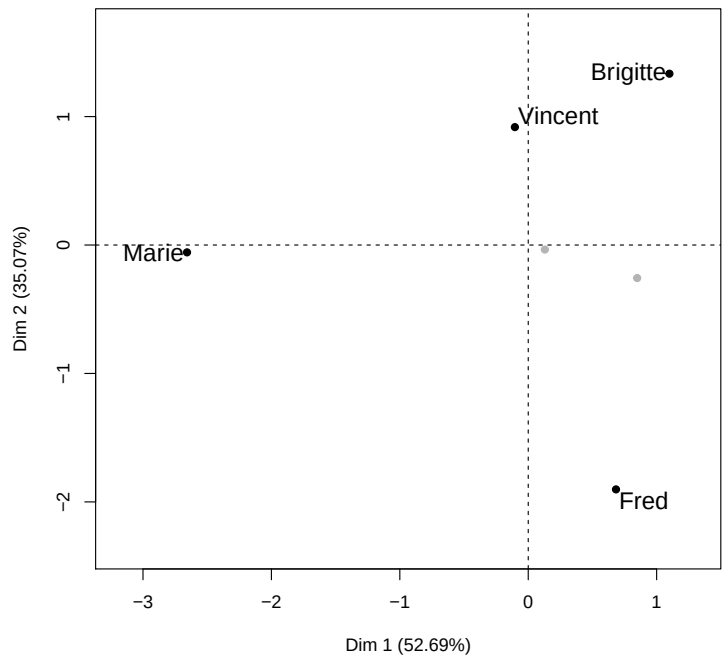
Les *cos*2 des patients sur les axes peuvent être retrouvés avec les distance des patients à l'origine :

```
## Brigitte    Marie  Vincent     Alex    Manue     Fred
##     1.76     2.68     1.07     0.92     1.27     2.03
```

Brigitte est très bien représentée sur le premier plan factoriel car son cos^2 sur ce plan vaut $0.966 = 0.3907 + 0.57504$.

De même on trouve que les cos^2 des 6 patients sur le premier plan factoriel sont :

```
##     Fred    Marie Brigitte  Vincent    Manue     Alex
##    0.994    0.982    0.966    0.743    0.487    0.022
```



- Les individus sont-ils globalement bien projetés sur ce plan ?
- Interpréter les distances entre Vincent et Marie et entre Vincent et Brigitte.

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 $\lambda_\alpha = \sum_{i=1}^{n} w_i f_{i\alpha}^2$ avec souvent $w_i = \frac{1}{n}$.
- La contribution relative d'un individu i à l'inertie de l'axe Δ_α est

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

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

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

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 des patients.

Coordonnées factorielles des patients

```
##             f1     f2
## Brigitte  1.10  1.334
## Marie    -2.66 -0.057
## Vincent  -0.10  0.918
## Alex      0.13 -0.035
## Manue     0.85 -0.257
## Fred      0.68 -1.903
```

Contributions des patients

```
##          Delta1 Delta2
## Brigitte  12.75 28.186
## Marie     74.44  0.052
## Vincent    0.11 13.352
## Alex       0.18  0.020
## Manue      7.59  1.046
## Fred       4.93 57.345
```

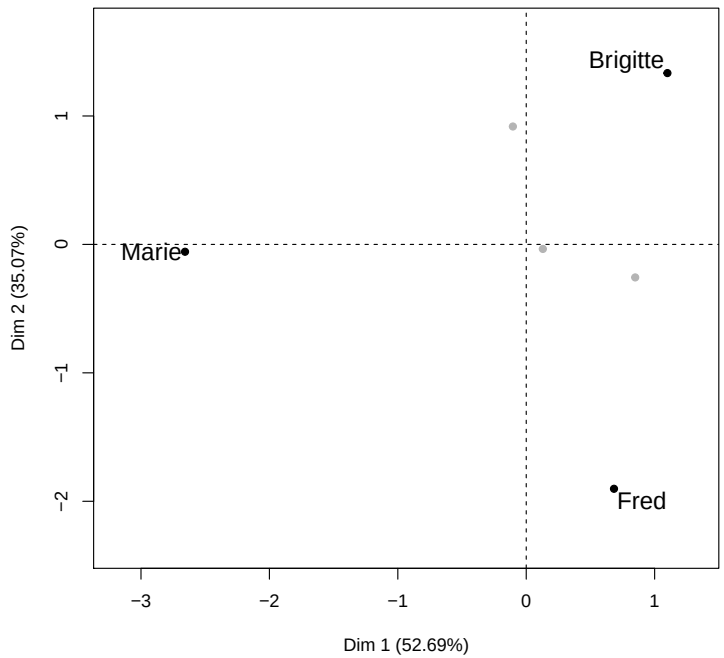
Deux premières valeurs propres

```
## lambda1 lambda2
##     1.6     1.1
```

- Calculer la contribution de Brigitte à l'inertie du premier axe. Calculer ensuite la contribution de Brigitte à l'inertie du premier plan factoriel.
- Vérifier que pour chaque axe, la somme des contributions relative est 1.

Les contributions des 6 patients à l'inertie du premier plan factoriel sont :

```
##    Marie     Fred Brigitte  Vincent    Manue     Alex
##    44.71    25.87    18.92     5.40     4.98     0.11
```



Ce sont bien les 3 patients les plus excentrés.

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 :

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

- 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 $(G_\alpha, G_{\alpha'})$:

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

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

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

Exemple des 3 variables décrivant les patients.

Coordonnées factorielles des variables

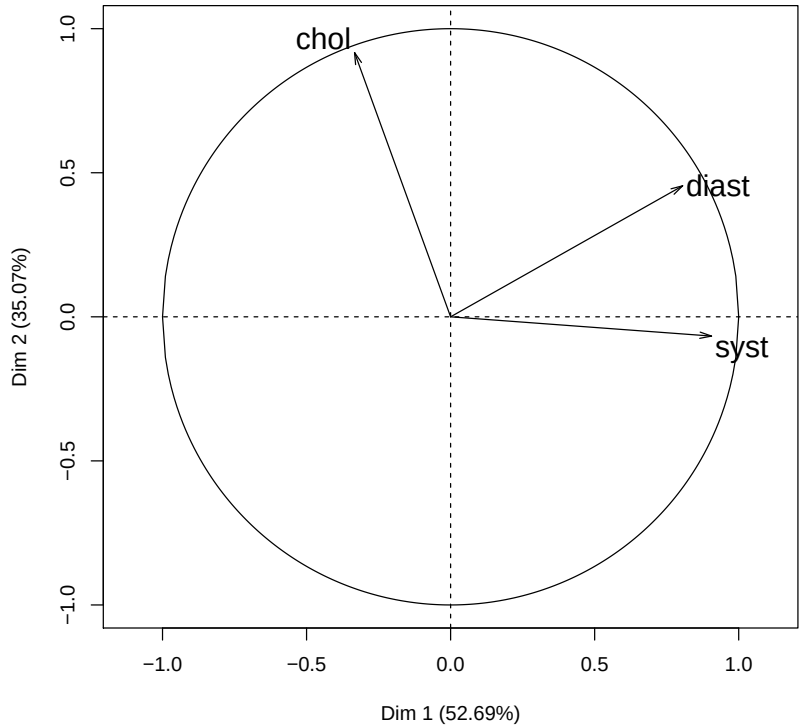
```
##          a1     a2
## diast  0.81  0.455
## syst   0.91 -0.067
## chol  -0.33  0.917
```

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 $\sqrt{0.8568} = 0.92563$.



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

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

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

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

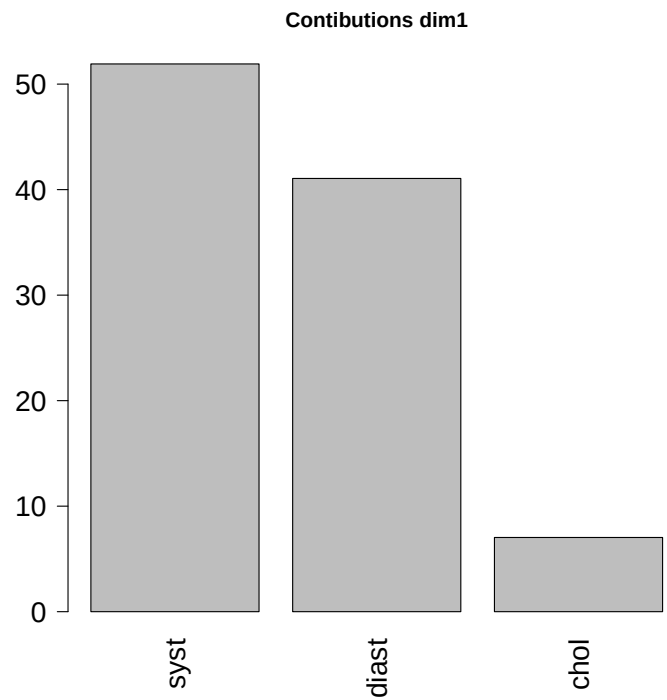
- La contribution relative d'une variable j au plan (G_α, G'_α) est

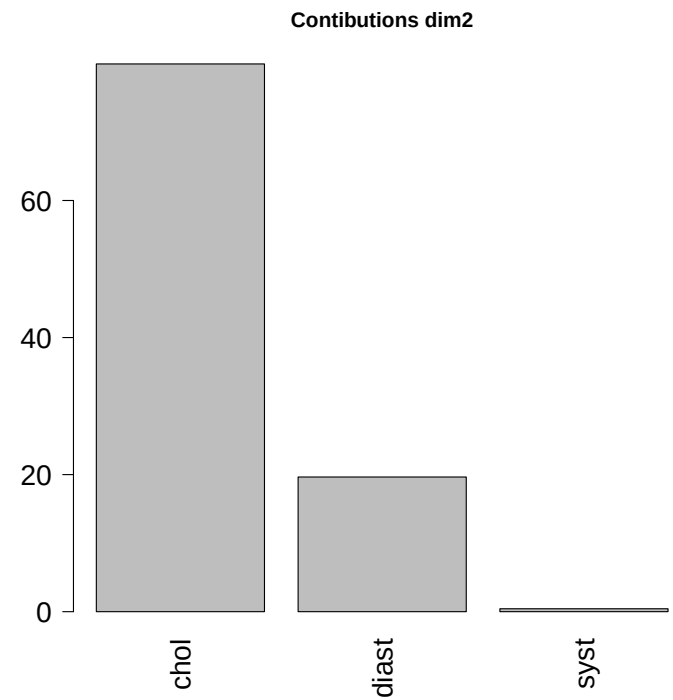
$$Ctr(j,(\alpha,\alpha')) = \frac{a_{j\alpha}^2 + a_{j\alpha'}^2}{\lambda_\alpha + \lambda'_\alpha}.$$

Exemple des 3 variables décrivant les patients.

Contributions relatives (en pourcentage) des 3 variables aux 2 axes :

```
##       G1    G2
## diast 41 19.65
## syst  52  0.42
## chol   7 79.92
```

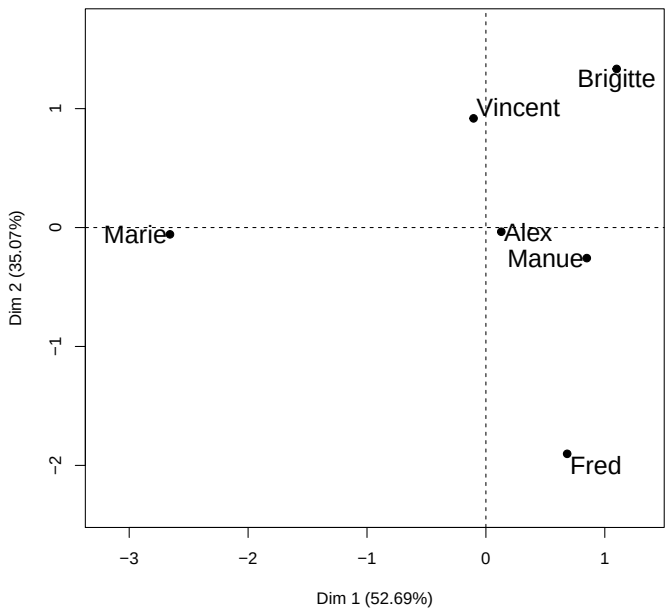


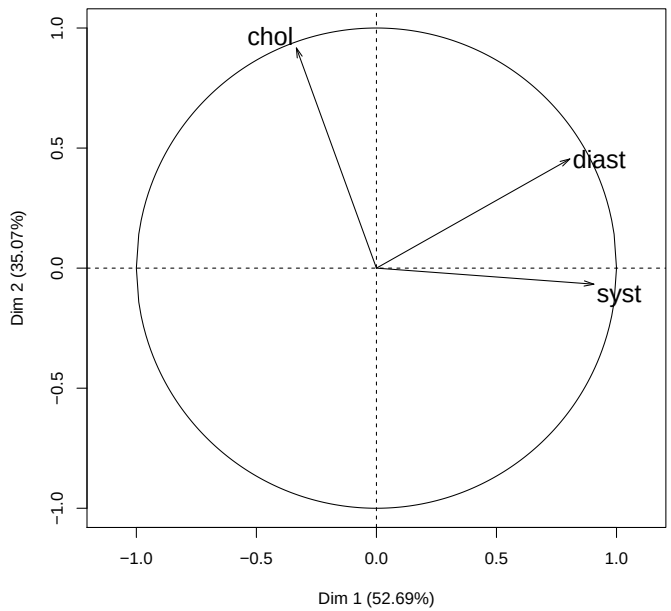


Interprétation du plan factoriel des individus à partir du cercle des corrélations.

$$a_{j\alpha} = cor(\mathbf{x}^j, \mathbf{f}^\alpha)$$

```
##       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.