Analyse des données

Apprentissage supervisé et non supervisé

[Apprentissage non supervisé : ACP]

3ème année ENSEIRB-MATMECA - CISD

Méthodes de Machine Learning (Apprentissage Automatique)

Non supervisées

$X^1, \ldots, X^p$: variables (quantitatives ou qualitatives)

Clustering

Création d'une nouvelle variable qualitative

Exemple : k-means, CAH, GMM,…

Réduction de dimension

Création de nouvelles variables quantitatives qui résument $X^1, \ldots, X^p$.

Exemples de méthodes linéaires :

- *ACP si les données sont quantitatives*
- *ACM si les données sont qualitatives*
- *ACPmixte si les données sont mixtes*

Exemple de méthodes non linéaire : AutoEncoders,…

Supervisées

$X^1, \ldots, X^p$: variables d'entrées

Y : variable de sortie (quantitative ou qualitative

Régression : Y quantitatif

Exemples de méthodes linéaires :

- *Régression linéaire simple et multiple si entrées quantitatives*
- *ANOVA si entrées qualitatives*
- *ANCOVA si les entrées sont mixtes*

Exemples de méthodes non linéaires :

- *Arbre de décision et forêts aléatoires (entrées mixtes)*
- *SVM, réseaux de neurones*

Classification : Y qualitatif

Exemples de méthodes linéaires :

- *Régression logistique (Y binaire)*
- *LDA et QDA (entrées quantitatives)*

Exemples de méthodes non linéaires :

- *KNN, réseaux de neurone (entrées quantitatives)*
- *Bayésien naif, arbres et forêts aléatoires (entrées mixtes)*

Introduction

Vocabulaire :

- ACP = Analyse en Composantes Principales.
- PCA = Principal Component Analysis.

Quelques ressources pour ce cours :

- Vidéo sur l'ACP : `https://youtu.be/i-mEyUa9U5k`
- Livre : Analyse des données avec R, F. Husson, S. Lê, J. Pagès, éditions PUR
- R package : FactoMineR
- Ressources pédagogiques : `https://marie-chavent.perso.math.cnrs.fr/teaching/`

Introduction

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

Exemple : données décrivant 8 eaux minérales sur 5 descripteurs sensoriels.

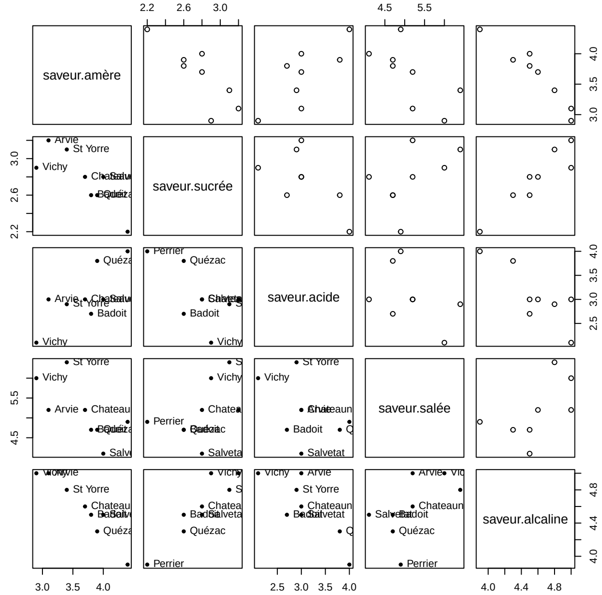
	saveur.amère	saveur.sucrée	saveur.acide	saveur.salée	saveur.alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Badoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3
Arvie	3.1	3.2	3.0	5.2	5.0
Chateauneuf	3.7	2.8	3.0	5.2	4.6
Salvetat	4.0	2.8	3.0	4.1	4.5
Perrier	4.4	2.2	4.0	4.9	3.9

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

L'objectif est alors de savoir :

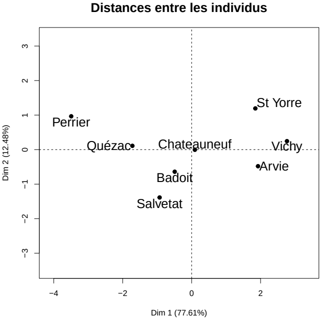
- quels individus se ressemblent,
- quelles variables sont liées.

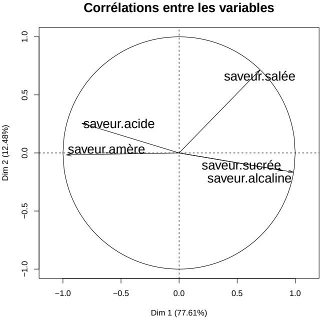
On peut faire de la statistique descriptive bivariée :



On peut faire de statistique descriptive multivariée, par exemple de l'ACP pour :

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





- construire 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

Plan

Données et exemples

L'ACP s'intéresse à des tableaux de données rectangulaires numériques où les individus sont en lignes et les variables en colonnes.

$$
\begin{array}{c|ccc|}
 & 1 \ldots & j & \ldots p \\
\hline
1 & & & \\
\vdots & & \vdots & \\
i & \ldots & x_{ij} & \ldots \\
\vdots & & \vdots & \\
n & & & \\
\hline
\end{array}
$$

On notera :

- $\mathbf{X} = (x_{ij})_{n \times p}$ le tableau des données brutes où $x_{ij} \in \mathbb{R}$ est la valeur du $i^{\text{ème}}$ individu sur la $j^{\text{ème}}$ variable.
- $\bar{x}^j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$ la moyenne de la variable j
- $s_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}^j)^2}$ l'écart-type de la variable æ

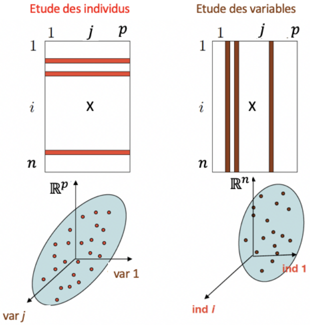
Deux nuages de points

Chaque individu est un point de $\mathbb{R}^p$ (ligne de $\mathbf{X}$) :

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

Chaque variable est un point de $\mathbb{R}^n$ (colonne de $\mathbf{X}$) :

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



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

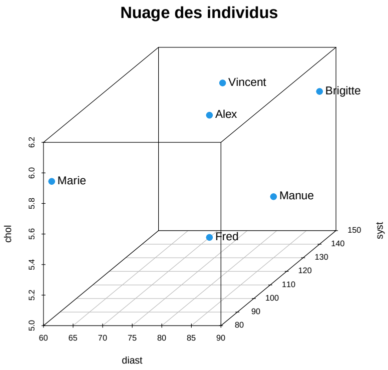
$\Rightarrow$ Deux nuages de points :

- Nuage des 6 individus : 6 points de $\mathbb{R}^3$.
- Nuage des 3 variables : 3 points dans $\mathbb{R}^6$.

Plan

Analyse du nuage des individus

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



Les données sont pré-traitées avant d'être analysées en ACP :

Matrice **Y** des données centrées

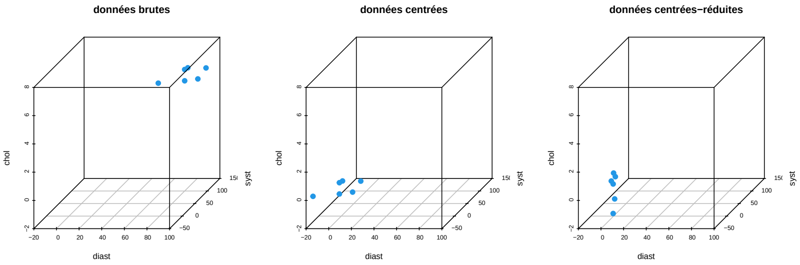
$$
\begin{array}{c|ccc|}
 & 1 \ldots & j & \ldots p \\
\hline
1 & & & \\
\vdots & & \vdots & \\
i & \ldots & y_{ij} = x_{ij} - \bar{x}^j & \ldots \\
\vdots & & \vdots & \\
n & & & \\
\hline\hline
\bar{\mathbf{y}} & \ldots & 0 & \ldots
\end{array}
$$

Matrice **Z** des données centrées-réduites

$$
\begin{array}{c|ccc|}
 & 1 \ldots & j & \ldots p \\
\hline
1 & & & \\
\vdots & & \vdots & \\
i & \ldots & z_{ij} = \frac{x_{ij} - \bar{x}^j}{s_j} & \ldots \\
\vdots & & \vdots & \\
n & & & \\
\hline\hline
\bar{\mathbf{z}} & \ldots & 0 & \ldots \\
s & \ldots & 1 & \ldots
\end{array}
$$

- Centrer les données permet d'avoir des colonnes (variables) de moyenne nulle.
- Réduire les données permet d'avoir des colonnes (variables) de variance 1.

Exemple : trois nuages de points des 6 patients.



- Centrer les données ne modifie pas les distances Euclidiennes entre les individus.
- Centrer-réduire les données les modifient.

En effet :

- Distance Euclidienne (au carré) entre deux individus i et i' décrits par les deux lignes $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ de $\mathbf{X}$:

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

- Distance Euclidienne (au carré) entre deux individus i et i' décrits par les deux lignes $\mathbf{z}_i$ et $\mathbf{z}_{i'}$ de $\mathbf{Z}$:

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

Distance Euclidienne entre Brigitte et Marie

Données brutes (**X**) :

	diast	syst	chol
Brigitte	90	140	6.0
Marie	60	85	5.9

Données centrées-réduites (**Z**)

	diast	syst	chol
Brigitte	1.5	0.48	0.78
Marie	-1.5	-2.16	0.52

Ecart-types :

	diast	syst	chol
moy	75	130	5.70
sd	10	21	0.38

Avec les données brutes **X** (ou centrées **Y**) :

$$\begin{aligned} d^2(Brigitte, Marie) &= (90-60)^2 + (140-85)2 + (6-5.9)^2 \\ &= 30^2 + 55^2 + 0.1^2 \end{aligned}$$

Avec les données standardisées **Z** :

$$\begin{aligned} d^2(Brigitte, Marie) &= (1.5+1.5)^2 + (0.48+2.16)^2 + (0.78-0.52)^2 \\ &= \frac{1}{10^2}(90-60)^2 + \frac{1}{20.8^2}(140-85)^2 + \frac{1}{0.383^2}(6-5.9)^2 \end{aligned}$$

A SAVOIR :

- 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.
- Vocabulaire : lorsqu'on centre-réduite les données, on dit aussi souvent que les normalise ou qu'on les standardise (même si il existe plusieurs autres manières de la faire).

En ACP on peut analyser :

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

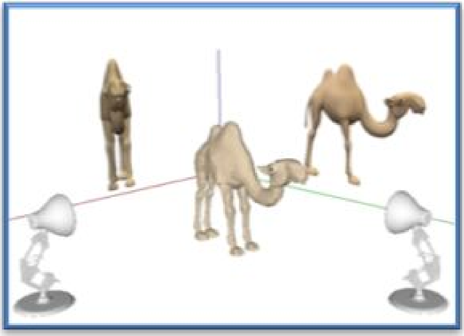
On distingue alors deux type d'ACP :

- 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.

Projection du nuage des individus

Trouver le sous-espace sur lequel le nuage des individus se projette avec la plus grande dispersion (variabilité) possible.



Ici, la dispersion d'un nuage de points est mesurée par son inertie.

L'inertie d'un nuage de n points $\mathbf{x}_i \in \mathbb{R}^p$ (n lignes d'une matrice $\mathbf{X}$) pondérés par w_i est définie ici par :

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

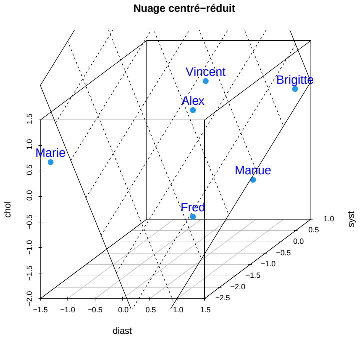
où $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})^T$, $\bar{\mathbf{x}} = (\bar{x}^1, \ldots, \bar{x}^p)^T$ et $\sum_{i=1}^n w_i = 1$.

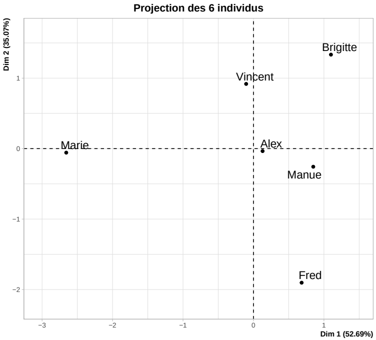
- Dans ce cours tous les individus auront le même poids $w_i = \frac{1}{n}, \forall i = 1, \ldots, n$.
- L'inertie totale du nuage des individus (données brutes) se réecrit :

  $$I(\mathbf{X}) = \sum_{j=1}^{p} s_j^2.$$

  où s_j^2 est la variance empirique de la variable j.
- Lorsque les données sont centrées-réduites, l'inertie totale est égale au nombre de variables : $I(\mathbf{Z}) = p$.

Comment trouver le premier plan de projection des individus ?





- Objectif : conserver au mieux les distances entre les individus (et donc la variabilité i.e. l'inertie du nuage de points).
- Solution : touver les axes sur lesquels les coordonnées des points projetés sont de variance maximum.

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.

En ACP, Δ_1 est l'axe de vecteur directeur $\mathbf{v}_1 \in \mathbb{R}^p$ qui maximise la variance des coordonnées des n individus projetés :

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

où $\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ est la matrice $p \times p$ des corrélations de terme général :

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

où $r_{jj'} = cor(\mathbf{x}^j, bx^{j'})$ est la corrélation entre les deux variables $\mathbf{x}^j$ et $\mathbf{x}^{j'}$.

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 coordonnées des n individus projetés :

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

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 : $< \mathbf{f}^1, \mathbf{f}^2 >= 0$.

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.

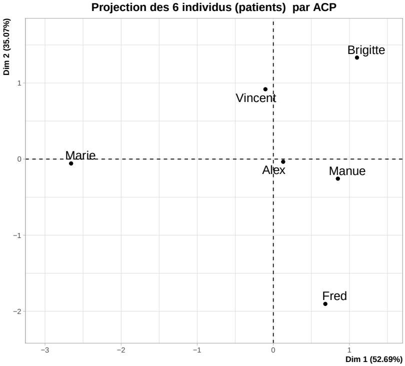
En résumé :

1. On effectue la décomposition en valeurs propres de la matrice des corrélations $\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ et on choisit q.
2. On calcule la matrice $\mathbf{F} = \mathbf{ZV}$ des q composantes 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 de 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.

$$
\mathbf{F} = \begin{array}{c|ccc|}
 & 1 \ldots & \alpha & \ldots q \\
\hline
1 & & & \\
\vdots & & \vdots & \\
i & \ldots & f_{i\alpha} & \ldots \\
\vdots & & \vdots & \\
n & & & \\
\hline\hline
\text{moy} & \ldots & 0 & \ldots \\
\text{var} & \ldots & \lambda_\alpha & \ldots
\end{array}
$$

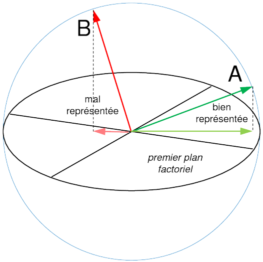
Exemple des 6 patients : matrice **F** des $q = 2$ premières CP

	f1	f2
Brigitte	1.10	1.33
Marie	-2.66	-0.06
Vincent	-0.10	0.92
Alex	0.13	-0.04
Manue	0.85	-0.26
Fred	0.68	-1.90



Plan

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



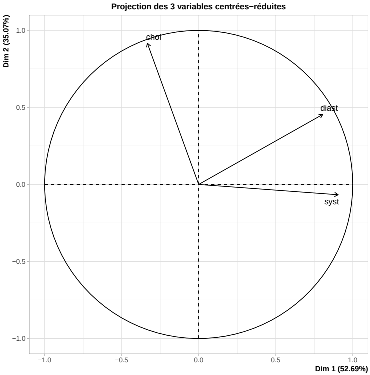
Exemple : les 3 colonnes de $\mathbf{Z}$ sont trois points sur la boule unité de $\mathbb{R}^6$.

Données standardisées $\mathbf{Z}$:

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

Martrice des corrélations $\mathbf{R}$:

	diast	syst	chol
diast	1.0	0.5	0.1
syst	0.5	1.0	-0.3
chol	0.1	-0.3	1.0



L'objectif est de trouver un plan sur lequel l'angle entre les variables projetées est le plus proche possible de l'angle entre entre les variables dans $\mathbb{R}^n$.

En effet, la corrélation entre deux variables $\mathbf{x}^j$ et $\mathbf{x}^{j'}$ s'interprète (en utilisant a métrique $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$) comme le cosinus de l'angle entre les deux variables $\mathbf{z}^j$ et $\mathbf{z}^{j'}$ noté $\theta(\mathbf{z}^j, \mathbf{z}^{j'})$:

$$\cos_{\mathbf{N}} \theta(\mathbf{z}^j, \mathbf{z}^{j'}) = \frac{< \mathbf{z}^j, \mathbf{z}^{j'} >_{\mathbf{N}}}{\|\mathbf{z}^j\|_{\mathbf{N}} \|\mathbf{z}^{j'}\|_{\mathbf{N}}} = < \mathbf{z}^j, \mathbf{z}^{j'} >_{\mathbf{N}} = cor(\mathbf{x}^j, \mathbf{x}^{j'}).$$

A SAVOIR :

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

Projection d'une variable standardisée (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 :

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

Remarque : on a muni ici $\mathbb{R}^n$ de la métrique $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$ car tous les individus ont un poids de $\frac{1}{n}$.

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.

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

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

- $\mathbf{u}_1$ est le vecteur propre associé à la plus grande valeur propre de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$,
- la plus première 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 $\mathbf{u}_1$:

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

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 :

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

On peut montrer que :

- $\mathbf{u}_2$ est le vecteur propre associé à la deuxième valeur propre de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$.
- la deuxième valeur propre de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ est aussi la deuxième 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 $\mathbf{u}_2$:

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

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 standardisées.

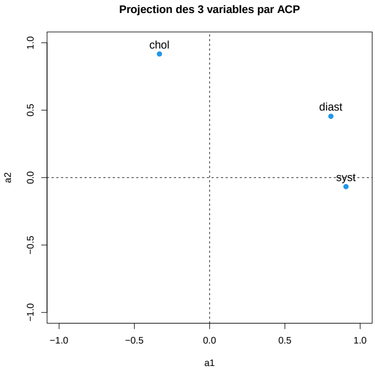
En résumé :

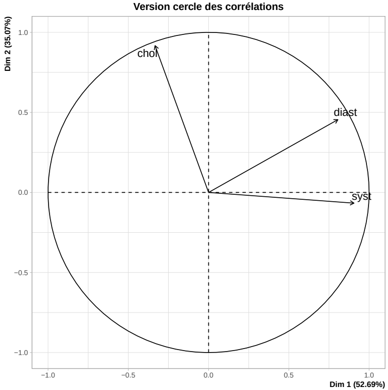
1. On effectue la décomposition en valeurs propres de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ et on choisit q.
2. On calcule la matrice $\mathbf{A} = \mathbf{Z}^T\mathbf{N}\mathbf{U}$ à partir de la matrice $\mathbf{U}$ des q premiers vecteurs propres de $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$.
   - Les colonnes $\mathbf{a}^\alpha = \mathbf{Z}^T\mathbf{N}\mathbf{u}_\alpha$ 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.

$$\mathbf{A}= \begin{array}{c|ccc|}
 & 1\ldots & \alpha & \ldots q \\ \hline
1 & & & \\
\vdots & & \vdots & \\
i & \ldots & a_{i\alpha} & \ldots \\
\vdots & & \vdots & \\
p & & & \\ \hline\hline
\text{norme} & \ldots & \sqrt{\lambda_\alpha} & \ldots
\end{array}$$

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
```





A SAVOIR :

On peut montrer que les coordonnées factorielles des variables (les loadings) sont aussi les corrélations entre les variables et les composantes principales :

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

Cette relation sera fondamentale pour l'interprétation des résultats en ACP.

Plan

Interprétation des résultats

```r
load("chol.rda")
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
```

```r
qr(cor(X))$rank
```

```
## [1] 3
```

```r
library(FactoMineR)
res <- PCA(X, graph = FALSE)
F <- res$ind$coord
F
```

```
##          Dim.1  Dim.2 Dim.3
## Brigitte  1.10  1.334 -0.33
## Marie    -2.66 -0.057 -0.36
## Vincent  -0.10  0.918  0.54
## Alex      0.13 -0.035  0.91
## Manue     0.85 -0.257 -0.91
## Fred      0.68 -1.903  0.15
```

```r
# r=3 composantes principales
```

```r
# Valeurs propres
res$eig[,1, drop=FALSE]
```

```
##        eigenvalue
## comp 1       1.58
## comp 2       1.05
## comp 3       0.37
```

```r
# Variances
apply(F, 2, var)*5/6
```

```
## Dim.1 Dim.2 Dim.3
##  1.58  1.05  0.37
```

```r
# Inertie des 3 CP
sum(res$eig[,1])
```

```
## [1] 3
```

```r
# On retrouve I(Z)=p=3
```

```r
# CP non corrélées
round(cor(F))
```

```
##       Dim.1 Dim.2 Dim.3
## Dim.1     1     0     0
## Dim.2     0     1     0
## Dim.3     0     0     1
```

```r
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
```

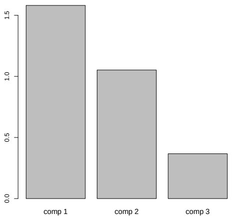
- $r = 3$ valeurs propres non nulles car $r = \min(n-1, p) = 3$,
- La somme des valeurs propres vaut $p = 3$ (l'inertie totale),
- $\frac{1.58}{3} = 53\%$ de l'inertie est expliquée par la première CP.
- 88 % de l'inertie est expliquée par les deux premières CP.
- 100 % de l'inertie est expliquée par 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 propres (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.

```r
# Eboulis des valeurs propres
barplot(res$eig[,1], cex.names=1.2)
```



- 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 perd que 12% de l'information (l'inertie) de départ.

Interprétation des plans factoriels des individus.

Règle : si deux individus sont bien projetés, alors la distance entre ces deux individus sur le plan factoriel est proche de leur vraie 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 $(\Delta_\alpha, \Delta_{\alpha'})$:

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

- Un individu bien projeté a un $\cos^2$ proche de 1.

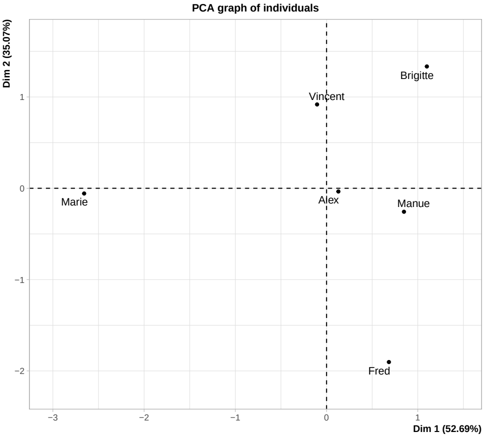
Exemple des 6 patients.

```r
res$ind$cos2[,1:2]

##            Dim.1   Dim.2
## 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

# cos2 avec les axes 1 et 2
```

```r
apply(res$ind$cos2[1:3, 1:2], 1, sum)

## Brigitte    Marie  Vincent
##     0.97     0.98     0.74

apply(res$ind$cos2[4:6,1:2], 1, sum)

##  Alex Manue  Fred
## 0.022 0.487 0.994

# cos2 avec le 1er plan factoriel
```

Alex est mal projeté (au centre du plan factoriel).

```r
plot(res, choix = "ind")
```



Interprétation ?

Interprétation des plans factoriels des variables.

Règle : si deux variables sont bien projetées, alors l'angle entre ces deux variables sur le plan factoriel est proche de leur vrai angle dans $\mathbb{R}^n$ et donne donc une idée de leur corrélation.

- 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 :

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

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

$$\cos^2_{\mathbf{N}}(\theta_{j(\alpha,\alpha')}) = a^2_{j\alpha} + a^2_{j\alpha'}.$$

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

- Une variable bien projetée a un $\cos^2$ proche de 1 et donc une flèche proche du cercle.

Exemple des 3 variables `diast`, `syst` et `chol`.

```r
res$var$cos2[,1:2]
```

```
##       Dim.1  Dim.2
## diast  0.65 0.2068
## syst   0.82 0.0045
## chol   0.11 0.8410
```

```r
# cos2 avec les axes 1 et 2
```

```r
apply(res$var$cos2[, 1:2], 1, sum)
```

```
## diast  syst  chol
##  0.86  0.82  0.95
```

```r
# cos2 avec le 1er plan factoriel
```

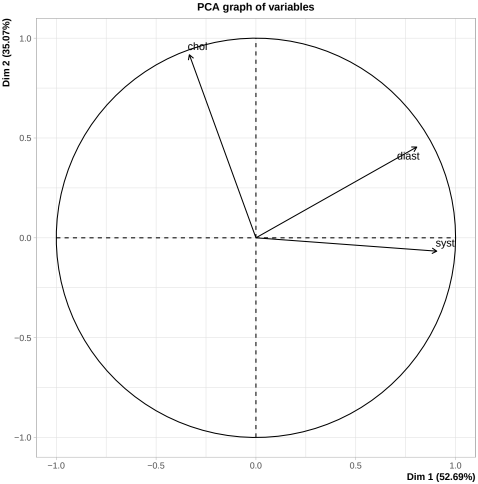
```r
sqrt(apply(res$var$cos2[, 1:2], 1, sum))
```

```
## diast  syst  chol
##  0.93  0.91  0.98
```

```r
# longueur des 3 flèches
```

Les trois variables `diast`, `syst` et `chol` sont bien projetées sur le premier plan factoriel. On peut donc interpréter le cosinus de leur angle comme une approximation correcte de leur corrélation.

```r
plot(res, choix="var")
```



Interprétation ?

Interprétation du plan factoriel des individus à partir de celui des variables.

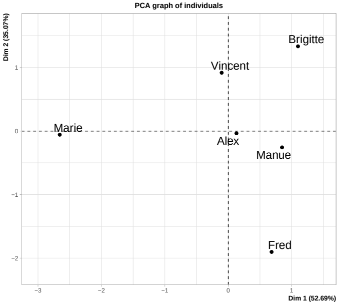
Règle : la coordonnée d'une variable sur un axe α est égale à sa corrélation avec la α-ème composante principale :

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

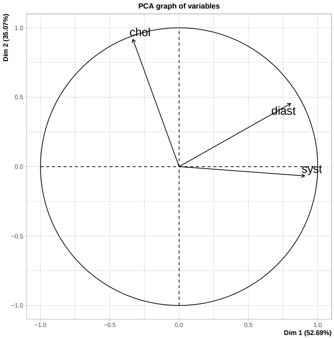
- Si la coordonnées $a_{j\alpha}$ de la variable j sur l'axe α est proche de 1 (resp. -1) , cette variable est corrélée positivement (resp. négativement) à la α-ème composante principale.
- Les individus ayant des coordonnées plus grandes (resp. plus petites) que la moyenne sur cet axe ont donc des valeurs plus grandes (resp. plus petite) que la moyenne sur la variable j.
- Les composantes principales étant centrées (moyenne nulle), leur position à droite, à gauche, en haut et en bas donne une indiquation de leurs valeurs sur les variables.

```
res$var$coord

##       Dim.1  Dim.2 Dim.3
## diast  0.81  0.455 -0.38
## syst   0.91 -0.067  0.42
## chol  -0.33  0.917  0.22

# variables diast et syst corrélées positivement à la première CP
# variable chol corrélée positivement à la deuxième CP
```

```r
plot(res, choix = "ind", cex=1.5)
```




```r
plot(res, choix = "var", cex=1.5)
```



Comment différentier les individus à doite, à gauche, en haut et en bas en fonction des variables `diast`, `syst` et `chol` ?

Plan

Et si les données ne sont pas quantitative ?

Méthode	Nature des données	Exemple de package R
ACP	Quantitatives	FactoMineR
ACM	Qualitatives	FactoMineR
ACPmixte	Mixtes	PCAmixdata

Vocabulaire :

- ACM = Analyse des Correspondances Multiples,
- MCA = Multiple Correspondance Analysis.

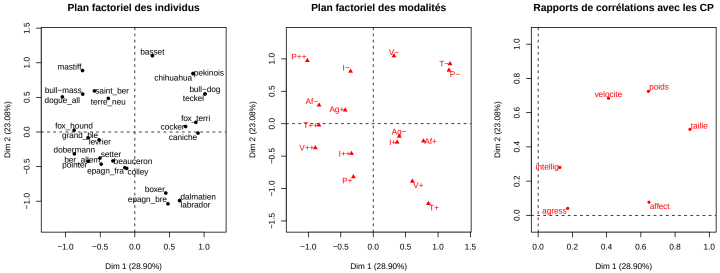
Exemple des races de chiens.

```
load("chiens.rda")
chiens
```

```
##            taille poids velocite intellig affect agress
## beauceron     T++    P+      V++       I+    Af+    Ag+
## basset         T-    P-       V-       I-    Af-    Ag+
## ber_allem     T++    P+      V++      I++    Af+    Ag+
## boxer          T+    P+       V+       I+    Af+    Ag+
## bull-dog       T-    P-       V-       I+    Af+    Ag-
## bull-mass     T++   P++       V-      I++    Af-    Ag+
## caniche        T-    P-       V+      I++    Af+    Ag-
## chihuahua      T-    P-       V-       I-    Af+    Ag-
## cocker         T+    P-       V-       I+    Af+    Ag+
## colley        T++    P+      V++       I+    Af+    Ag-
## dalmatien      T+    P+       V+       I+    Af+    Ag-
## dobermann     T++    P+      V++      I++    Af-    Ag+
## dogue_all     T++   P++      V++       I-    Af-    Ag+
## epagn_bre      T+    P+       V+      I++    Af+    Ag-
## epagn_fra     T++    P+       V+       I+    Af-    Ag-
## fox_hound     T++    P+      V++       I-    Af-    Ag+
## fox_terri      T-    P-       V+       I+    Af+    Ag+
## grand_ble     T++    P+       V+       I-    Af-    Ag+
## labrador       T+    P+       V+       I+    Af+    Ag-
## levrier       T++    P+      V++       I-    Af-    Ag-
## mastiff       T++   P++       V-       I-    Af-    Ag+
## pekinois       T-    P-       V-       I-    Af+    Ag-
## pointer       T++    P+      V++      I++    Af-    Ag-
## saint_ber     T++   P++       V-       I+    Af-    Ag+
## setter        T++    P+      V++       I+    Af-    Ag-
## teckel         T-    P-       V-       I+    Af+    Ag-
## terre_neu     T++   P++       V-       I+    Af-    Ag-
```

$n = 27$ chiens et $p = 6$ variables qualitatives à 2 ou 3 modalités.

```r
library(FactoMineR)
res <- MCA(chiens, graph = FALSE)
par(mfrow=c(1,3))
plot(res, choix="ind", invisible = "var", graph.type = "classic", title = "Plan factoriel des individus")
plot(res, choix="ind", invisible = "ind", graph.type = "classic", title = "Plan factoriel des modalités")
plot(res, choix="var", graph.type = "classic", title = "Rapports de corrélations avec les CP")
```



A SAVOIR :

- Le rapport de corrélation est une mesure de liaison entre une variable quantitative et une variable qualitative.
- Il varie entre 0 et 1 et donne la part de la variance de la variable quantitative expliquée par les modalités de la variable qualitative
- Il ne doit pas être confondu avec la corrélation linéaire entre deux variables quantitatives.

ATTENTION : les pourcentages d'inertie expliquée par les axes sont toujours petits en ACM.

A SAVOIR : l'ACM permet de recoder des données qualitatives en données quantitatives.

```
# Recodage des 6 variables qualitatice par les 4 premières CP qui sont numériques
F <- res$ind$coord[, 1:4]
F
```

```
##           Dim 1  Dim 2  Dim 3  Dim 4
## beauceron -0.32 -0.418 -0.101 -0.211
## basset     0.25  1.101 -0.191  0.293
## ber_allem -0.49 -0.464 -0.498  0.577
## boxer      0.45 -0.882  0.692  0.260
## bull-dog   1.01  0.550 -0.163 -0.350
## bull-mass -0.75  0.547  0.498  0.655
## caniche    0.91 -0.016 -0.577  0.628
## chihuahua  0.84  0.844 -0.470 -0.086
## cocker     0.73  0.079  0.662  0.190
## colley    -0.12 -0.526 -0.335 -0.658
## dalmatien  0.65 -0.990  0.459 -0.186
## dobermann -0.87 -0.315 -0.452  0.510
## dogue_all -1.05  0.507  0.165  0.063
## epagn_bre  0.48 -1.037  0.062  0.603
## epagn_fra -0.14 -0.516  0.117 -0.469
## fox_hound -0.88  0.025 -0.362 -0.015
## fox_terri  0.88  0.139  0.054  0.286
## grand_ble -0.52 -0.113  0.044  0.241
## labrador   0.65 -0.990  0.459 -0.186
## levrier   -0.68 -0.083 -0.596 -0.462
## mastiff   -0.76  0.888  0.588  0.130
## pekinois   0.84  0.844 -0.470 -0.086
## pointer   -0.67 -0.424 -0.686  0.064
## saint_ber -0.58  0.594  0.894 -0.134
## setter    -0.50 -0.377 -0.289 -0.725
## teckel     1.01  0.550 -0.163 -0.350
## terre_neu -0.38  0.485  0.661 -0.580
```

Encore d'autres méthodes...

Linéaires :

Méthode	Particularité	Exemple de package R
AFC	tableau de contingence	FactoMineR
AFM	tableaux multiples	FactoMineR
ACP sparse	sélection de variables	elasticnet

- AFC = Analyse Factoriel des Correspondance (CA en englais),
- AFM = Multiple Factoriel Multiple (MFA en englais).

Non linéaires :

Méthode	Particularité	Fonctions R
MDS	Matrice de dissimilarités	cmdscale (STAT), isoMDS (MASS)
NLPCA	AutoEncoder	nlpca (pcaMethods)
t-SNE	Embeddings	tsne (M3C)

- MDS = MultiDimentional Scaling,
- NLPCA = Non Linear PCA,
- t-SNE = t-distributed stochastic neighbor embedding

Et encore d'autres....