



# Data exploration

## Analyse en Composantes Principales

Observer simultanément des individus d'une population sur  $p > 2$  variables

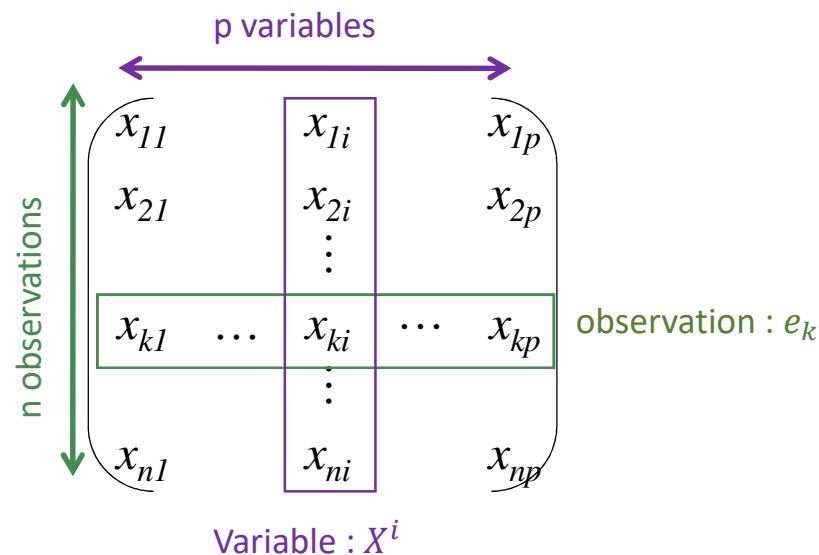
- Etudier le lien entre les variables
- Faire une représentation graphique d'un nuage de points à  $p$  dimensions
- Qualifier les observations du jeu de donnée



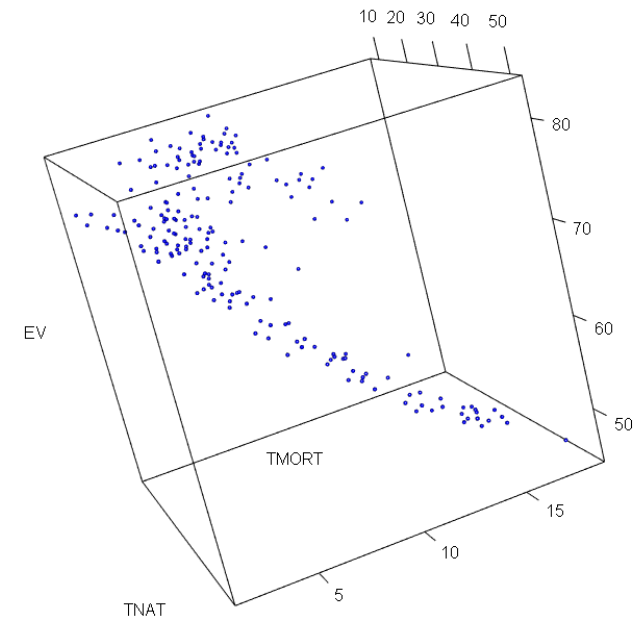
## ACP

# Nuage de points en dimension $p$

Un jeu de données est un tableau avec en ligne les observations et en colonne les variables  $X_1, \dots, X_p$ . On peut donc représenter chaque ligne du tableau comme un point dont les coordonnées sont les valeurs prises par les variables.



Sur cet exemple chaque point est un pays caractérisé par son espérance de vie, et ses taux de natalité et de mortalité.



Si  $p > 3$ , il est impossible de visualiser le nuage de points. L'objectif de l'ACP est de trouver une projection du nuage de points en dimension 2 ou 3 de façon à perdre le moins possible d'information.

- Qu'est-ce que l'information?
- Si projection alors produit scalaire?
- Quel lien avec les variables  $X_1, \dots, X_p$ ?





## ACP

# Comment mesurer l'information?

## Centre de gravité

Le centre de gravité est le point dont les coordonnées sont définies par les valeurs moyennes des variables,

$$G = (\bar{X}_1, \dots, \bar{X}_p)$$

## Inertie

L'information contenue dans un nuage de points correspond à l'inertie de celui-ci, c.-à-d. la somme des distances au carré entre les observations et le centre de gravité du nuage,

$$I = \sum_{k=1}^n \|e_k - G\|^2$$

où  $\|.$  désigne la norme euclidienne.

## Propriétés de l'inertie

L'inertie peut s'exprimer comme la trace de la matrice de variance-covariance, c.-à-d. la somme des variances des variables,

$$I = \text{tr}(V) = \sum_{k=1}^p s_k^2 \quad \text{où} \quad V = \begin{pmatrix} s_1^2 & \cdots & c_{X_1 X_p} \\ \vdots & \ddots & \vdots \\ c_{X_1 X_p} & \cdots & s_p^2 \end{pmatrix}.$$

La matrice de variance-covariance est symétrique définie positive.

L'inertie peut aussi s'écrire comme la somme des valeurs propres de  $V$ ,

$$I = \lambda_1 + \cdots + \lambda_p$$



## ACP

# Centrer et réduire les variables

L'analyse en composantes principales nécessite de calculer des distances entre observations,

$$\|e_k - e_{k'}\|^2 = \sum_{i=1}^p (x_{ki} - x_{k'i})^2.$$

Si les variables n'ont pas le même ordre de grandeur, certaines variables à valeurs faibles « disparaîtrons » de l'information au profit de celles ayant de fortes valeurs.

	Pop. (milliers)	Taux nat. (pour mille)	Esp. vie	Nb. enfants
Argentine	41050	16,87	75,87	2,19
Arménie	3099	15,47	74,44	1,77
Australie	21731	12,56	81,99	1,85
Autriche	8407	9,01	80,55	1,40

distance entre  
l'Argentine et  
l'Arménie

$$(41050-3099)^2 + (16,87-15,47)^2 + (75,87-74,44)^2 + (2,19-1,77)^2 = 1440278405$$

$$(41050-3099)^2 = 1440278401$$

Les variables Taux nat., Esp. vie et Nb. enfants ne comptent pas dans le calcul de la distance

De la même façon, la quantification de l'information au travers de l'inertie,  $I = \sum_{i=1}^p s_i^2$ , privilégie les variables fortement dispersés.

**Il est donc important de centrer et réduire les variables**

$$X_i \leftarrow \frac{X_i - \bar{X}_i}{s_i}, i = 1, \dots, p$$



## ACP

*Quel produit scalaire?*

Le produit scalaire entre deux variables  $X_i$  et  $X_j$  est défini par,

$$\langle X_i, X_j \rangle = \frac{1}{n} \sum_{k=1}^n x_{ki} x_{kj},$$

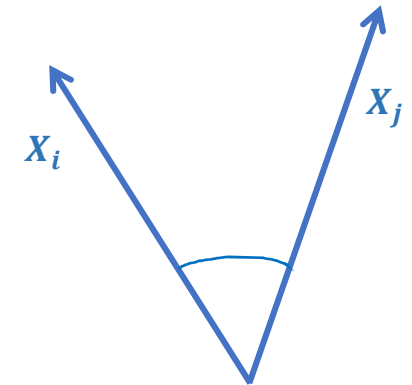
d'où la norme

$$\|X_i\|^2 = \frac{1}{n} \sum_{k=1}^n (x_{ki})^2$$

Si les variables sont centrées alors  $\langle X_i, X_j \rangle = c_{X_i X_j}$  et  $\|X_i\|^2 = s_i^2$

D'après la formule du cosinus,

$$\cos(\widehat{X_i, X_j}) = \frac{\langle X_i, X_j \rangle}{\|X_i\| \|X_j\|} = \frac{c_{X_i X_j}}{s_i s_j} = r_{X_i X_j}$$



- $|r_{X_i X_j}| = 1 \Leftrightarrow$  les variables sont colinéaires  
     corrélées positivement si  $r_{X_i X_j} = 1$   
     corrélées négativement si  $r_{X_i X_j} = -1$
- $r_{X_i X_j} = 0 \Leftrightarrow$  les variables sont orthogonales  
      $\Leftrightarrow$  les variables ne sont pas linéairement corrélées



# ACP

## Principe de l'ACP

Le principe de l'ACP est de trouver des espaces de petites dimensions sur lesquels les projections des observations minimisent la déformation de la réalité.

On cherche donc un sous-espace  $F_q$  de  $\mathbb{R}^p$  de dimension  $q$  ( $q=2,3,..$ ) sur lequel projeté le nuage de points. Les axes de ce sous-espace sont des combinaisons linéaires des axes d'origine (c.-à-d. les variables). Les nouveaux axes s'appellent les *composantes principales*.

### Principe

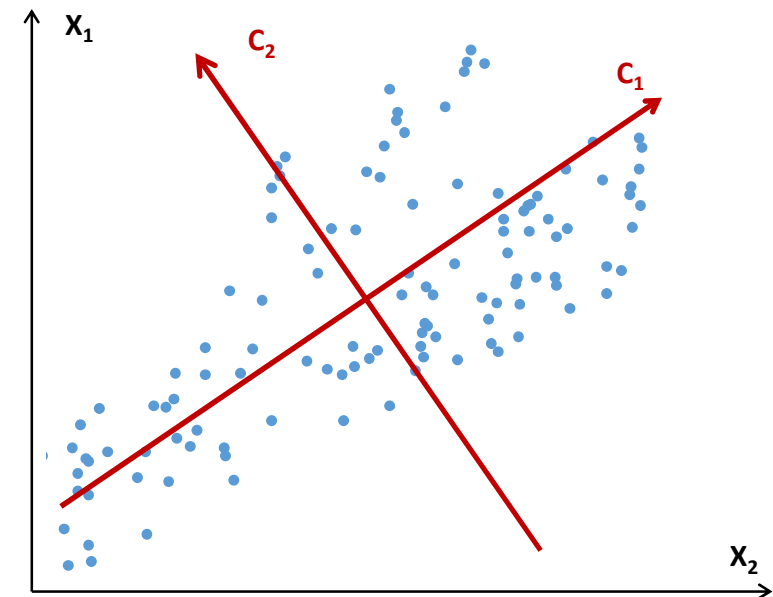
- La 1<sup>ère</sup> composante principale ( $C_1$ ) doit « capturer » le maximum d'information

*Il reste un résidu d'information non expliquée*

- La 2<sup>ème</sup> composante principale ( $C_2$ ) est calculée sur ce résidu telle que

- ✓ Elle capture un maximum d'information
- ✓ Elle soit non corrélée linéairement à  $C_1$  (orthogonalité)

- Sur le même principe, calcul de  $C_3, C_4, ..., C_p$



Il s'agit d'un changement de repère pour passer du repère initial formé par les variables à un repère orthogonale tel que les nouveaux axes sont ordonnés par quantité d'information décroissante.

**Nb. composantes principales = Nb. variables initiales**



## ACP

# Comment perdre le moins d'information possible?

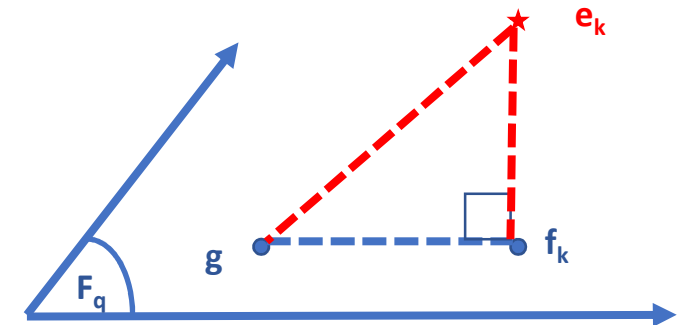
Soit  $e_k$  un point du nuage et notons  $f_k$  sa projection orthogonale sur le sous-espace  $F_q$ .

On cherche  $F_q$  tel que la distance entre  $F_q$  et les individus soit minimale.

Minimiser :  $\sum_{k=1}^n \|f_k - e_k\|^2$

D'après Pythagore,

$$\|f_k - e_k\|^2 = \|g - e_k\|^2 - \|f_k - g\|^2$$



Donc le problème revient à maximiser

$$\sum_{k=1}^n \|f_k - g\|^2 \quad \text{car } \|g - e_k\|^2 \text{ ne dépend pas de } F_q$$

autrement dit maximiser l'inertie du nuage projeté. On cherche à garder un maximum de dispersion dans la projection.



## ACP

# Solution du problème

Le sous-espace qui minimise l'inertie du nuage projeté est défini par :

$$F_q = \text{vect}(u_1, \dots, u_q)$$

où  $u_k$  est le **vecteur propre** unitaire de la matrice  ${}^tV$  associée à la  $k^{\text{ème}}$  plus grande valeur propre.

- ✓ L'inertie du nuage projeté sur  $u_k$  est  $\lambda_k$
- ✓ L'inertie du nuage projeté sur  $F_q$  est  $\lambda_1 + \dots + \lambda_q$
- ✓ L'inertie totale est  $I = \lambda_1 + \dots + \lambda_q$

Les vecteurs propres sont appelés les *axes principaux*

- ✓ Le premier axe principal  $u_1$  est associé à la plus grande valeur propre  $\lambda_1$
- ✓ Le deuxième axe principal  $u_2$  est associé à la deuxième valeur propre  $\lambda_2$
- ✓ Etc...

L'ACP est un changement de repère dans lequel les 1<sup>ers</sup> axes contiennent un maximum d'information

Cf. Exercice TD sur le calcul des coordonnées de  $X_k$  dans le nouveau repère

La projection des individus sur un axe principal est une nouvelle variable appelée *composante principale*

- ✓ La première composante  $c_1$  représente les coordonnées des projections des individus sur l'axe  $u_1$
- ✓ La deuxième composante  $c_2$  représente les coordonnées des projections des individus sur l'axe  $u_2$
- ✓ Etc...





ACP

## Combien d'axes retient-on?

Il y a deux règles pour le choix du nombre d'axes :

- garder un maximum d'information contenu dans ces axes (pourcentage cumulé d'inertie)
- couper sur le dernier grand saut d'information entre les axes (elbow rule)

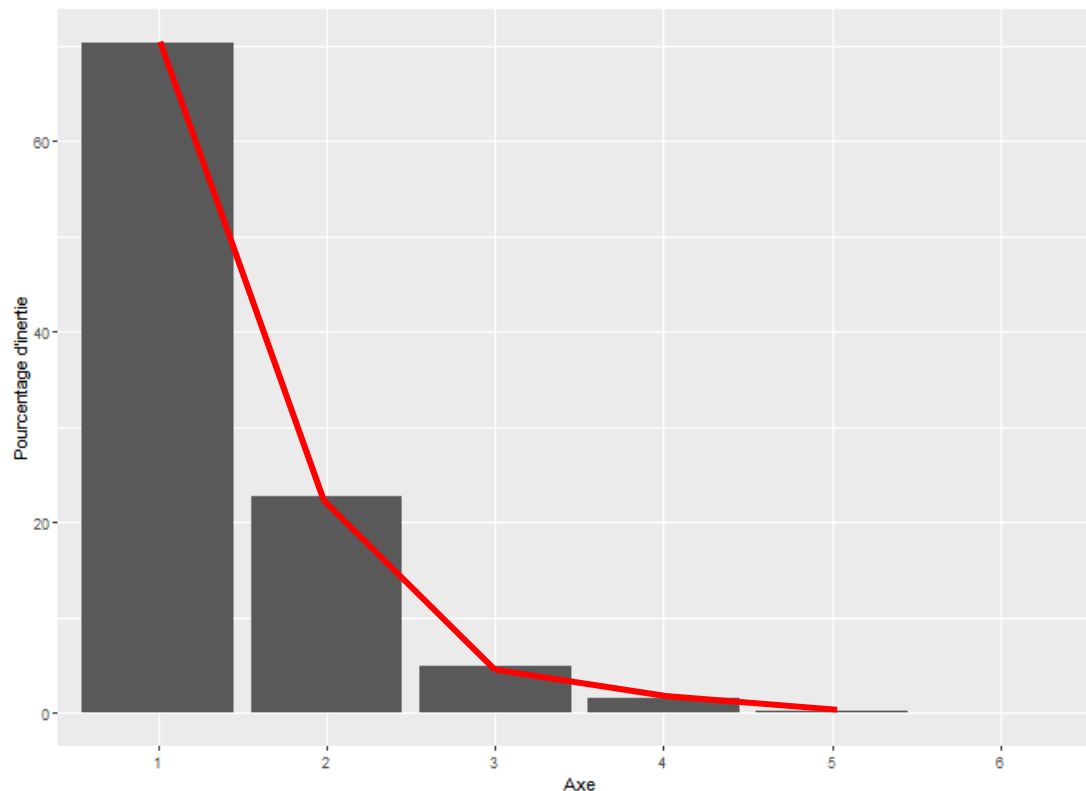


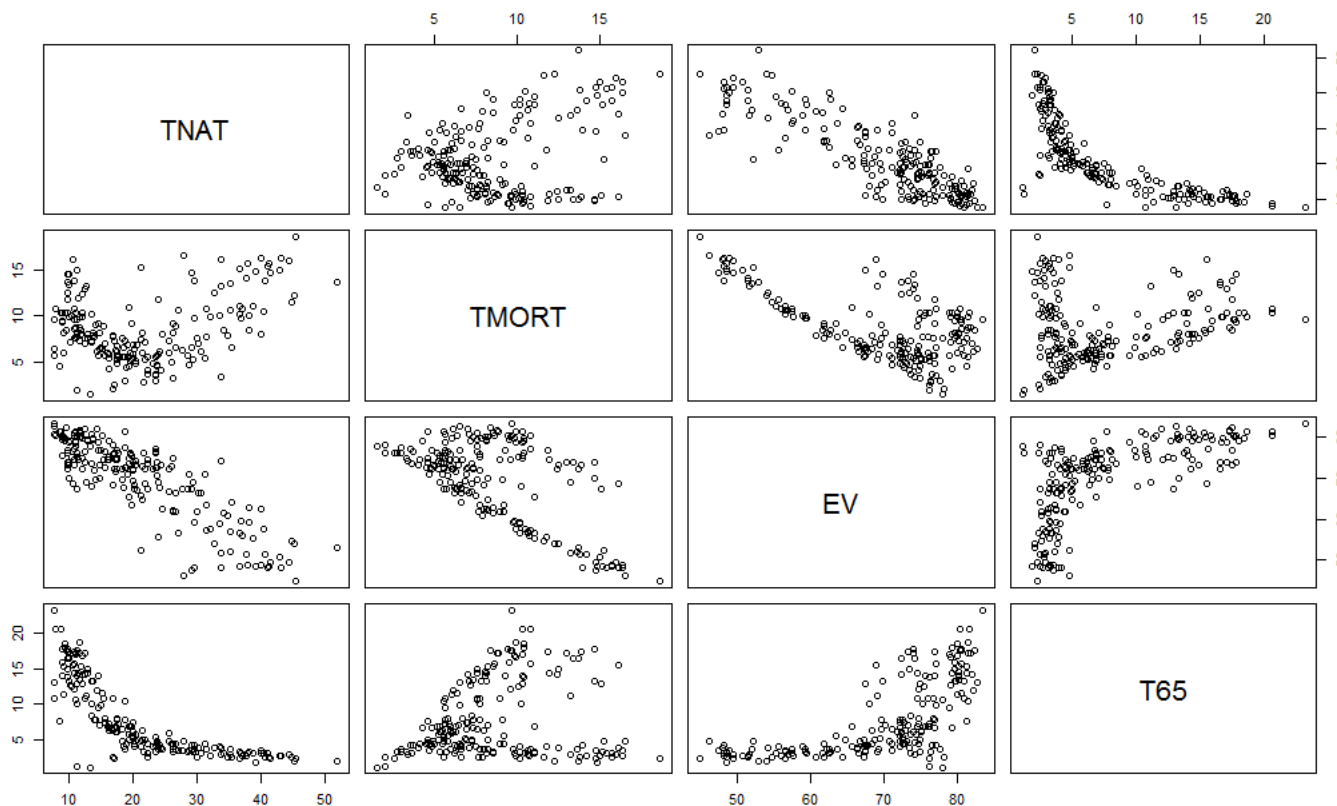
Tableau des valeurs propres

Axe	%	Cum. %
1	70.3	70.3
2	22.8	93.1
3	4.9	98.0
4	1.6	99.6
5	0.2	99.8
6	0.2	100.0



## ACP

# Exemple de la démographie mondiale



Pays caractérisés par 4 variables :

- TNAT : Taux de natalité
- TMORT : Taux de mortalité
- EV : Espérance de vie
- T65 : Taux >65 ans



## ACP

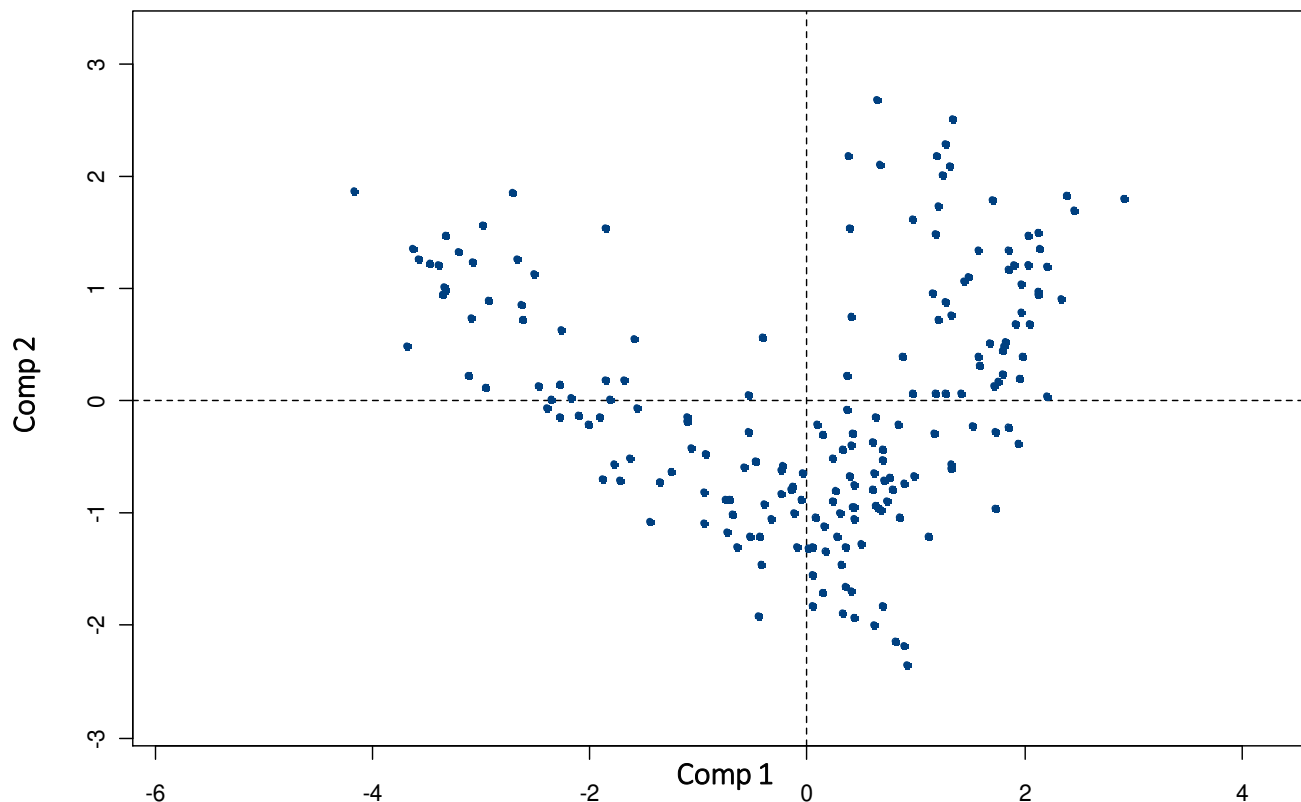
# Représentation des observations

	eigenvalue	percentage of variance	cumulative % of variance
comp 1	2.66302177	66.5755442	66.57554
comp 2	1.19799267	29.9498168	96.52536
comp 3	0.12720887	3.1802217	99.70558
comp 4	0.01177669	0.2944172	100.00000

**4 variables**  
(TNAT, TMORT, EV, T65)

⇔

**4 composantes principales**  
(combinaisons linéaires des 4 variables  
initiales)



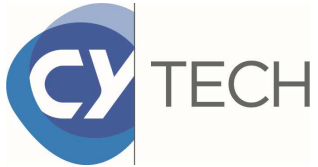
La 1<sup>ère</sup> composante principale contient 66,6% de l'inertie.

La 2<sup>ème</sup> composante principale contient 29,9% de l'inertie.

La représentation des pays sur le plan principal retranscrit 96,5% de l'information.

**Comment interpréter ce graphique?**

Comment qualifier un pays en haut à droite par exemple?



## ACP

# Représentation des variables

\$cor = coordonnée=corrélation

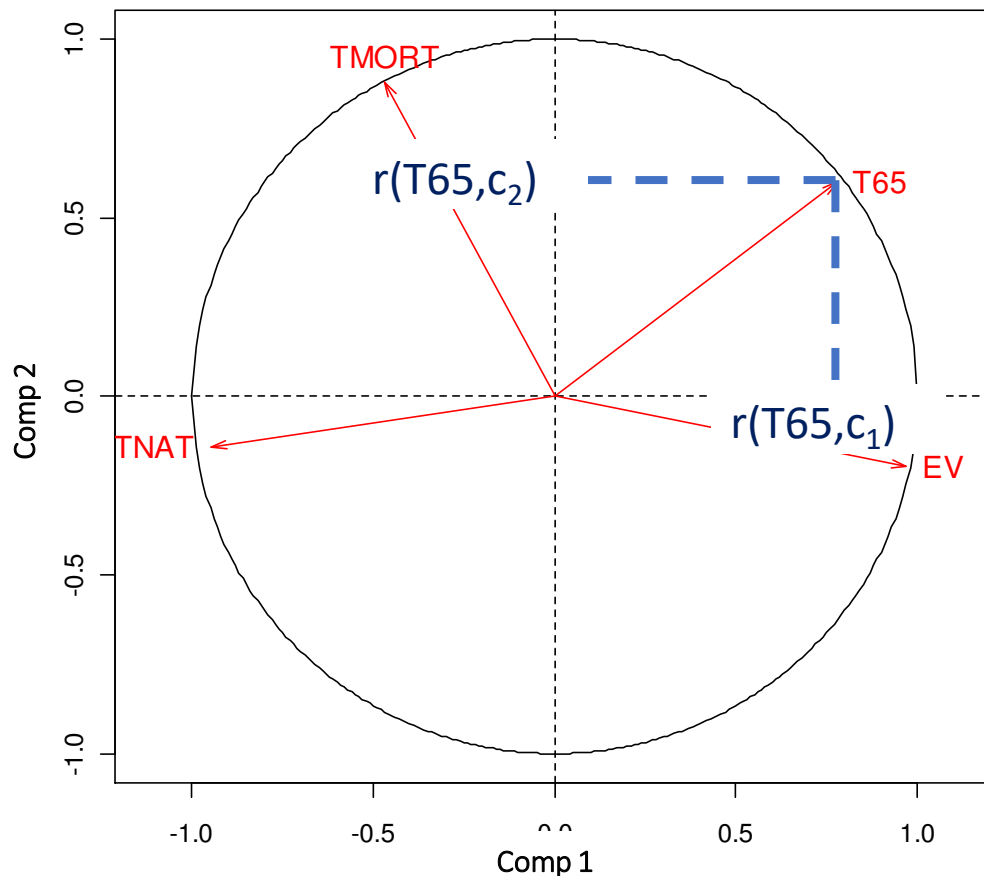
\$cos2 = répartition de la variable sur les 4 axes principaux

La variable TNAT est représentée à 89.82% sur  $C_1$  et 1.98% sur  $C_2$  etc..

\$contrib = contribution de la variable à la construction de l'axe

La variable TMORT ne contribue pas (8%) à la construction de  $c_1$ .

Représentation des variables



## \$cor

	Dim.1	Dim.2	Dim.3	Dim.4
TNAT	-0.9477642	-0.1409135	...	
TMORT	-0.4674138	0.8814408		
EV	0.9692397	-0.1966503		
T65	0.7790144	0.6021020		

## \$cos2

	Dim.1	Dim.2
TNAT	0.8982571	0.01985663
TMORT	0.2184757	0.77693792
EV	0.9394256	0.03867136
T65	0.6068634	0.36252677

## \$contrib

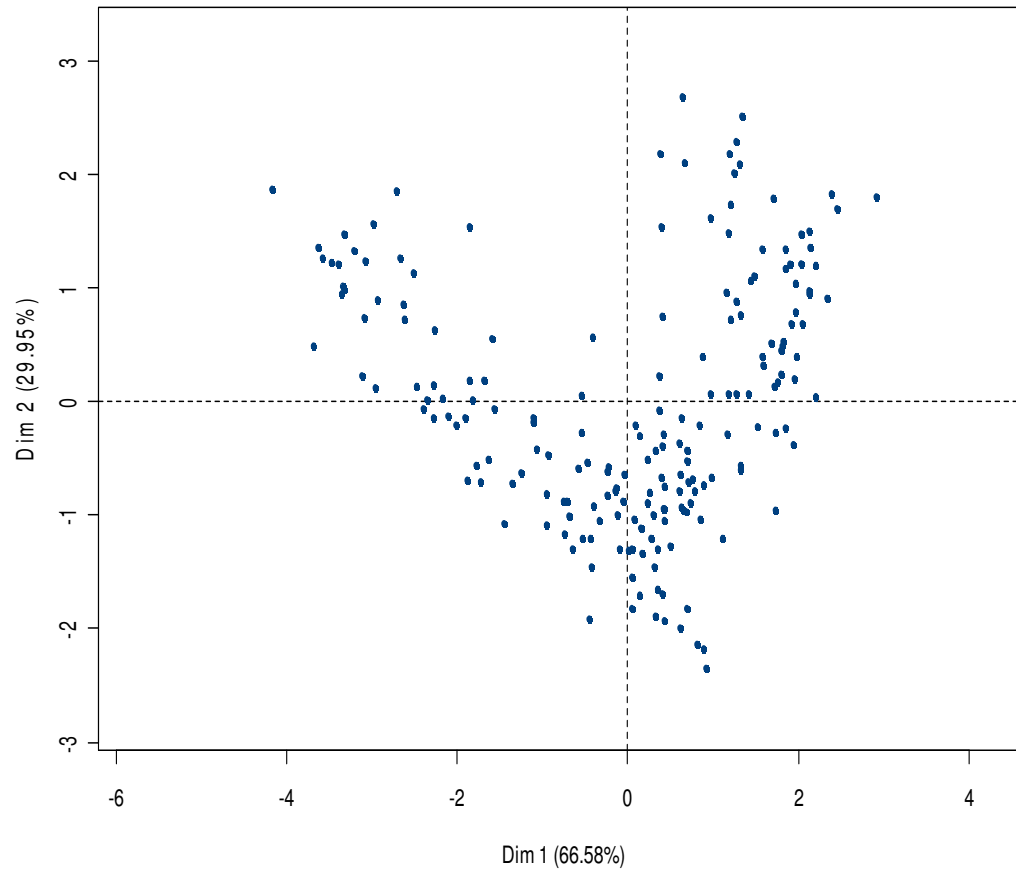
	Dim.1	Dim.2
TNAT	33.73074	1.657492
TMORT	8.20405	64.853311
EV	35.27668	3.228013
T65	22.78853	30.261184



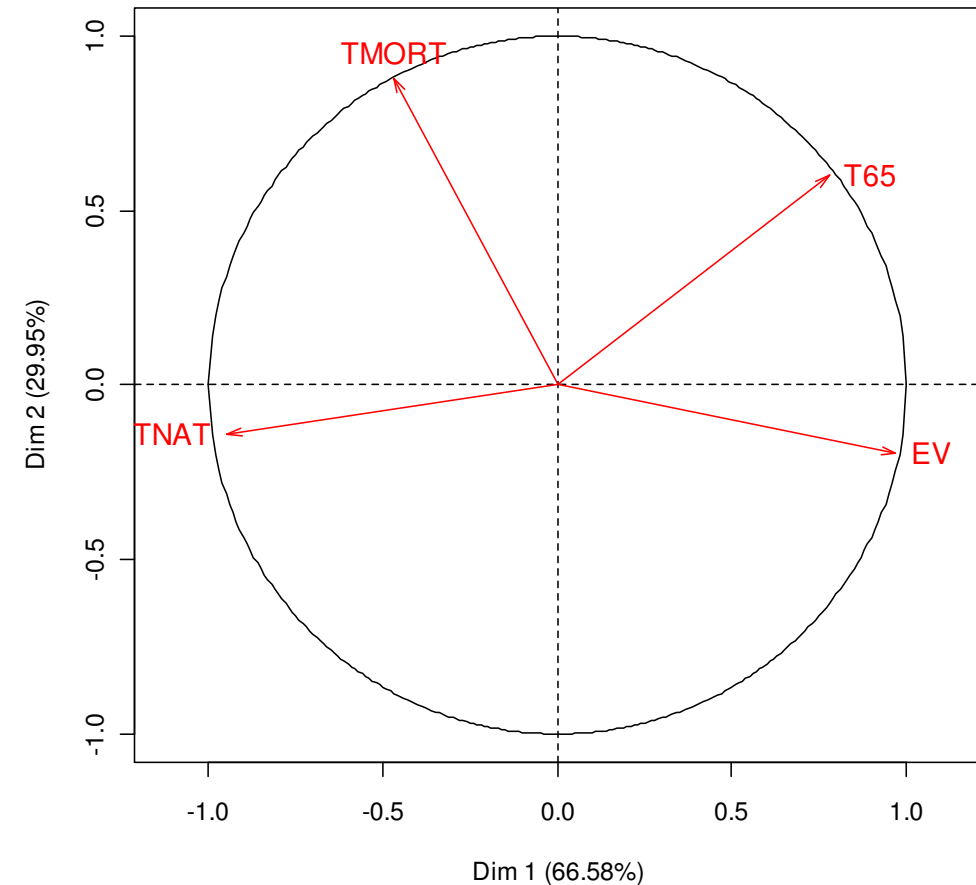
# ACP

## Interprétation

Représentation des individus



Représentation des variables



- TNAT et EV sont corrélés négativement  
les pays avec un fort taux de natalité ont une espérance de vie courte
- TMORT et T65 sont non corrélés



ACP

# Validité des représentations graphiques

- La projection perd le moins d'information possible

⇒ vérifier le % d'inertie expliquée par l'axe

⇒ conserver le nombre d'axes nécessaire pour avoir une inertie expliquée correcte

## Autre utilisation de l'ACP = réduire la dimension d'un problème

L'ACP est très souvent utilisée en amont de méthodes de *machine learning* pour réduire le nombre de variables. L'objectif n'est plus l'interprétation des données sur un graphique.

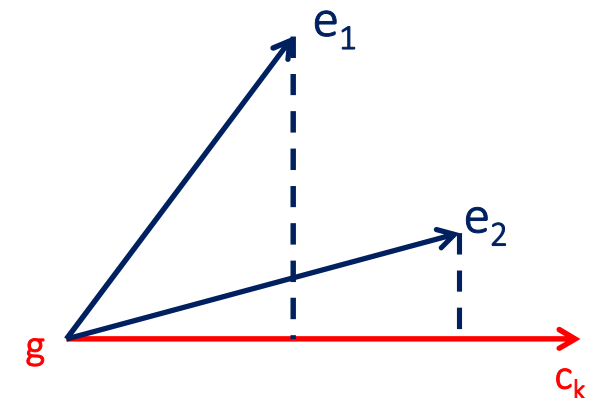
- Les variables sont bien représentées si elles sont proches du cercle. A contrario celles qui sont proches de l'origine sont peu corrélées avec les axes

⇒ pas d'interprétation possible pour ces variables

- Les individus sont bien représentés s'ils ne sont pas trop éloignés de l'axe sur lequel on les projette

⇒ vérifier le cosinus entre l'individu et l'axe (proche de 1)

⇒ valable si l'individu loin du centre de gravité



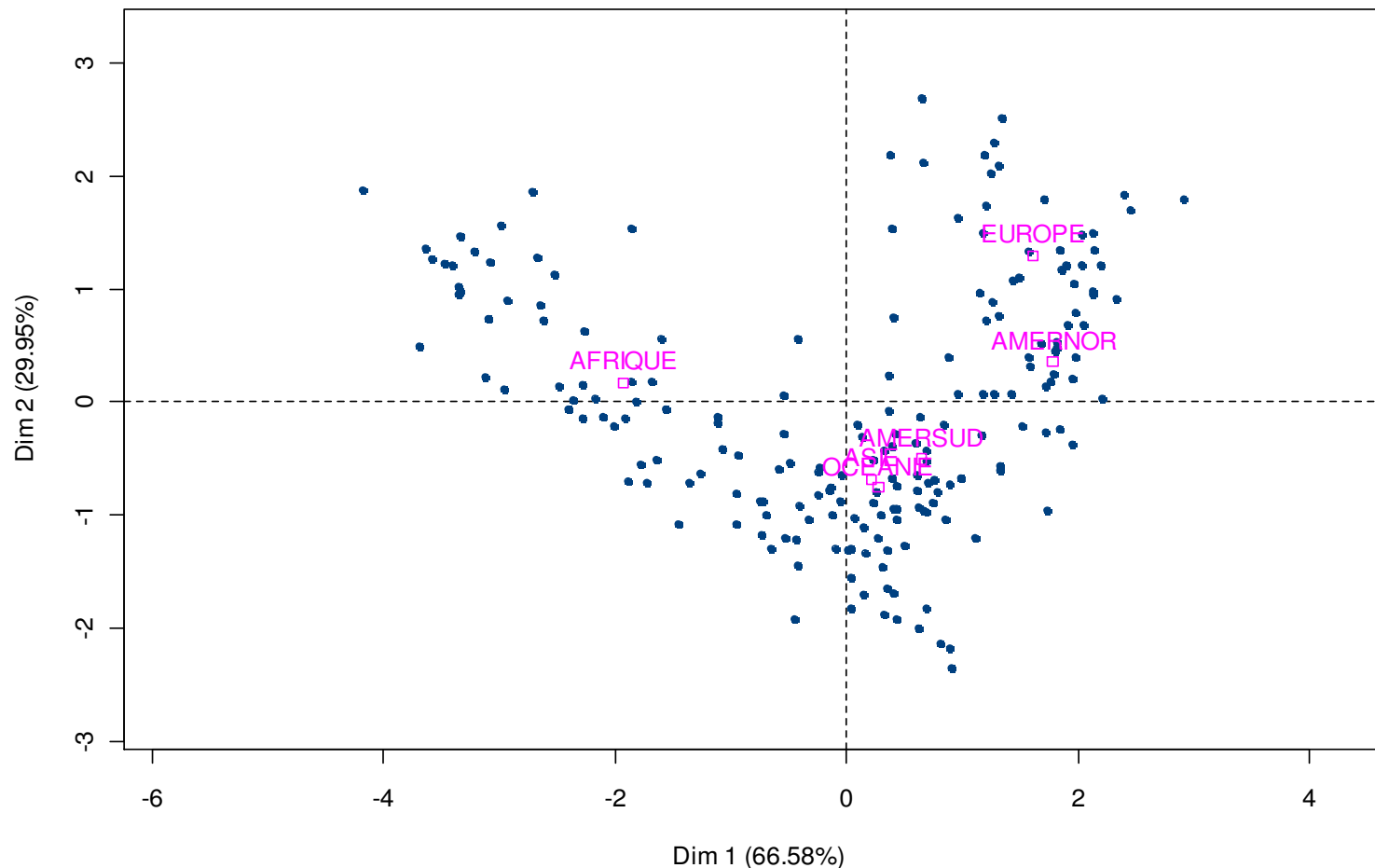


ACP

# Ajout de variable ou d'observation

Il est possible d'ajouter des individus ou des variables aux représentations graphiques.  
Ceux-ci ne participent pas à la construction des axes

Représentation des individus



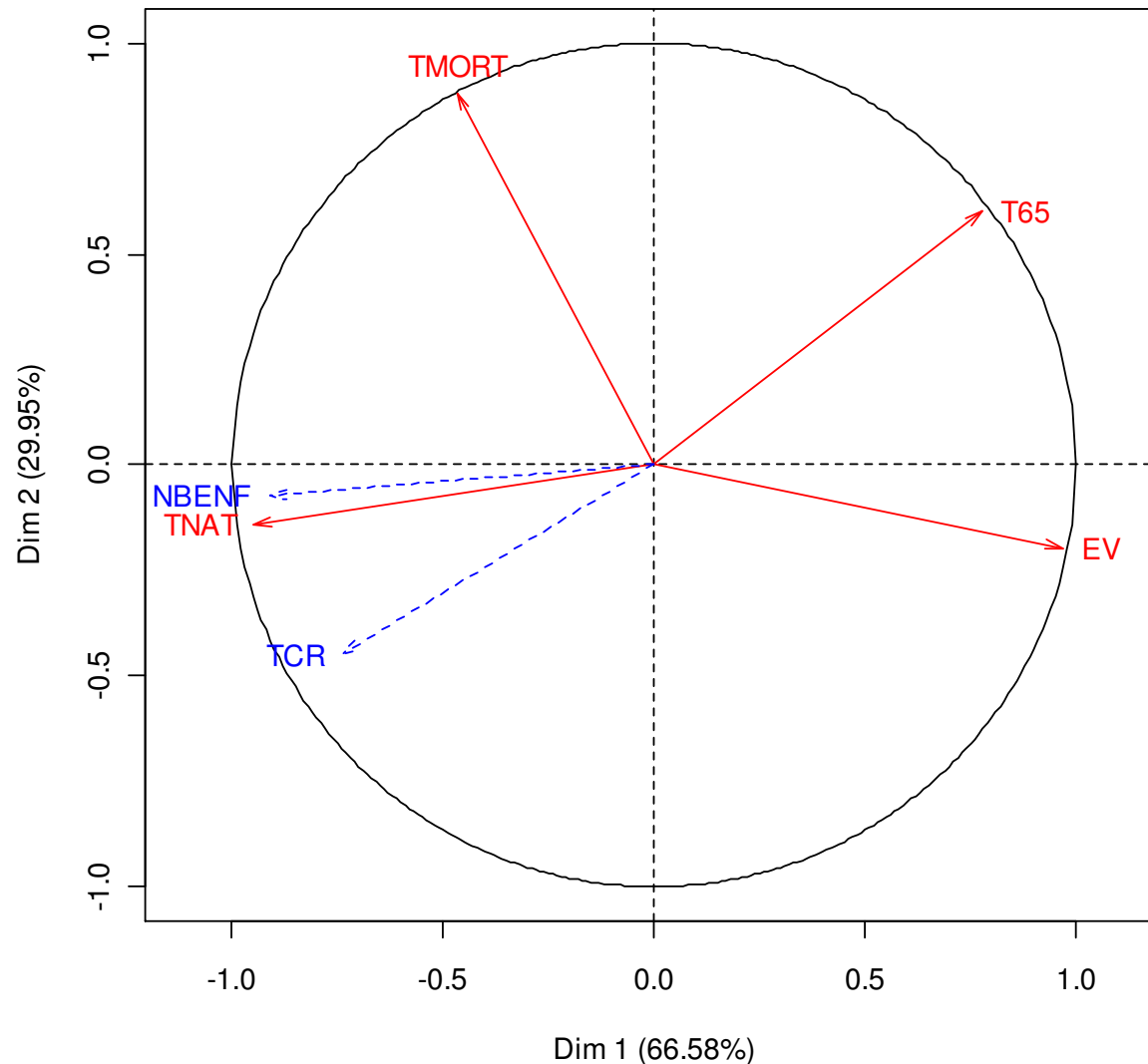
**Ajout d'une variable qualitative**  
une modalité = un nouveau point qui est le centre de gravité des individus présentant cette modalité



## ACP

# Ajout de variable ou d'observation

Représentation des variables



Ajout des variables nombre d'enfants par femme et taux de croissance.

Cet ajout n'a pas modifié le calcul des composantes principales. Il s'agit juste d'une projection des variables dans le cercle de corrélation.

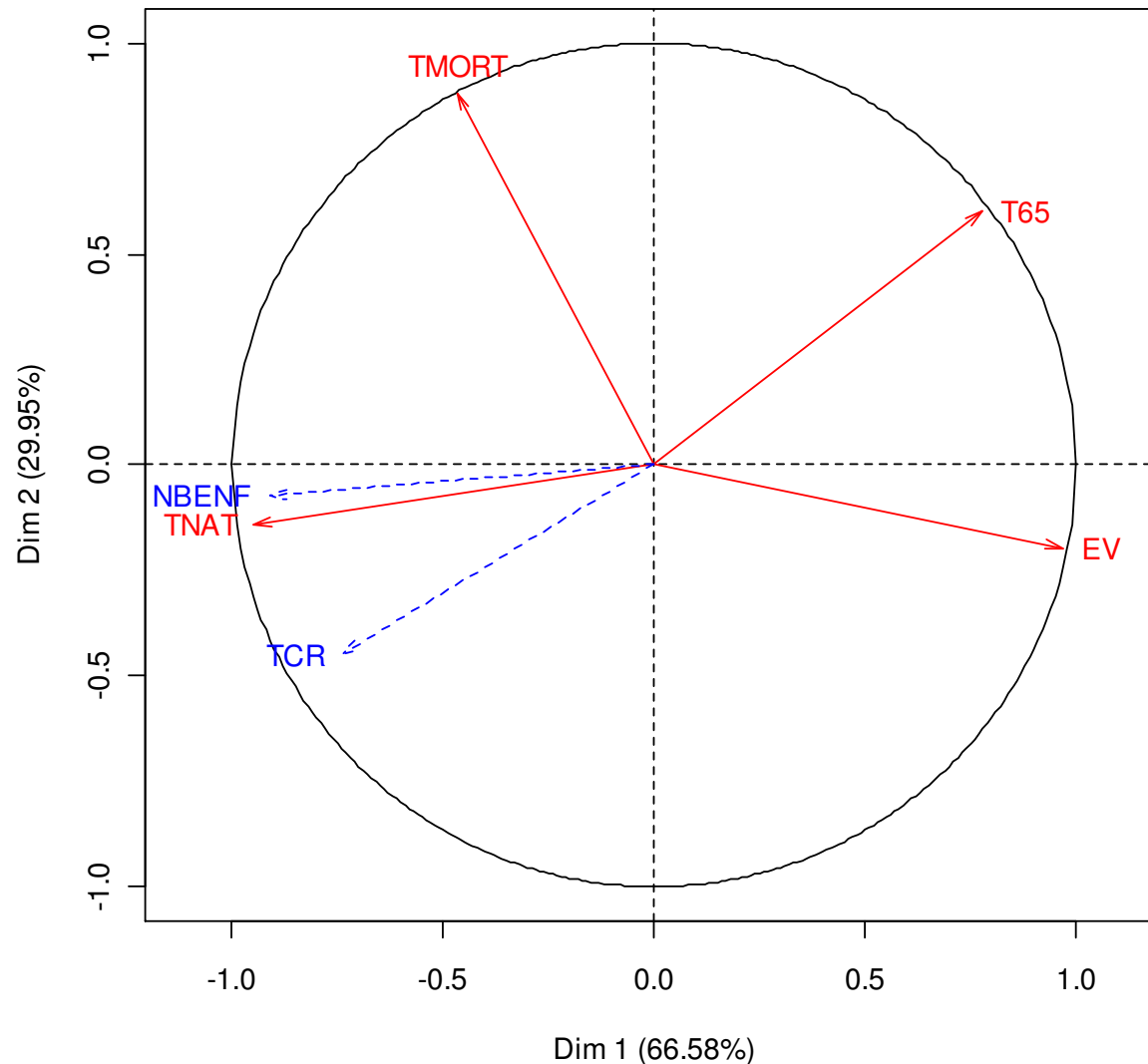




## ACP

# Ajout de variable ou d'observation

Représentation des variables



Ajout des variables nombre d'enfants par femme et taux de croissance.

Cet ajout n'a pas modifié le calcul des composantes principales. Il s'agit juste d'une projection des variables dans le cercle de corrélation.



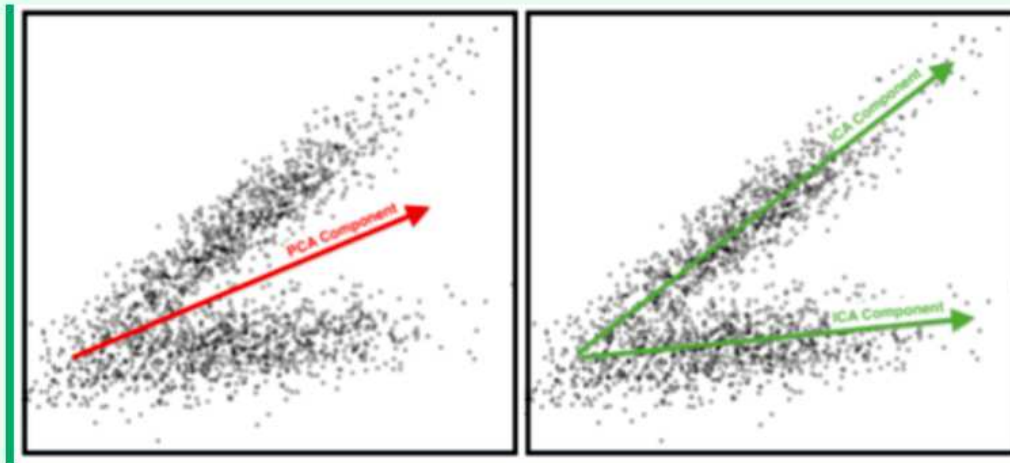
# ACP

## Alternatives à l'ACP

Analyse en composantes principales indépendants  
Analyse en composantes principales par noyaux

### Extensions of PCA

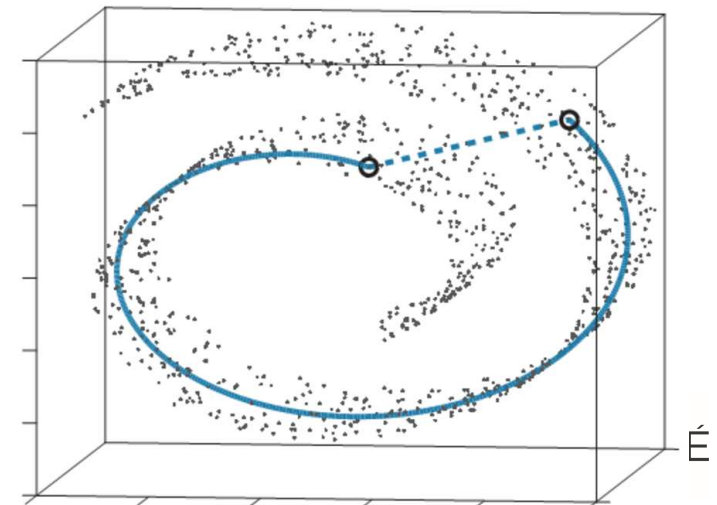
In some cases the orthogonality of the principal components prevent it from extracting informative features. For these cases **independent component analysis** (ICA) is a better choice. An examples case where ICA works much better than PCA is illustrated below.



Finally, PCA works better on data with linear relationships. For non-linear transformation, Kernel PCA can be applied for better results. This [link](https://www.commonlounge.com/discussion/9bcc188644cd4bc9b7542dab93fa8bd7/history) shows an

<https://www.commonlounge.com/discussion/9bcc188644cd4bc9b7542dab93fa8bd7/history>

L'algorithme **t-SNE (t-distributed stochastic neighbor embedding)** est une technique de réduction de dimension pour la visualisation de données . Il s'agit d'une méthode **non linéaire** (contrairement à l'ACP) permettant de représenter un ensemble de points d'un espace à grande dimension dans un espace de deux ou trois dimensions, les données peuvent ensuite être visualisées avec un nuage de points. L'algorithme t-SNE tente de trouver une configuration optimale pour respecter les proximités entre points : deux points qui sont proches (resp. éloignés) dans l'espace d'origine devront être proches (resp. éloignés) dans l'espace de faible dimension.



<https://openclassrooms.com/fr/courses/4379436>