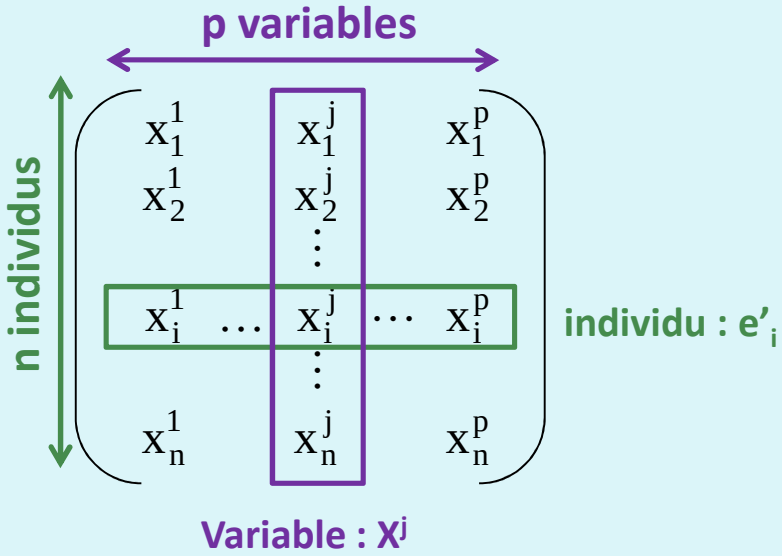
Analyse en composantes principales (A.C.P.)

L' ACP est une méthode descriptive permettant de représenter graphiquement l'essentielle de l'information contenu dans le tableau des données quantitatives



VARIABLE = Elément de $\mathbb{R}^n$

$\Rightarrow$ Visualisation des variables en fonction de leurs corrélations

INDIVIDU = Elément de $\mathbb{R}^p$

$\Rightarrow$ Visualisation des individus selon une distance appropriée

Perte d'information

Projections

en 2D ou 3D

Représentation graphique en dimension > 3D ???

Comment évaluer les ressemblances entre individus?

- Quels sont les individus qui se ressemblent?
- Quels sont ceux qui sont différents?
- Y-a-t'il des groupes homogènes d'individus?
- Peut-on définir une typologie des individus?

Notion de distance entre individus

$i^{ème}$ individu : $e_i = (x_i^1, \ldots, x_i^p)$

$j^{ème}$ individu : $e_j = (x_j^1, \ldots, x_j^p)$

distance entre e_i et e_j

$$\| e_i - e_j \|^2 = \sum_{k=1}^{p} (x_i^k - x_j^k)^2$$

	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

Problème des unités

(ordre de grandeurs des variables)

Centrage et réduction

$$x_i^k \leftarrow \frac{x_i^k - \bar{x}^k}{s_k}$$

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

Comment évaluer les liaisons entre variables?

- Quelles sont les variables liées positivement?
- Quelles sont les variables qui s'opposent?
- Y-a-t'il des groupes de variables liées?
- Peut-on définir une typologie des variables?

Traditionnellement, on utilise le coefficient de corrélation linéaire pour étudier le lien entre deux variables :

$$r(X^k, X^h) = \frac{\mathrm{cov}(X^k, X^h)}{\sqrt{\mathrm{var}(X^k)\,\mathrm{var}(X^h)}}$$

Pourquoi?

Produit scalaire entre variables

Produit scalaire : $\langle X^k, X^h \rangle = E[X^k X^h]$

Norme : $\|X\|^2 = E[X^2]$

Variables centrées

$\langle X^k, X^h \rangle = \mathrm{cov}(X^k, X^h)$

$\|X\|^2 = \mathrm{var}(X)$

$$\cos(\hat{X^k, X^h}) = \frac{\langle X^k, X^h \rangle}{\|X^k\| \|X^h\|} = r(X^k, X^h)$$

- $|r(X^k, X^h)| = 1 \Leftrightarrow$ les variables colinéaires
  - corrélées positivement si $r(X^k, X^h) = 1$
  - corrélées négativement si $r(X^k, X^h) = -1$
- $r(X^k, X^h) = 0 \Leftrightarrow$ les variables sont orthogonales
  $\Leftrightarrow$ les variables ne sont pas linéairement corrélées

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

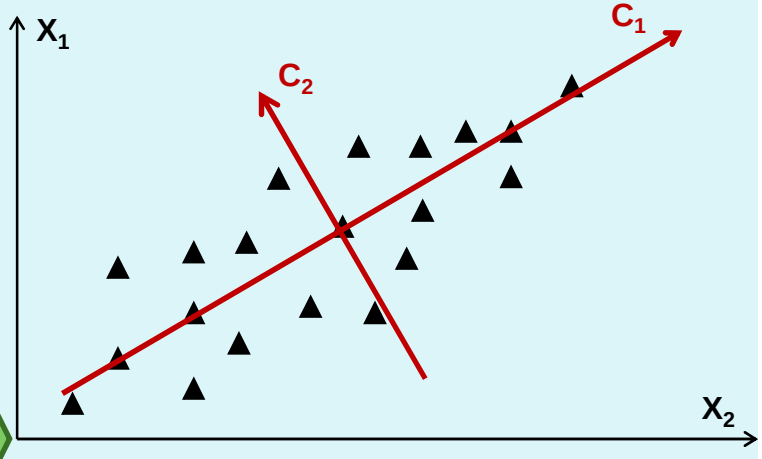
On cherche donc un sous-espace F_k de $\mathbb{R}^p$ de dimension k (k=2,3,..) sur lequel projeté le nuage de points, c-a-d, on cherche k nouvelles variables combinaisons linéaires des p variables initiales tel que le nuage projeté garde le plus d'information possible.

nouvelles variables = composantes principales

- La 1ère composante principale (C_1) doit « capturer » le maximum d'information

Il reste un résidu d'information non expliquée

- La 2ème 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



On stoppe le processus quand résidu d'information négligeable

Nb. composantes principales = Nb. variables initiales

Comment mesurer l'information?

L'inertie est la dispersion du nuage de points

$$I = \frac{1}{n}\sum_{i=1}^{n} \| e_i - g \|^2$$

où g centre de gravité du nuage

N.B. Si les variables sont centrées-réduites, I=p

$$V = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ & s_2^2 & \cdots & s_{2p} \\ & & \ddots & \vdots \\ & & & s_p^2 \end{pmatrix}$$

Matrice des variance-covariances

$I = tr(V)$

Comment perdre le moins d'information possible?

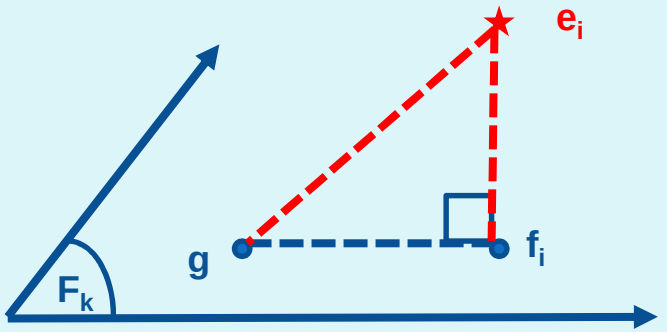
Soit f_i la projection orthogonale de e_i sur F_k.

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

$$\text{minimiser} \quad : \quad \sum_{i=1}^{n} p_i \| e_i - f_i \|^2$$

D'après Pythagore, cela revient à maximiser l'inertie du nuage projeté

$$\text{maximiser} \quad : \quad \sum_{i=1}^{n} p_i \| f_i - g \|^2$$



N.B. p_i poids du $i^{ème}$ individu. En général $p_i=1/n$.

Solution du problème

$F_k=Vect(u_1,u_2,\dots u_k)$
où u_k est le vecteur propre unitaire de V associé à la $k^{ième}$ plus grand valeur propre λ_k.

- L'inertie du nuage projeté sur u_k est λ_k
- L'inertie du nuage projeté sur F_k est $\lambda_1+\dots+\lambda_k$
- L'inertie totale est $I=\lambda_1+\dots+\lambda_p$

Les vecteurs propres sont appelés les **axes principaux**

- Le premier axe principal u_1 est associé à la plus grande valeur propre λ_1
- Le deuxième axe principal u_2 est associé à la deuxième valeur propre λ_2
- Etc,…

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

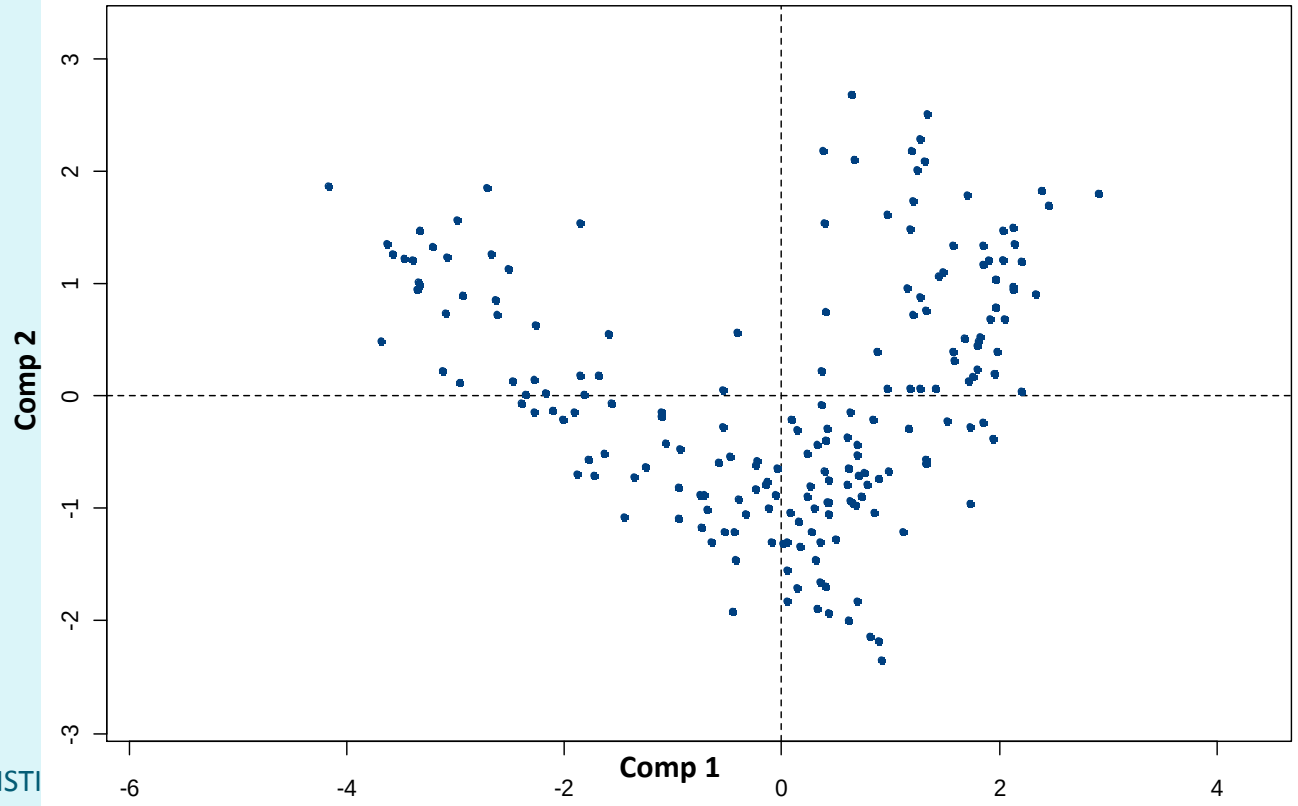
- La première composante c_1 représente les coordonnées des projections des individus sur l'axe u_1
- La deuxième composante c_2 représente les coordonnées des projections des individus sur l'axe u_2
- Etc…

Représentation graphique des individus

Exemple de la démographie mondiale avec les variables : TNAT, TMORT, EV, T65

	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
⇔
4 composantes principales



Représentation des individus (pays) sur les 2ères composantes principales

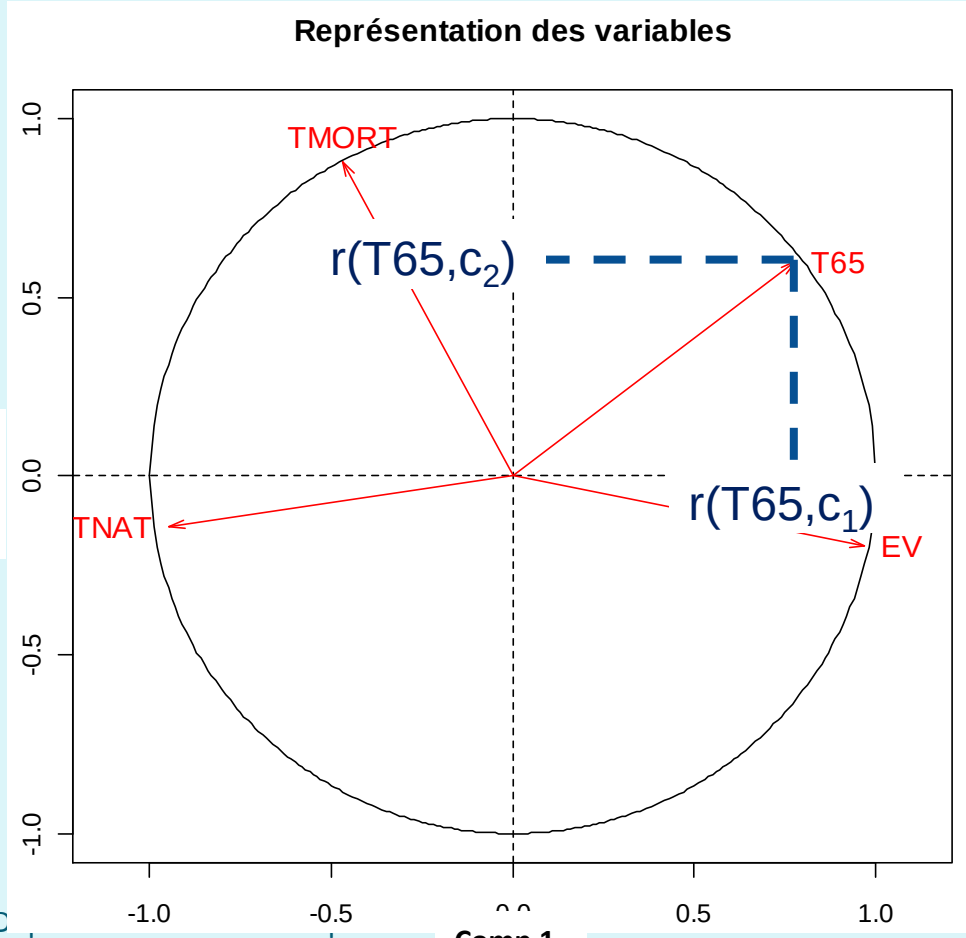
Comment interpréter ce graphique?

Représentation graphique des variables

- $cor = coordonnée=corrélation
- $cos2= répartition de la variable sur les 4 axes principaux

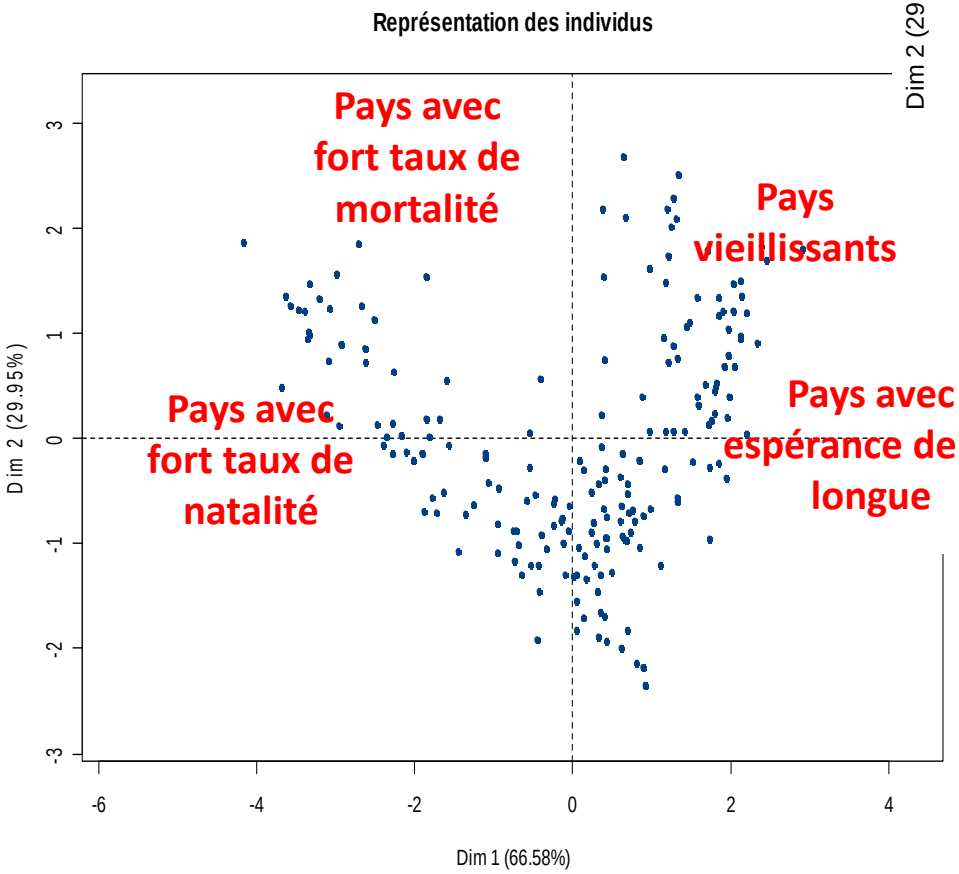
  La variable TNAT est représentée à 89.82% sur C_1 et 1.98% sur C_2 etc..
- $contrib = contribution de la variable à la construction de l'axe

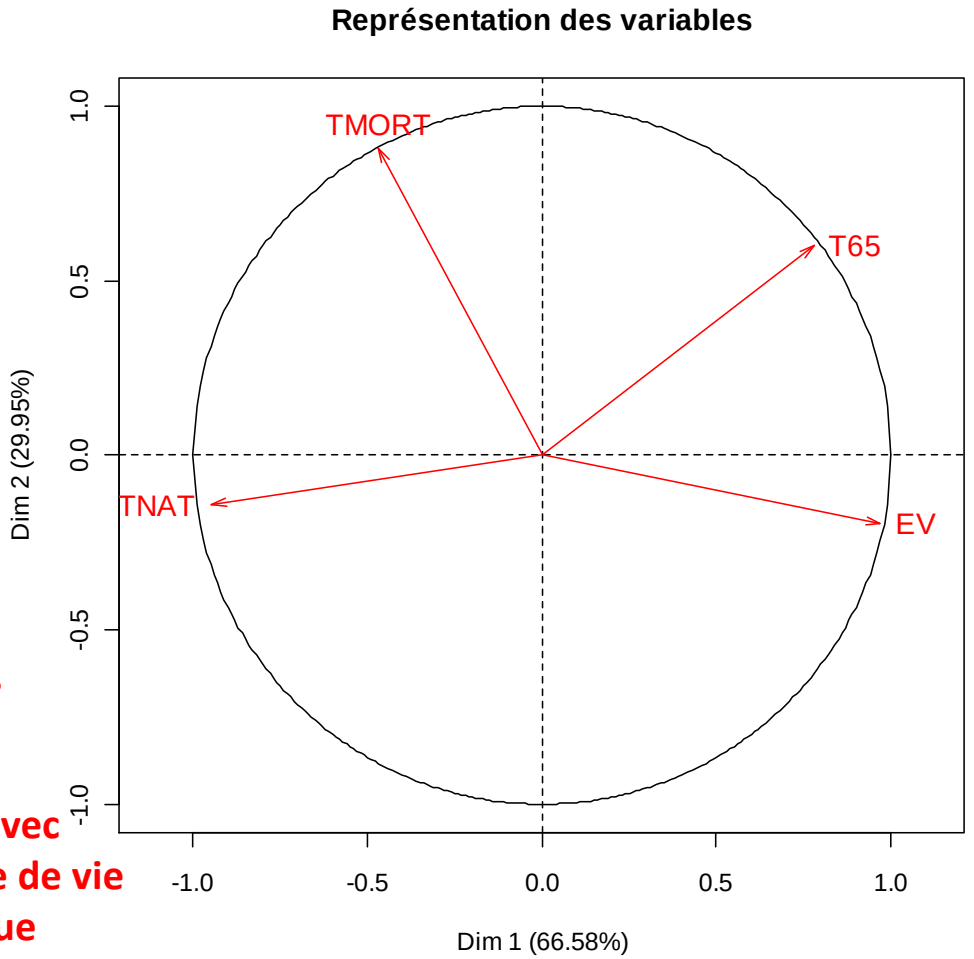
  La variable TMORT ne contribue pas (8%) à la construction de c_1.




```
$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
```

Interprétation





- 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

Validité des représentations graphiques

• La projection perd le moins d'information possible

⇒ *vérifier le % d'inertie expliquée pas l'axe*

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

Exemple démographie :
66,6% d'inertie expliquée par l'axe c_1
29,9% d'inertie expliquée par l'axe c_2
⇒ 96,5% d'inertie expliquée par le plan (c_1,c_2)

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

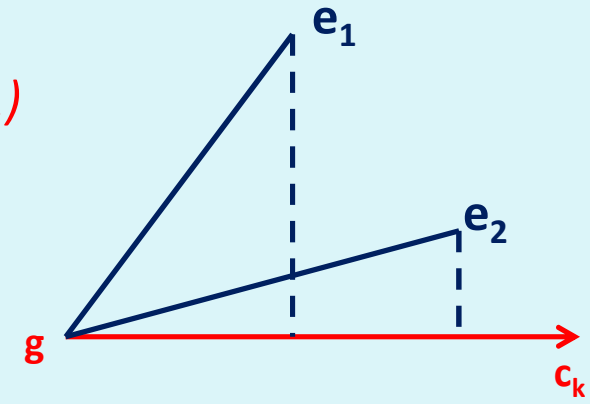
• Les variables sont bien représentées si elles sont proches du cercle. A contrario celles qui sont proches de l'origine sont peu corrélées avec les axes

⇒ *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*

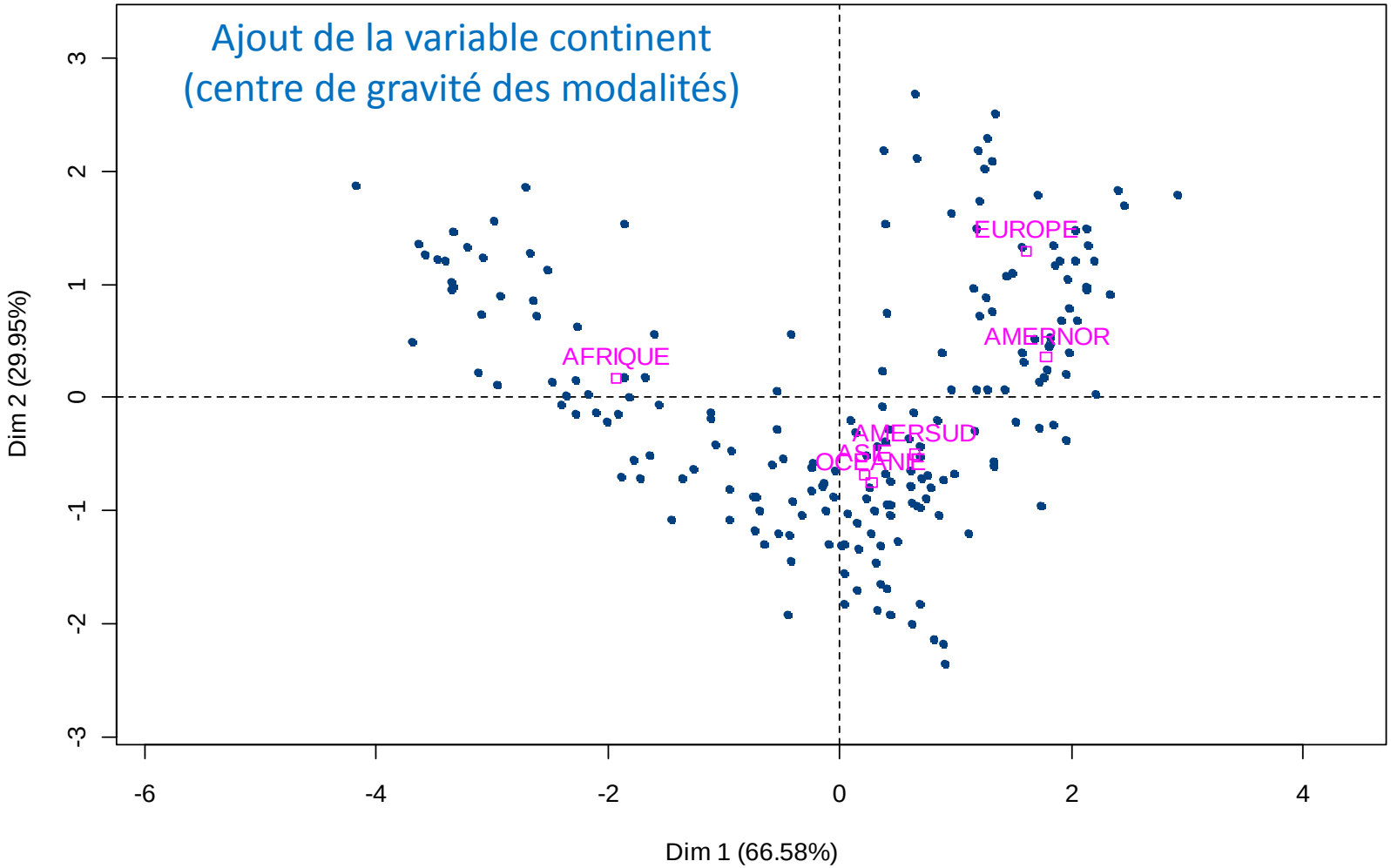


• Eliminer les individus ayant une contribution trop importante dans la construction de l'axe

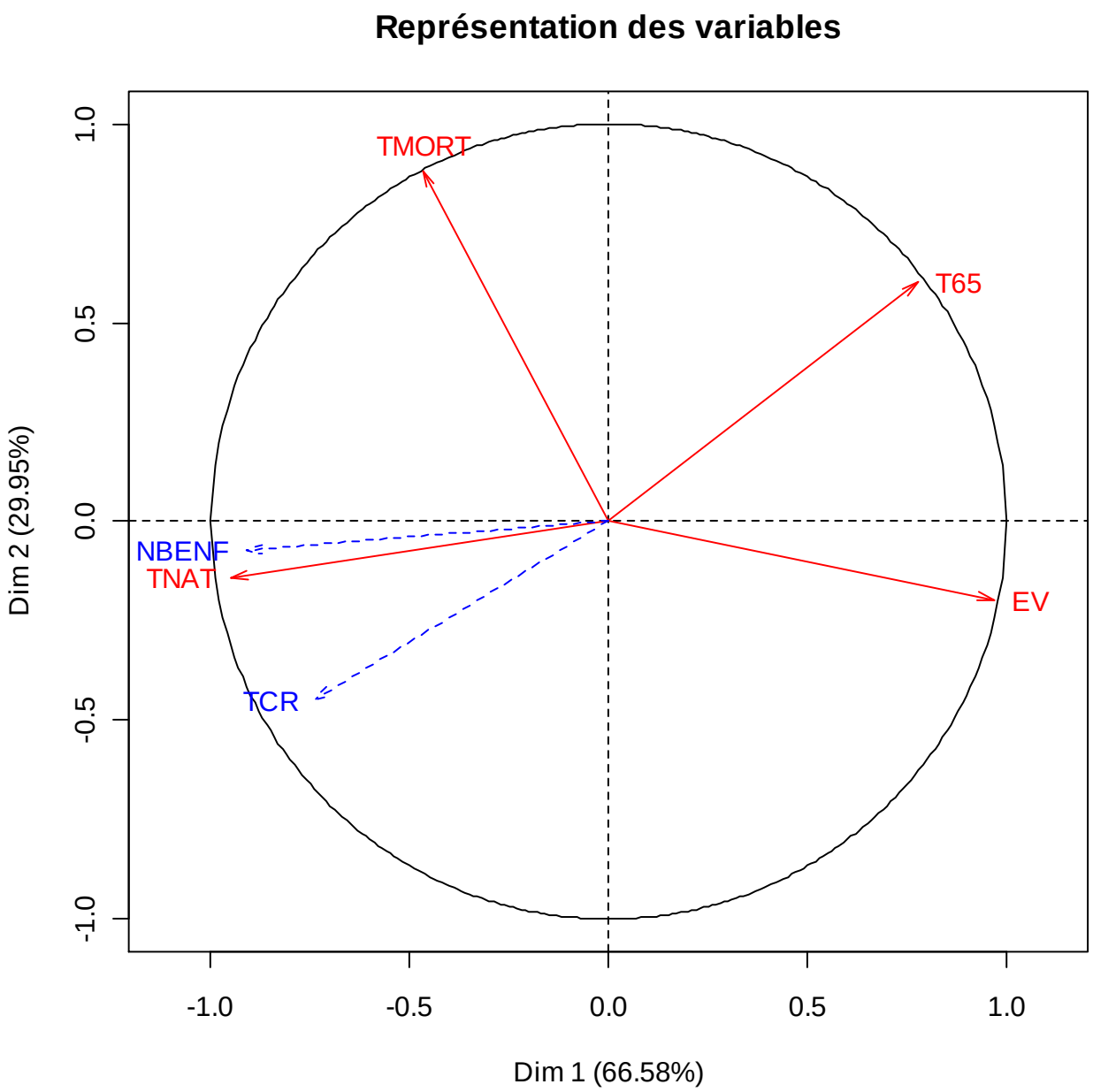
⇒ *vérifier la contribution des individus*

Ajout de variables ou individus

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



Ajout de variables ou individus



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

Résultats ACP avec Excel

Coordonnées factorielles des colonnes					Contributions			
	G1	G2	G3	G4	CTR1	CTR2	CTR3	CTR4
TNAT	0,95	-0,14	0,29	-0,01	33,7%	1,7%	64,3%	0,3%
TMORT	0,47	0,88	-0,04	-0,06	8,2%	64,9%	1,2%	25,7%
EV	-0,97	-0,20	0,13	-0,08	35,3%	3,2%	12,7%	48,8%
T65	-0,78	0,60	0,17	0,05	22,8%	30,3%	21,7%	25,2%
NBENF	0,91	-0,07	0,37	-0,03				
TCR	0,74	-0,45	0,30	0,00