L'ANALYSE EN COMPOSANTES PRINCIPALES (A.C.P.)

Pierre-Louis GONZALEZ

INTRODUCTION

Données :

n individus observés sur p variables quantitatives.

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

Résultats :

⇨ **Visualisation des individus**

(Notion de distances entre individus)

⇨ **Visualisation des variables**

(en fonction de leurs corrélations)

INTERPRÉTATION DES RÉSULTATS

① Mesurer la qualité des représentations obtenues :

- critère global
- critères individuels

② « Donner des noms aux axes »

Expliquer la position des individus

③ Utilisation éventuelle de variables supplémentaires

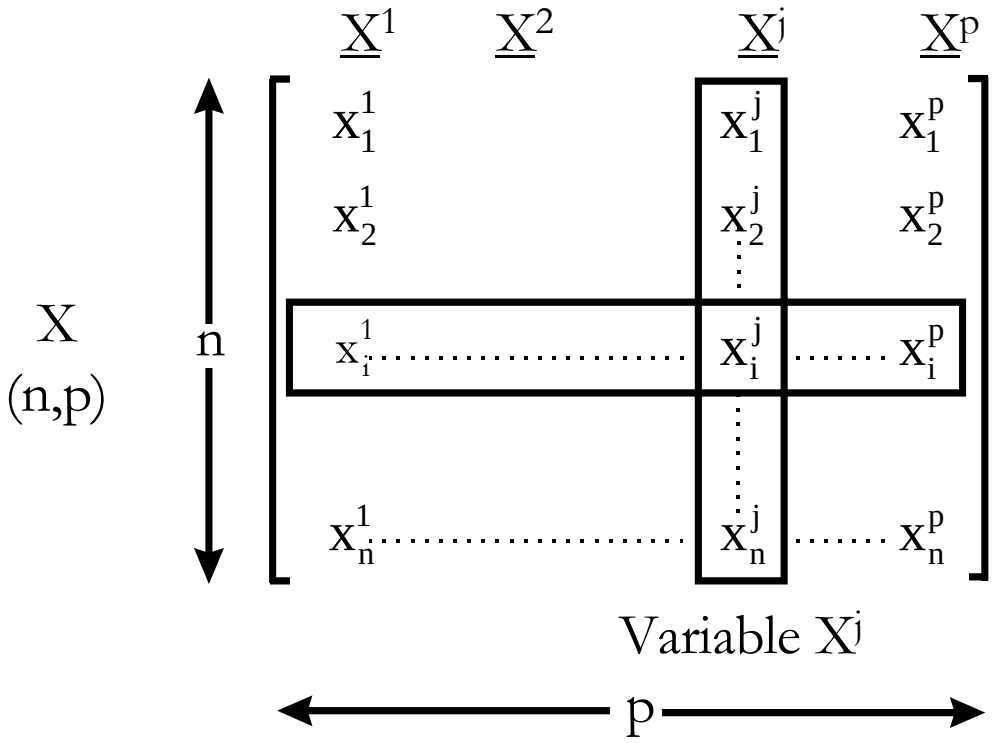
(illustratives)

I. L'ANALYSE EN COMPOSANTES PRINCIPALES

LE PROBLÈME

1. LES DONNÉES

p variables quantitatives observées sur n individus.



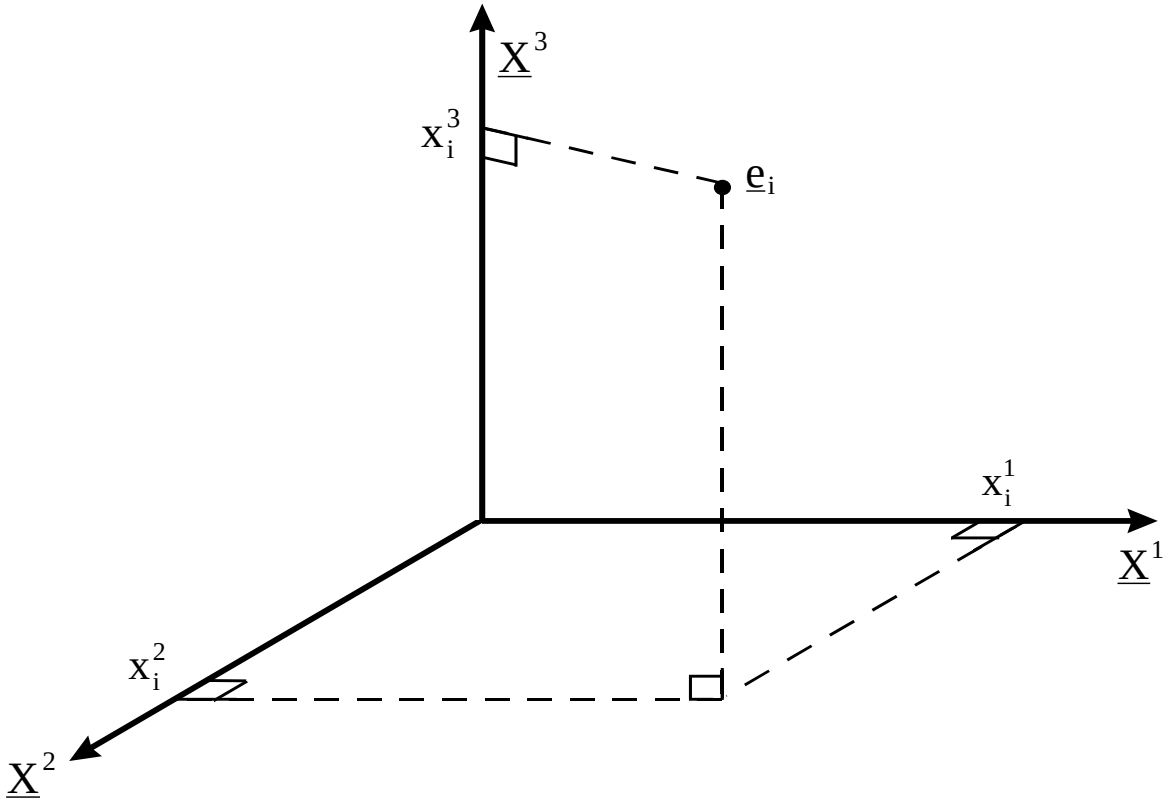
INDIVIDU = Élément de R^p

VARIABLE = Élément de R^n

On cherche à représenter le nuage des individus.

A chaque individu noté e_i, on peut associer un point dans R^p = espace des individus.

A chaque variable du tableau X est associé un axe de R^p.



Impossible à visualiser dès que p > 3.

2. PRINCIPE DE L'A.C.P.

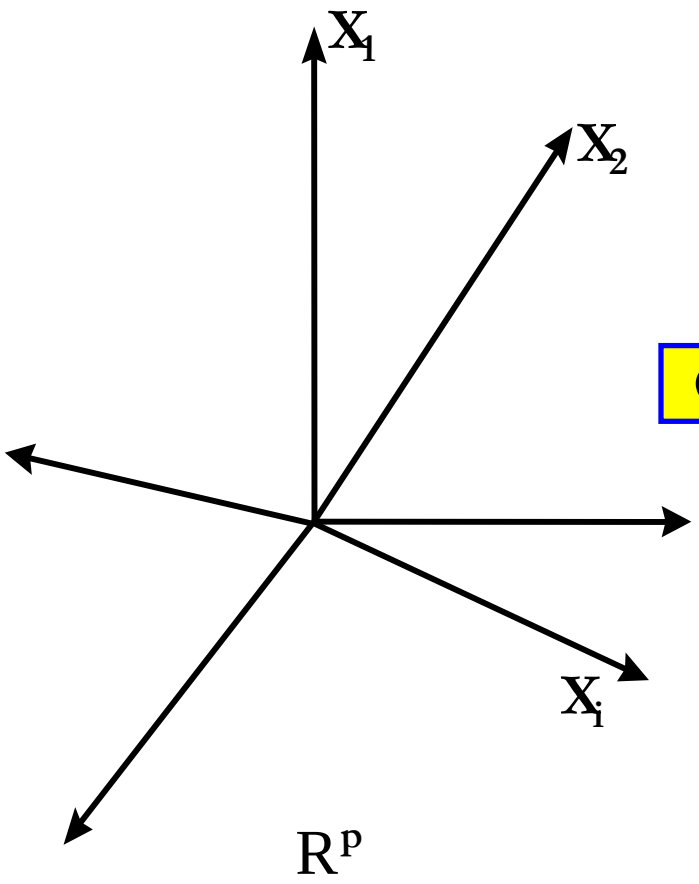
On cherche une représentation des n individus , dans un sous-espace F_k de R^p de dimension k (k petit 2, 3 …; par exemple un plan)

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

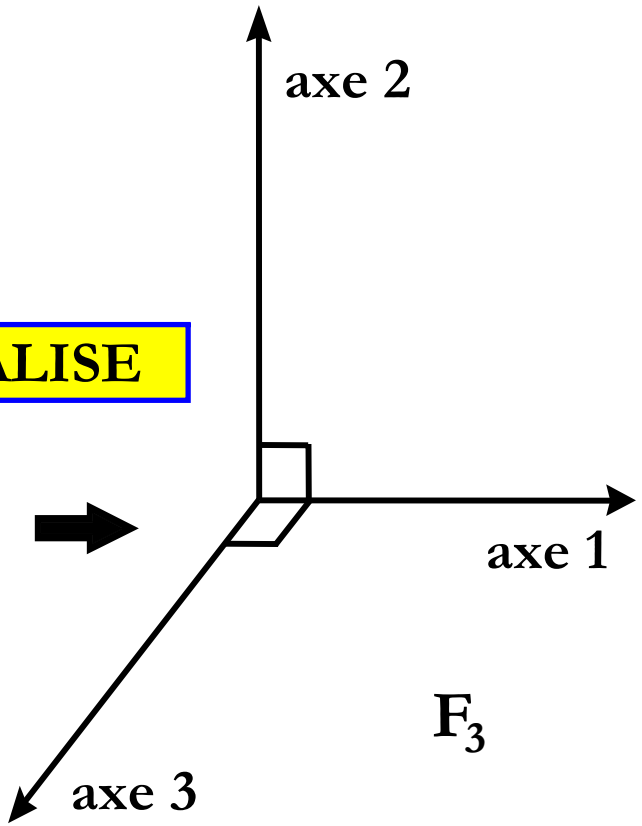
Ces variables seront appelées «***composantes principales*** »,

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

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



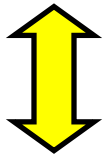
ON VISUALISE



axes principaux

« Perdre le moins d'information possible »

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

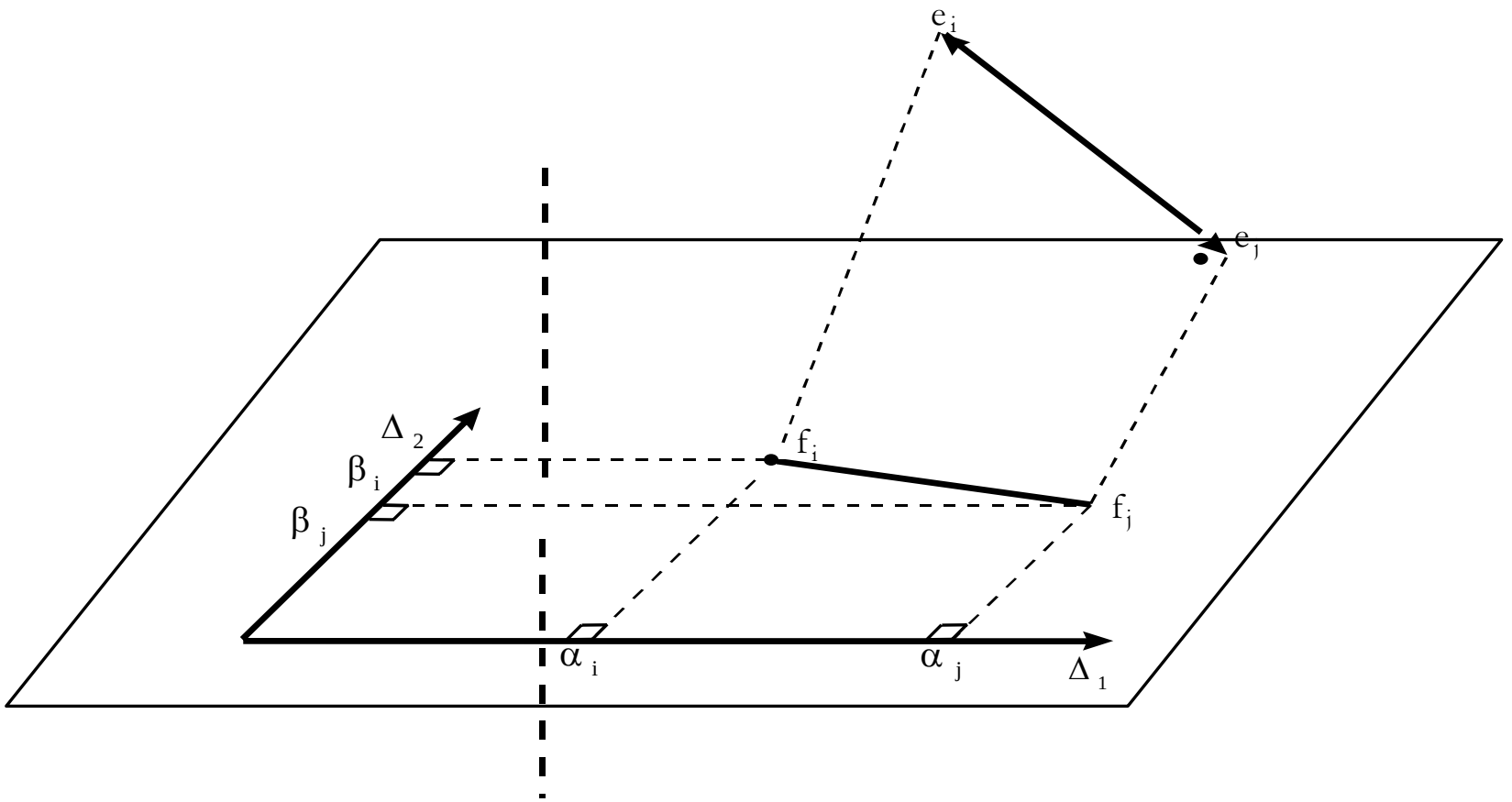


② F_k est le sous-espace tel que le nuage projeté ait une **inertie** (dispersion) maximale.

① **et** ② sont basées sur les notions de :

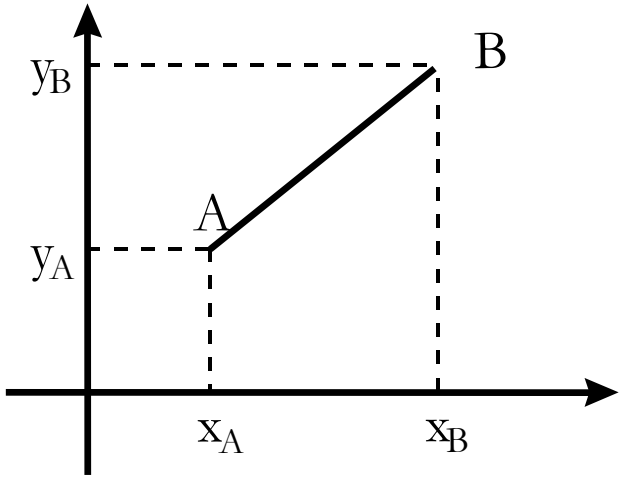
distance

projection orthogonale



La distance entre f_i et f_j est inférieure ou égale à celle entre e_i et e_j

3. LE CHOIX DE LA DISTANCE ENTRE INDIVIDUS



Dans le plan:

$$d^2(A,B) = (x_B - x_A)^2 + (y_B - y_A)^2$$

Dans l'espace R^p à p dimensions, on généralise cette notion : la distance euclidienne entre deux individus s'écrit:

$$e_i = (x_i^1\ x_i^2\ \dots\ x_i^p) \qquad e_j = (x_j^1\ x_j^2\ \dots\ x_j^p)$$

$$d^2(e_i, e_j) = (x_i^1 - x_j^1)^2 + (x_i^2 - x_j^2)^2 + \dots (x_i^p - x_j^p)^2$$

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

Le problème des unités ?

Pour résoudre ce problème, on choisit de transformer les données en données centrées-réduites.

L'observation X_i^k est alors remplacée par :

UNITÉS D'ÉCART TYPE:

$$\frac{X_i^k - \overline{X}^k}{S_k}$$

où : $\overline{X}^k$ = moyenne de la variable $\mathbf{X^k}$

s_k = écart-type de la variable $\mathbf{X^k}$

Exemple :

Puissance moyenne de 30 voitures = 92 ch Ecart-type = 24 ch

La Renault 21 TXI a une puissance de 140 ch

La Renault 21 TXI a une puissance de : $\frac{140-92}{24} = 2$

2 écarts-type au-dessus de la moyenne.

4. INERTIE TOTALE

$$I_g = \sum_{i=1}^{n} \frac{1}{n} d^2(e_i, \underline{g})$$

ou de façon plus générale

$$I_g = \sum_{i=1}^{n} p_i \, d^2(e_i, \underline{g})$$

avec $\sum_{i=1}^{n} p_i = 1$

L'inertie est la somme pondérée des carrés des distances des individus au centre de gravité $\underline{g}$

L'inertie mesure la dispersion totale du nuage de points.

L'inertie est donc aussi égale à la somme des variances des variables étudiées.

En notant V la matrice de variances-covariances :

$$V = \begin{pmatrix} s_1^2 & s_{12} \cdots\cdots & s_{1p} \\ \vdots & & \vdots \\ & s_2^2 & \\ \vdots & & \vdots \\ s_{p1} & & s_p^2 \end{pmatrix}$$

$$I_g = \sum_{i=1}^{p} s_i^2$$

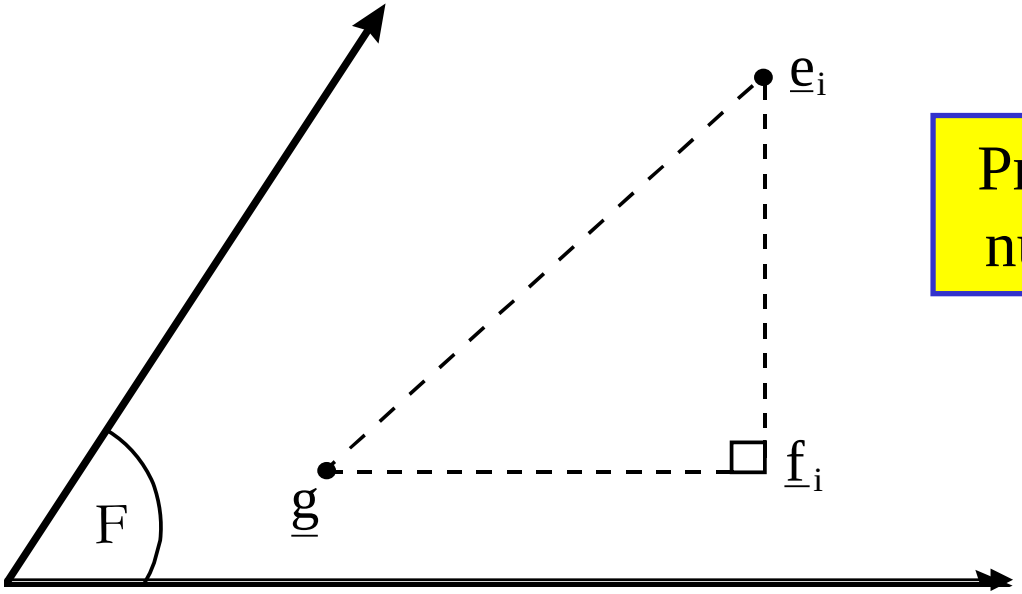
$$I_g = Tr(V)$$

Remarque

Dans le cas où les variables sont centrées réduites, la variance de chaque variable vaut 1.

L'inertie totale est alors égale à p (nombre de variables).

Équivalence des deux critères concernant la perte d'information



Projection orthogonale du nuage sur un sous-espace

Soit F un sous-ensemble de $\mathbf{R}^p$

$\underline{f}_i$ la projection orthogonale de $\underline{e}_i$ sur F

$$\left\|\underline{e}_i - \underline{g}\right\|^2 = \left\|\underline{e}_i - \underline{f}_i\right\|^2 + \left\|\underline{f}_i - \underline{g}\right\|^2 \qquad \forall i = 1 \ldots n$$

On va chercher F tel que :

① $\sum_{i=1}^{n} p_i \left\| \underline{e}_i - \underline{f}_i \right\|^2$ soit minimal

ce qui revient d'après le théorème de Pythagore à **maximiser** :

② $\sum_{i=1}^{n} p_i \left\| \underline{f}_i - \underline{g} \right\|^2$

$$\left\|\underline{e}_i - \underline{g}\right\|^2 = \left\|\underline{e}_i - \underline{f}_i\right\|^2 + \left\|\underline{f}_i - \underline{g}\right\|^2 \qquad \forall i = 1 \ldots n$$

Donc : $\underbrace{\sum_{i=1}^{n} p_i \left\|\underline{e}_i - \underline{g}\right\|^2}$ − $\underbrace{\sum_{i=1}^{n} p_i \left\|\underline{e}_i - \underline{f}_i\right\|^2}$ = $\underbrace{\sum_{i=1}^{n} p_i \left\|\underline{f}_i - \underline{g}\right\|^2}$

Inertie totale

minimiser cette quantité (carrés des distances entre points individus et leurs projections)

⇔

maximiser l'inertie du nuage projeté

II. LA SOLUTION DU PROBLÈME POSÉ

La recherche **d'axes portant le maximum d'inertie** équivaut à la construction de nouvelles variables (auxquelles sont associés ces axes) de **variance maximale**.

En d'autres termes, on effectue un changement de repère dans R^p de façon à se placer dans un nouveau système de représentation où le premier axe apporte le plus possible de l'inertie totale du nuage, le deuxième axe le plus possible de l'inertie non prise en compte par le premier axe, et ainsi de suite.

Cette réorganisation s'appuie sur la **diagonalisation de la matrice de variances-covariances.**

1. SOLUTION

Axes principaux

On appelle axes principaux d'inertie les axes de direction les vecteurs propres de V normés à 1.

Il y en a p.

Le premier axe est celui associé à la plus grande valeur propre . On le note u^1

Le deuxième axe est celui associé à la deuxième valeur propre . On le note u^2

...

Composantes principales

À chaque axe est associée une variable appelée composante principale.

La composante $\mathbf{c^1}$ est le vecteur renfermant les cordonnées des projections des individus sur l'axe 1.

La composante $\mathbf{c^2}$ est le vecteur renfermant les cordonnées des projections des individus sur l'axe 2.

Pour obtenir ces coordonnées, on écrit que chaque composante principale est une combinaison linéaire des variables initiales.

Exemple

$$\underline{c}^1 = u_1^1 \, \underline{x}^1 + u_2^1 \, \underline{x}^2 + \ \ldots \ u_p^1 \, \underline{x}^p$$

2. PROPRIÉTÉS DES COMPOSANTES PRINCIPALES

① La variance d'une composante principale est égale à l'inertie portée par l'axe principal qui lui est associé.

1ère composante $\mathbf{c^1}$ variance : λ_1

2ème composante $\mathbf{c^2}$ variance : λ_2

3ème composante $\mathbf{c^3}$ variance : λ_3

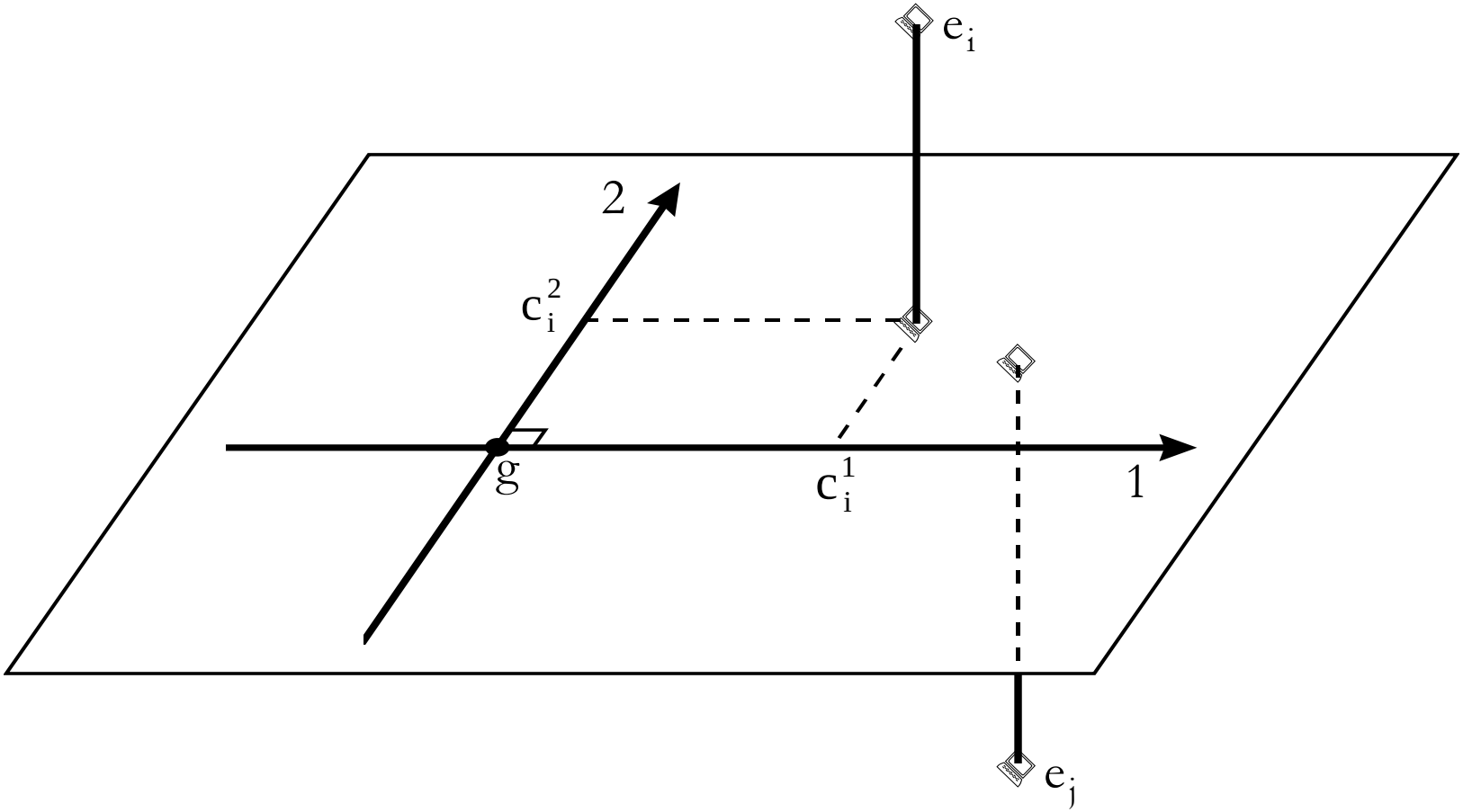
② Les composantes principales sont non corrélées deux à deux.

En effet, les axes associés sont orthogonaux.

3. REPRÉSENTATION DES INDIVIDUS

La $j^{ème}$ composante principale $\underline{c}^j = \begin{pmatrix} c_1^j \\ c_2^j \\ \vdots \\ c_n^j \end{pmatrix}$ fournit les coordonnées des n individus sur le **$j^{ème}$ axe principal**.

Si on désire une **représentation plane** des individus, la meilleure sera celle réalisée grâce aux **deux premières composantes principales.**

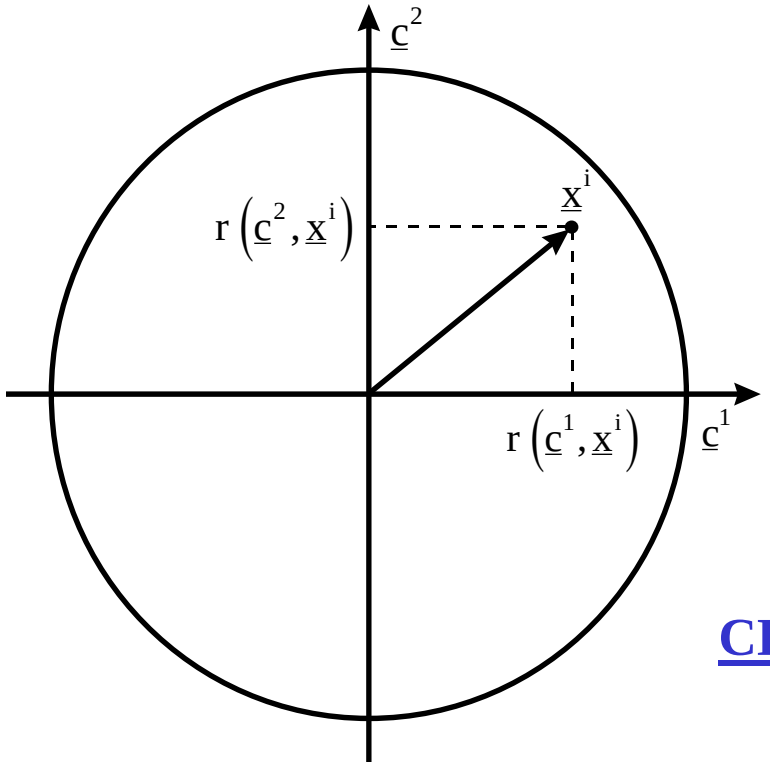


Attention à la qualité de représentation de chaque individu!

4. REPRÉSENTATION DES VARIABLES

Les « proximités » entre les composantes principales et les variables initiales sont mesurées par les covariances, et surtout **les corrélations**.

$r\left(\underline{c}^{j}, \underline{x}^{i}\right)$ est le **coefficient de corrélation linéaire** entre $\underline{c}^{j}$ et $\underline{x}^{i}$



CERCLE DES CORRÉLATIONS

5. INTERPRETATION DES « PROXIMITÉS » ENTRE VARIABLES

On utilise un **produit scalaire** entre variables permettant d'associer aux paramètres courants : écart-type, coefficient de corrélation linéaire des représentations géométriques.

$$\left\langle \underline{x}^{i}, \underline{x}^{j} \right\rangle = \frac{1}{n} \sum_{k=1}^{n} x_{k}^{i} \; x_{k}^{j}$$

On suppose les **variables centrées**.

$$\left\langle \underline{x}^i , \underline{x}^j \right\rangle = \mathrm{Cov}\left(\underline{x}^i , \underline{x}^j \right)$$

$$\left\| \underline{x}^i \right\|^2 = \left\langle \underline{x}^i , \underline{x}^i \right\rangle = \frac{1}{n} \sum_{k=1}^{n} \left(x_k^i \right)^2$$

$$\left\| \underline{x}^i \right\|^2 = s_i^2$$ Variance de $\underline{x}^i$

$$\left\| \underline{x}^i \right\| = s_i$$ Écart-type de $\underline{x}^i$

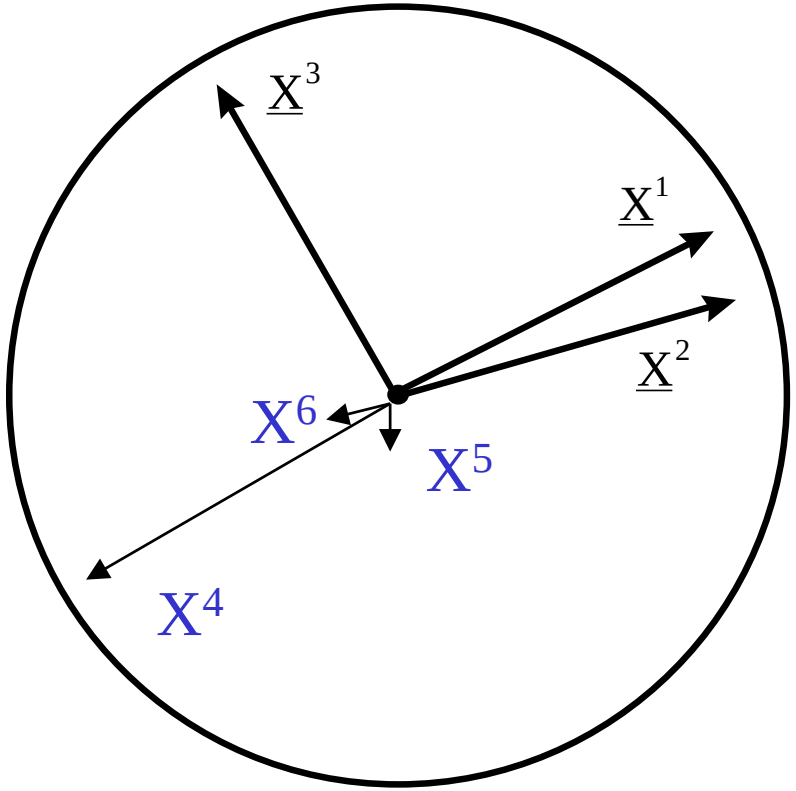
Coefficient de corrélation linéaire

$$\cos\widehat{\left(\underline{X}^i, \underline{X}^j\right)} = \frac{\left\langle \underline{x}^i, \underline{x}^j \right\rangle}{\left\| \underline{X}^i \right\| \left\| \underline{X}^j \right\|} = \frac{\mathrm{Cov}\left(\underline{X}^i, \underline{X}^j\right)}{s_i \, s_j} = r\left(\underline{X}^i, \underline{X}^j\right)$$

Le cosinus de l'angle formé par les variables X^i et X^j est le coefficient de corrélation linéaire de ces deux variables

X^1 et X^2 ont une corrélation proche de 1.

X^1 et X^3 ont une corrélation proche de 0.



CERCLE DES CORRÉLATIONS

III. VALIDITÉ DES REPRÉSENTATIONS

1. CRITÈRE GLOBAL

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots \lambda_p}$$

mesure la part d'inertie expliquée par l'axe i.

Exemple :

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i}$$

est la part d'inertie expliquée par le premier plan principal.

Ce critère (souvent exprimé en pourcentage) mesure le degré de reconstitution des carrés des distances.

La réduction de dimension est d'autant plus forte que les variables de départ sont plus corrélées.

Combien d'axes ?

Différentes procédures sont complémentaires:

① **Pourcentage d'inertie souhaité : a priori**

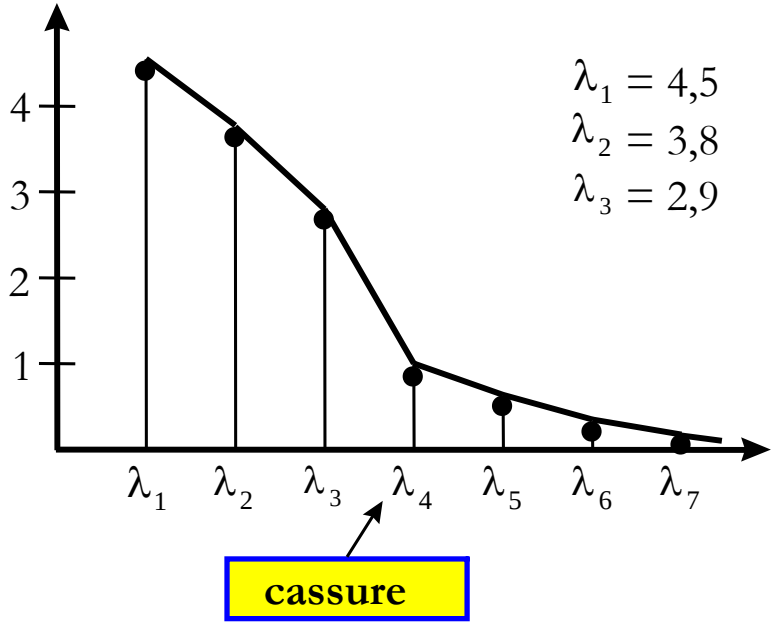
② Diviser l'inertie totale par le nombre de variables initiales

⇨ inertie moyenne par variable : I.M.

Conserver tous les axes apportant une inertie supérieure à cette valeur I.M. (inertie > 1 si variables centrées réduites).

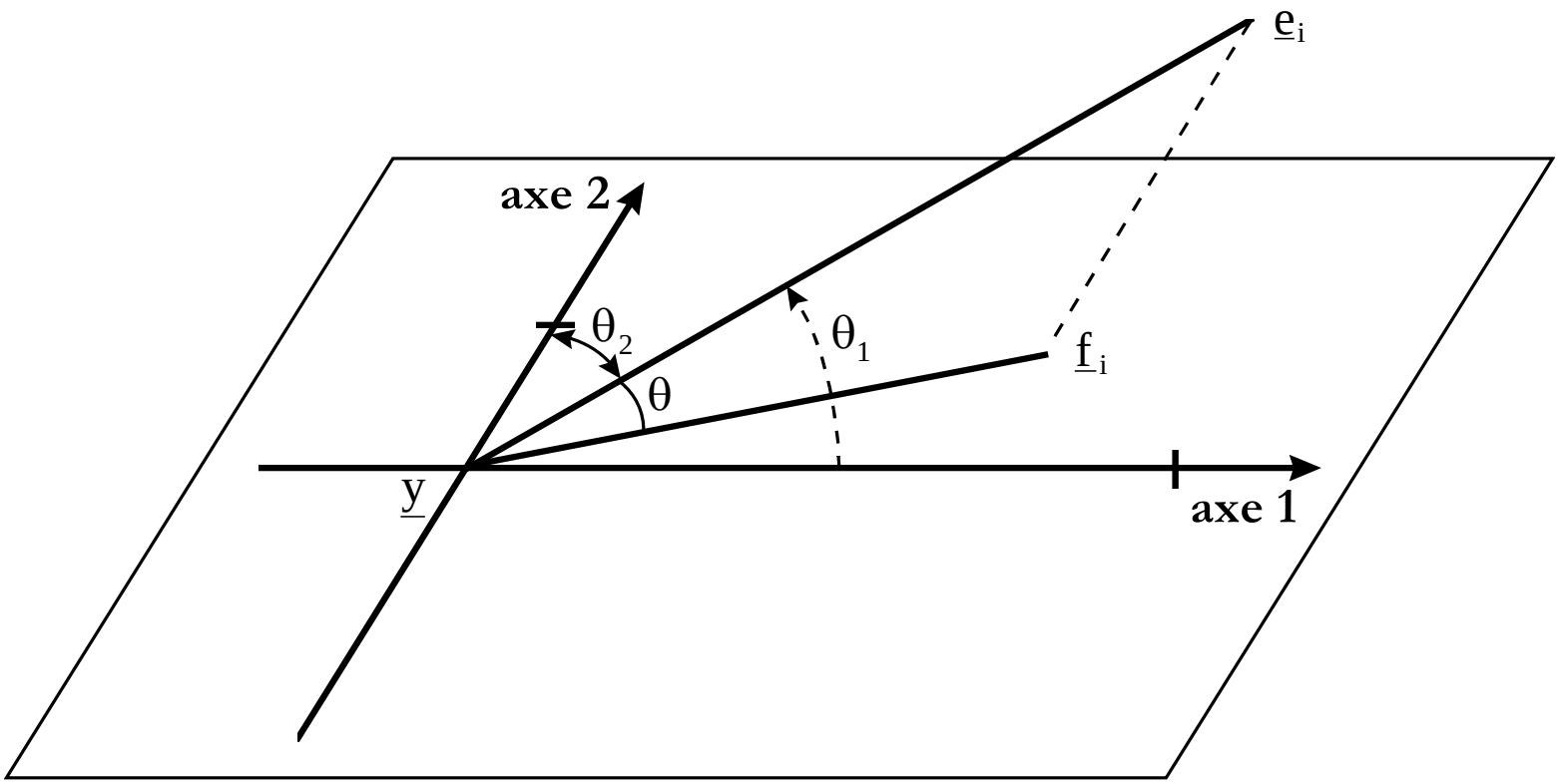
③ Histogramme

Conserver les axes associés aux valeurs propres situées avant la cassure.



2. CRITÈRES INDIVIDUELS

Cosinus carrés



$$\cos^2\theta = \cos^2\theta_1 + \cos^2\theta_2$$

Pour chaque individu , la qualité de sa représentation est définie par le carré du cosinus de l'angle entre l'axe de projection et le vecteur $\underline{e}_i$. **Plus la valeur est proche de 1, meilleure est la qualité de représentation**

En général, les qualités de représentation sont données axe par axe. Pour avoir la qualité de représentation dans un plan, on additionne les critères correspondant aux axes étudiés.

Ce critère n'a pas de signification pour les individus proches de l'origine.

Quand on détecte un individu pour lequel le cosinus carré est faible, on doit tenir compte de sa distance à l'origine avant d'indiquer qu'il est mal représenté

Contributions

Il est très utile aussi de calculer pour chaque axe la **contribution apportée** par les divers individus à cet axe.

Considérons la k^ième composante principale $\underline{c}^k$, soit $\underline{c}_i^k$ la valeur de la composante pour le i^ème individu.

$$\sum_{i=1}^{n} \frac{1}{n} \left(\underline{c}_i^k\right)^2 = \lambda_k$$

La **contribution** de l'individu $\underline{e}_i$

à la composante n° k est définie par

$$\frac{\frac{1}{n}\left(c_i^k\right)^2}{\lambda_k}$$

Remarque :

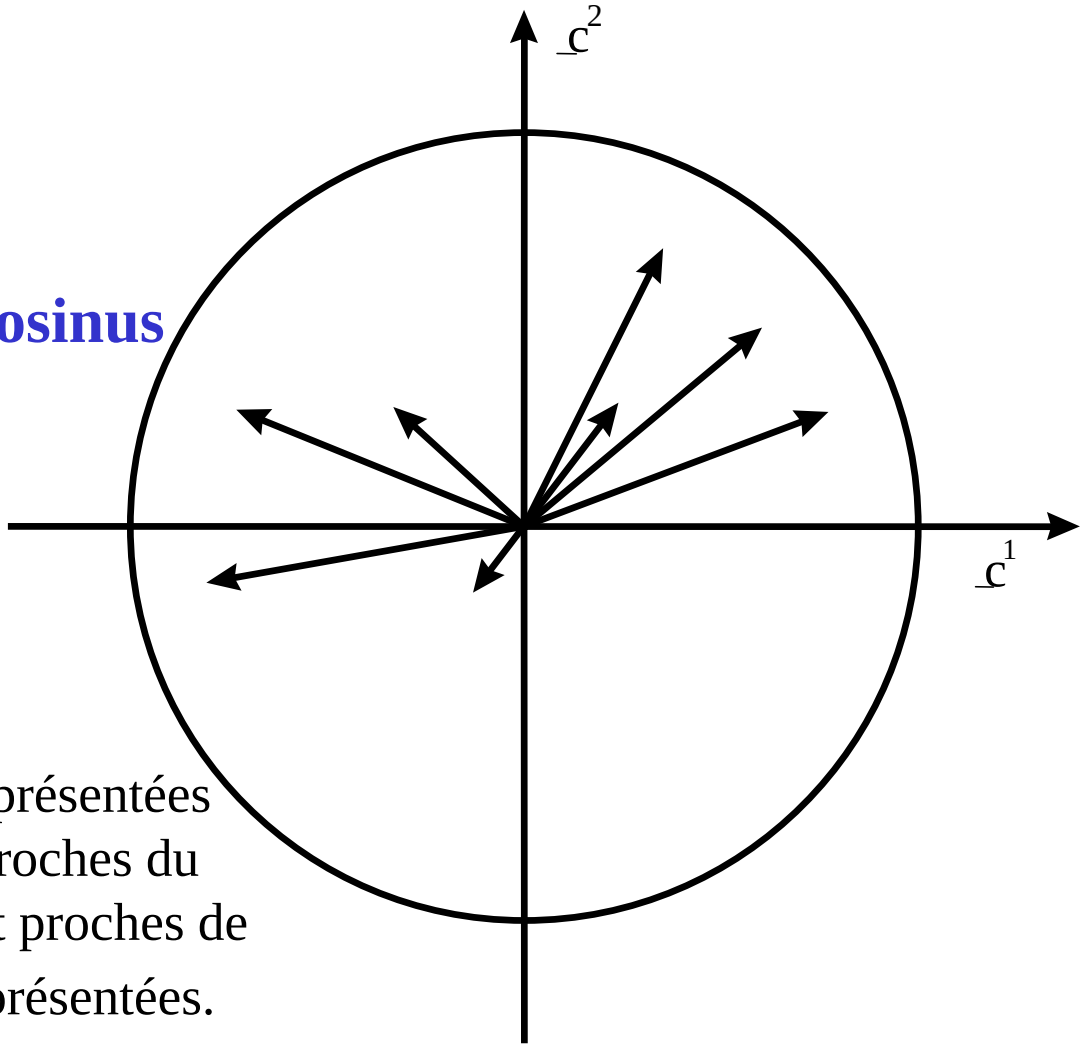
- Il n’est pas souhaitable qu’un individu ait une contribution excessive (car facteur d’instabilité) ➔ éliminer les individus dont la contribution est trop importante.

- Problème des enquêtes par sondage

3. REPRÉSENTATION DES VARIABLES

Le cercle des corrélations est la projection du nuage des variables sur le plan des composantes principales.

corrélation = cosinus



Les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représentées.

4. INTERPRÉTATION EXTERNE : VARIABLES ET INDIVIDUS SUPPLÉMENTAIRES (ILLUSTRATIFS)

4.1 Variables

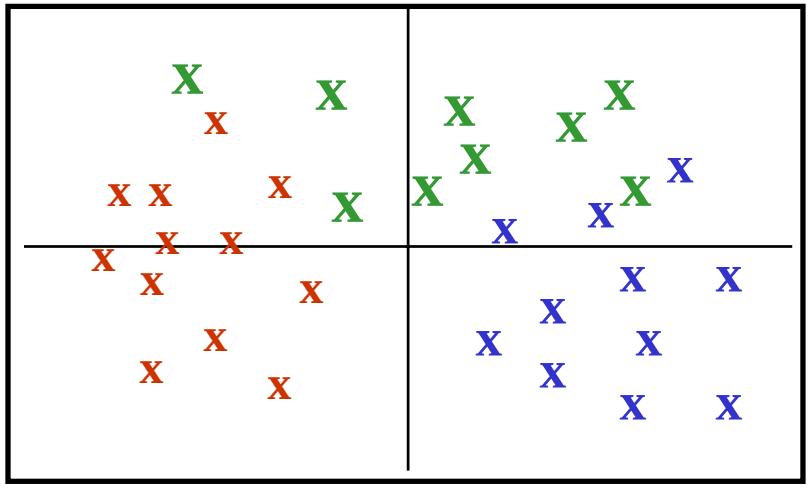
- **Variable quantitative:**

On calcule le coefficient de corrélation entre la variable supplémentaire et les composantes principales.

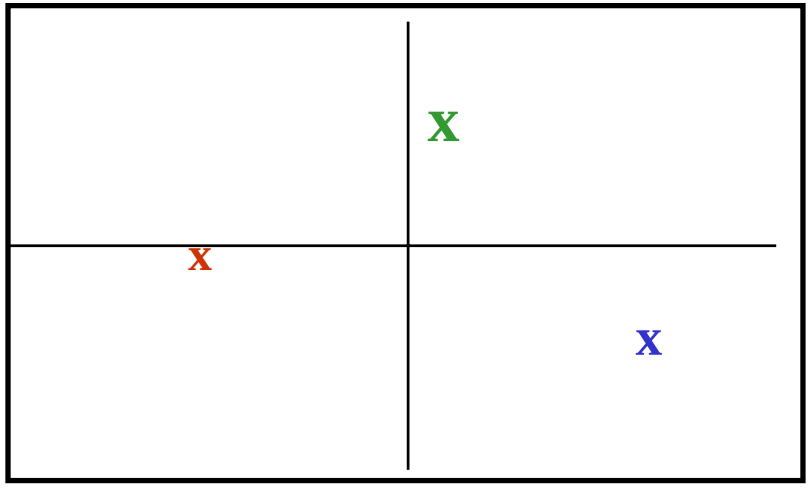
Ceci permet sa représentation sur le **cercle des corrélations**.

Variable qualitative

Identification des individus de chaque catégorie de la variable



Représentation de chaque catégorie par son centre de gravité.



Calcul du **rapport de corrélation** entre la variable qualitative supplémentaire et chaque composante principale (test de Fischer-Snedecor) ou **valeur-test** dans SPAD.

Individus

Individu de poids nul ne participant pas à l'analyse (fichier test).

Appliquer aux coordonnées de l'individu les expressions définissant les composantes principales.