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

Module : Analyse de données

Analyse en composante principale

L’Analyse en Composantes Principales (ACP) est une méthode d’analyse de données. Elle cherche à synthétiser l’information contenue dans un tableau croisant des individus et des variables quantitatives. Produire un résumé d’information au sens de l’ACP c’est etablir une similarité entre les individus, chercher des groupes d’individus homogènes, mettre en évidence une typologie d’individus. Quant aux variables c’est mettre en évidence des bilans de liaisons entre elles, moyennant des variables synthétiques et mettre en évidence une typologie de variables. L’ACP cherche d’une façon générale à établir des liaisons entre ces deux typologies.

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

- **Mesurer la qualité des représentations obtenues :**
  - critère global
  - critères individuels

« Donner des noms aux axes »

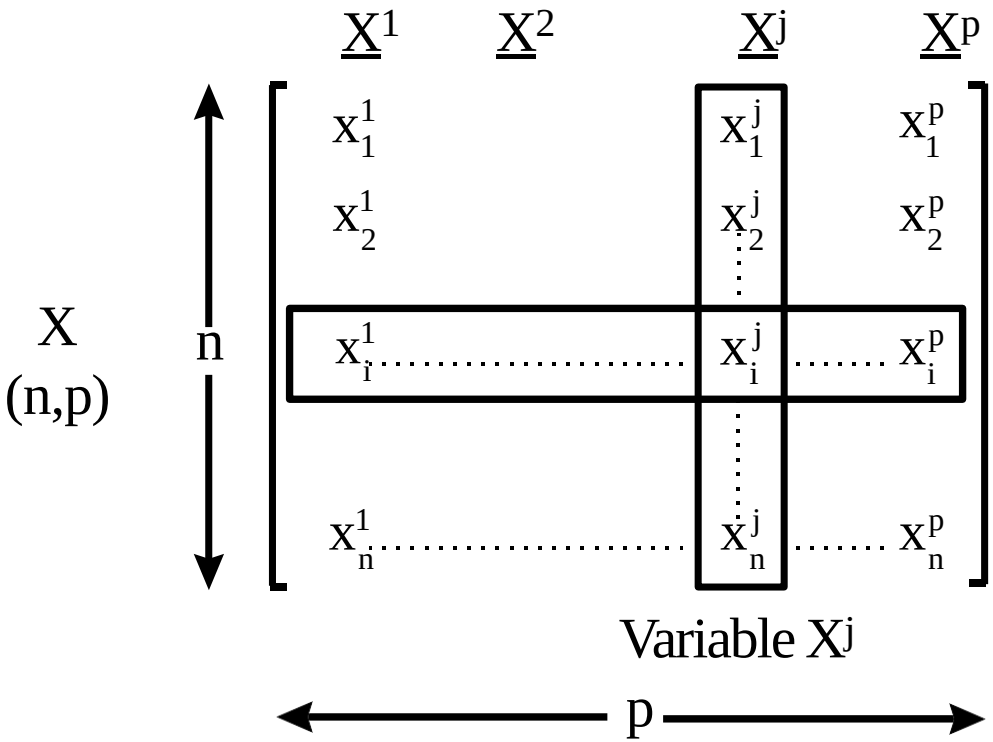
Expliquer la position des individus

I. L'ANALYSE EN COMPOSANTES PRINCIPALES

LE PROBLÈME

1. LES DONNÉES

p variables quantitatives observées sur n individus.



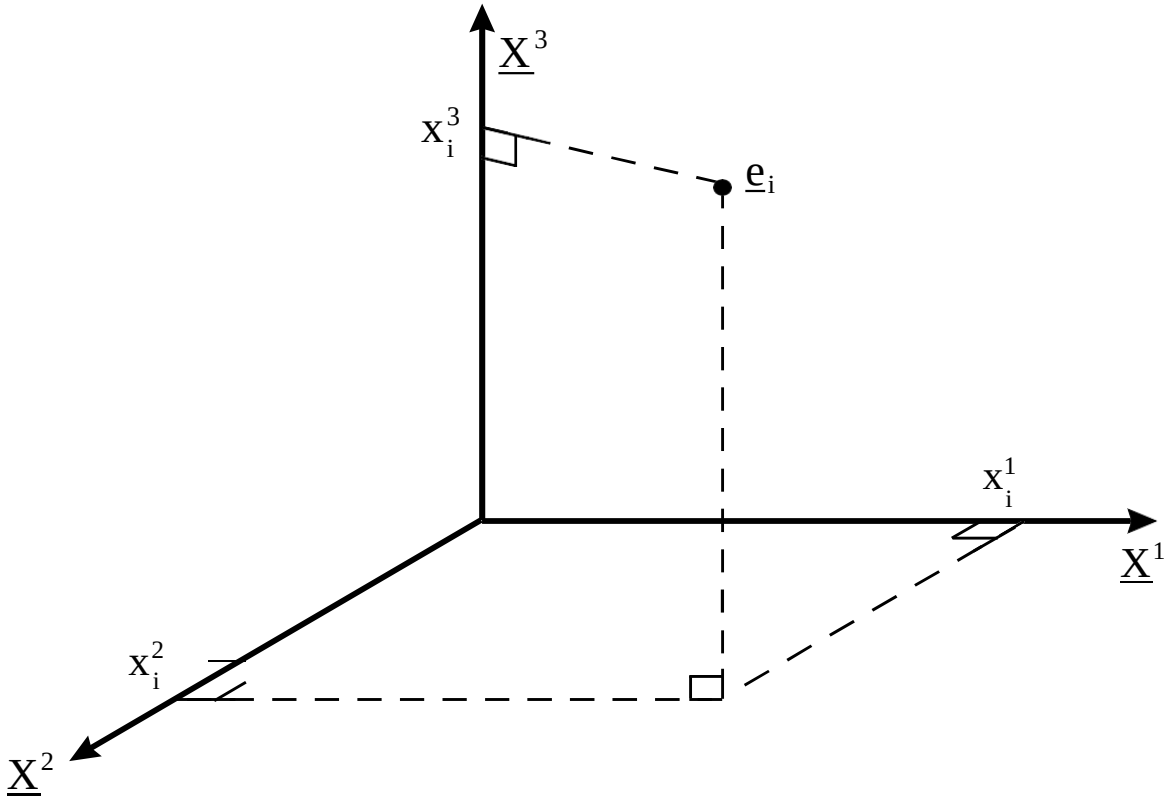
INDIVIDU = Élément de R^p

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.

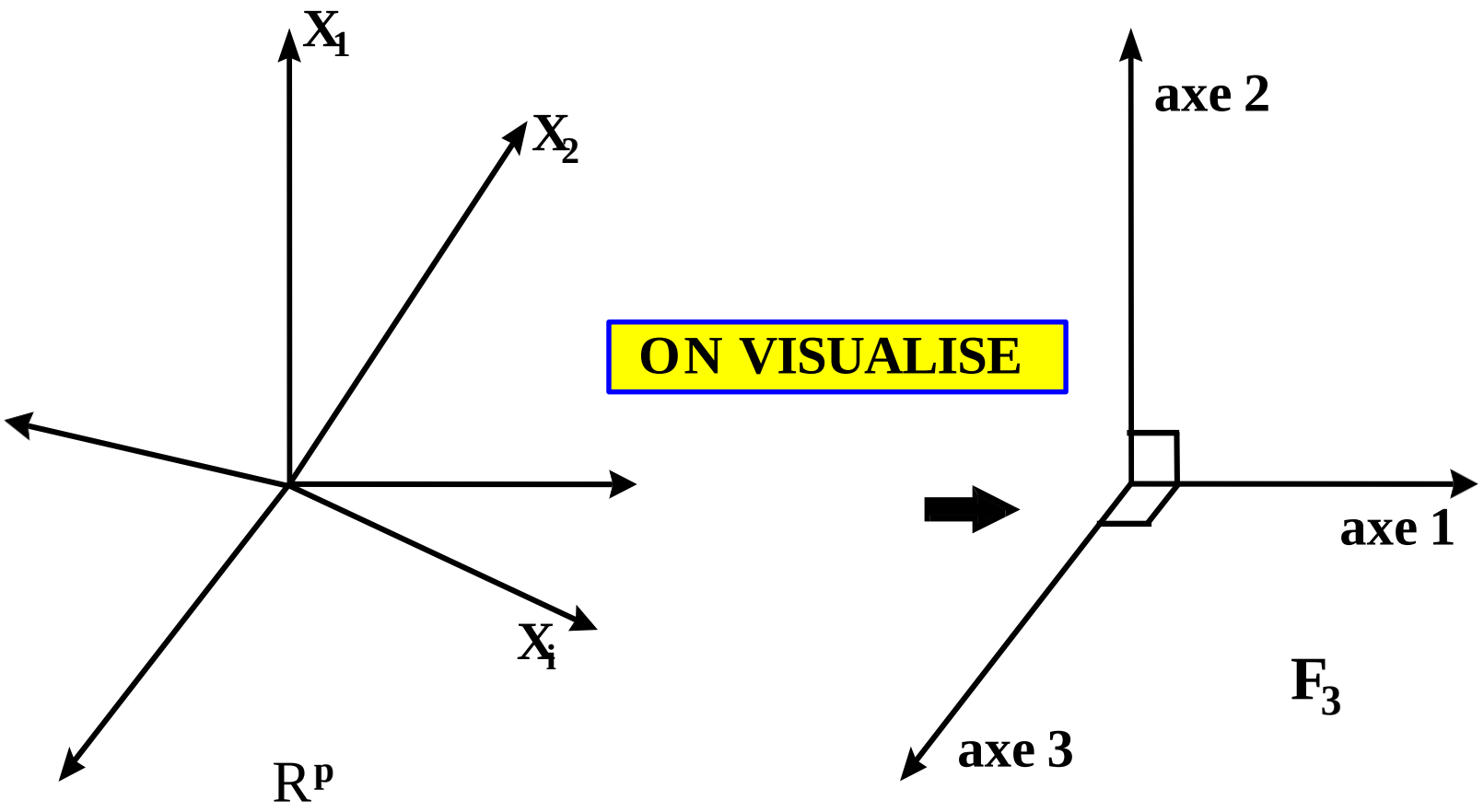


Impossible à visualiser dès que p > 3.

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

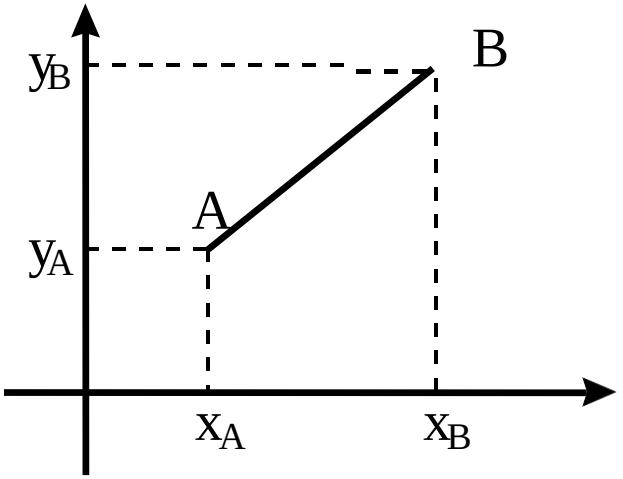
On cherche une représentation des n individus , dans un sous-espace F_k de R^p de dimension k (k petit 2, 3 …; par exemple un plan)

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



axes principaux

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.

4. INERTIE TOTALE

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

ou de façon plus générale

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

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

L'inertie est la somme pondérée des carrés des distances des individus au centre de gravité

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

On appelle inertie la quantité d'information contenue dans un tableau de données.

Une inertie nulle signifie que tous les individus sont presque identiques.

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

En notant V la matrice de variances-covariances :

$$V = \begin{pmatrix} var(v_1) & cov(v_1,v_2) \cdots \cdots & cov(v_1,v_j) \cdots\cdots & cov(v_1,v_p) \\ cov(v_2,v_1) & var(v_2) \cdots \cdots & cov(v_2,v_j) \cdots\cdots & cov(v_2,v_p) \\ \vdots & \vdots & \vdots & \\ cov(v_j,v_1) & cov(v_j,v_2) \cdots \cdots & var(v_j) \cdots\cdots & cov(v_j,v_p) \\ \vdots & \vdots & \vdots & \\ cov(v_p,v_1) & cov(v_p,v_2) & cov(v_p,v_j) \cdots\cdots & var(v_p) \end{pmatrix}$$

$$I_g = \sum_{i=1}^{p} \text{Var (vj)}$$

$$I_g = \text{Tr}(V)$$

Remarque

Dans le cas où les variables sont centrées réduites,

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

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 ou sur la matrice de corrélation**

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

$$\underline{c}^1 = u^1_1 \underline{x}^1 + u^1_2 \underline{x}^2 + \ldots u^1_p \underline{x}^p$$

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.**

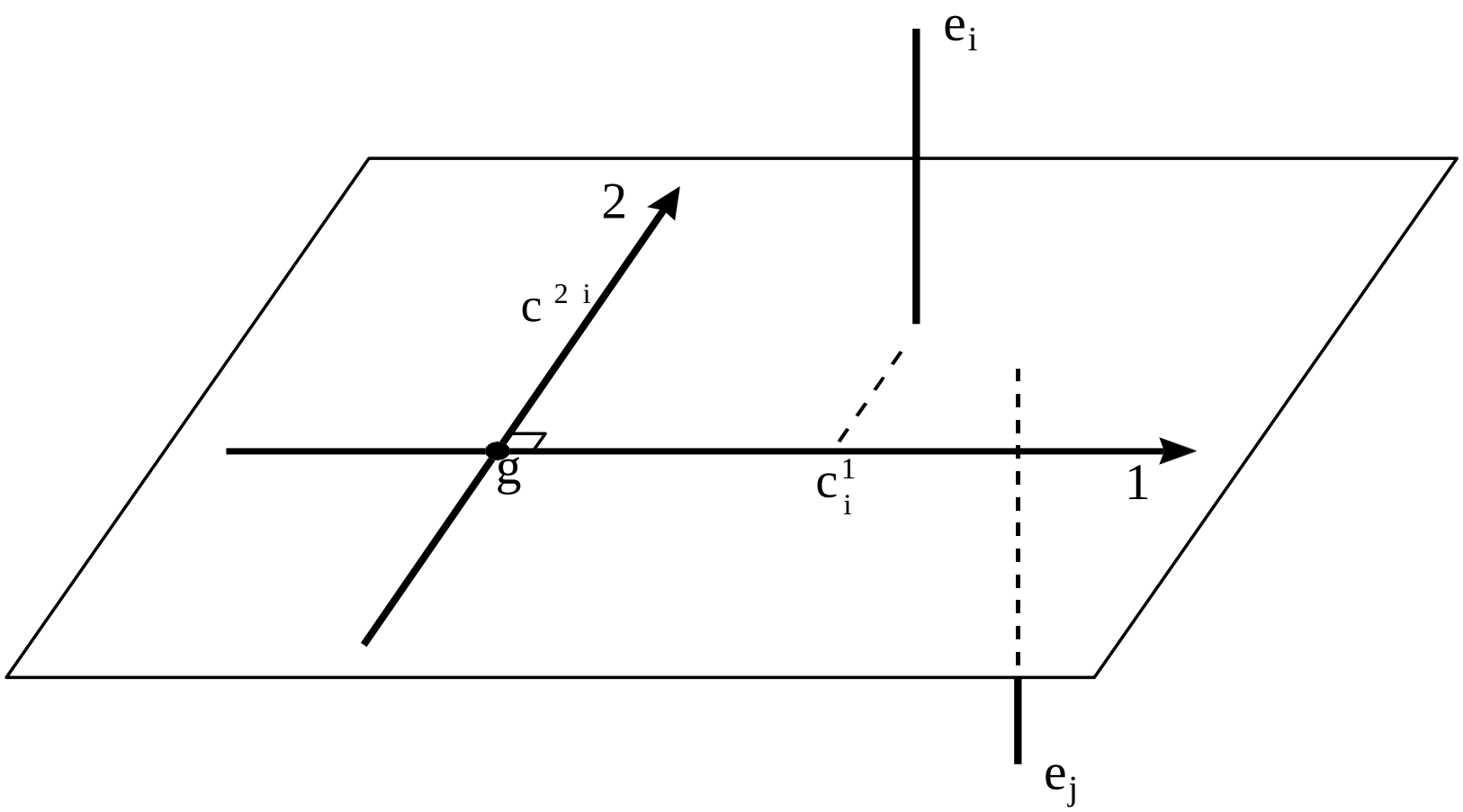
En effet, les axes associés sont orthogonaux.

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.**

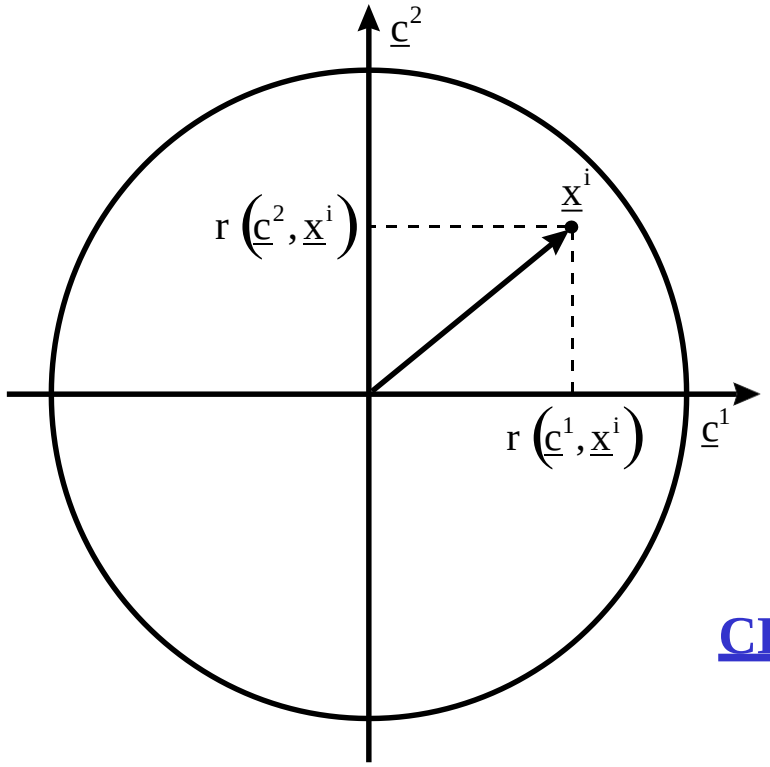


Attention à la qualité de représentation de chaque individu!

4. REPRÉSENTATION DES VARIABLES

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

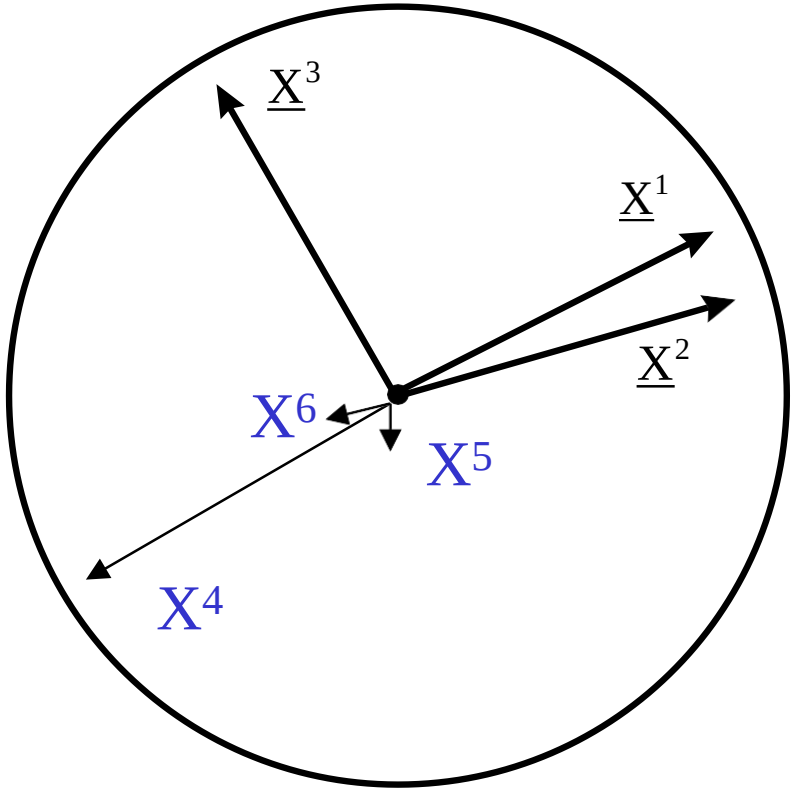
$r\left(\underline{c}^j, \underline{x}^i\right)$ est le **coefficient de corrélation linéaire** entre $\underline{c}^j$ et $\underline{x}^i$



CERCLE DES CORRÉLATIONS

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**

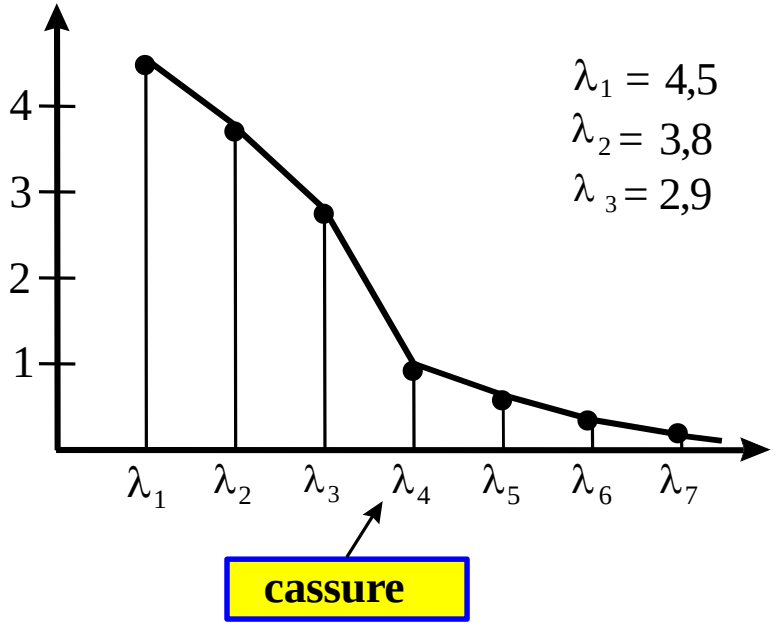
2 Diviser l'inertie totale par le nombre de variables initiales

$\Rightarrow$ 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.



Choix du nombre d'axes à retenir

Deux critères empiriques pour sélectionner le nombre d'axes :

Critère du coude : sur l'eboulis des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement

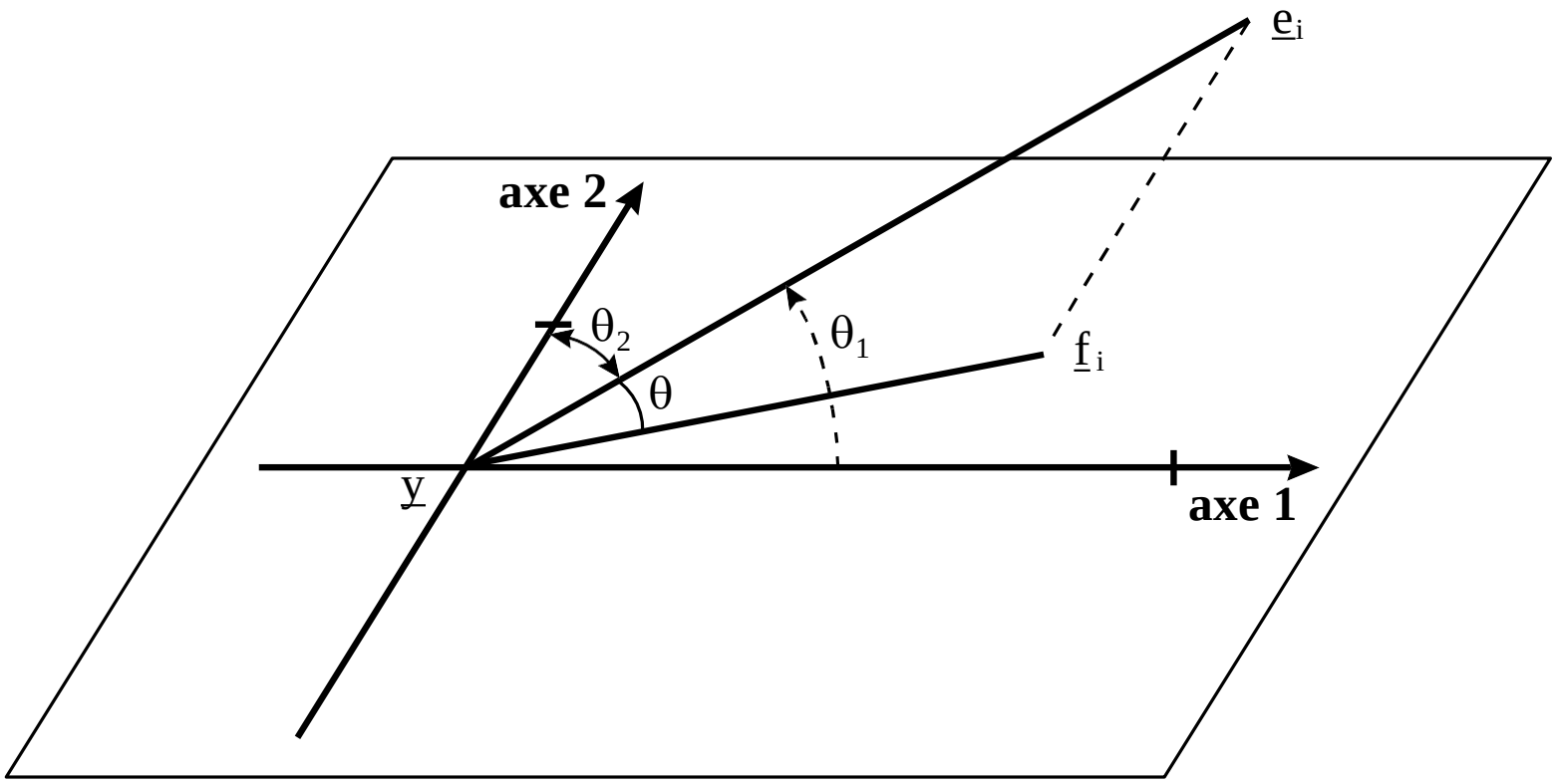
Critère de Kaiser: on ne retient que les axes dont l'inertie est supérieure à l'inertie moyenne I/p (un peu étroit).

Kaiser en ACP normée: I/p= 1 : On ne retiendra que les axes associés à des valeurs propre supérieures à 1

Dans la pratique, on retient en fait les q axes que l'on sait interpréter

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 $\underline{e}_i$

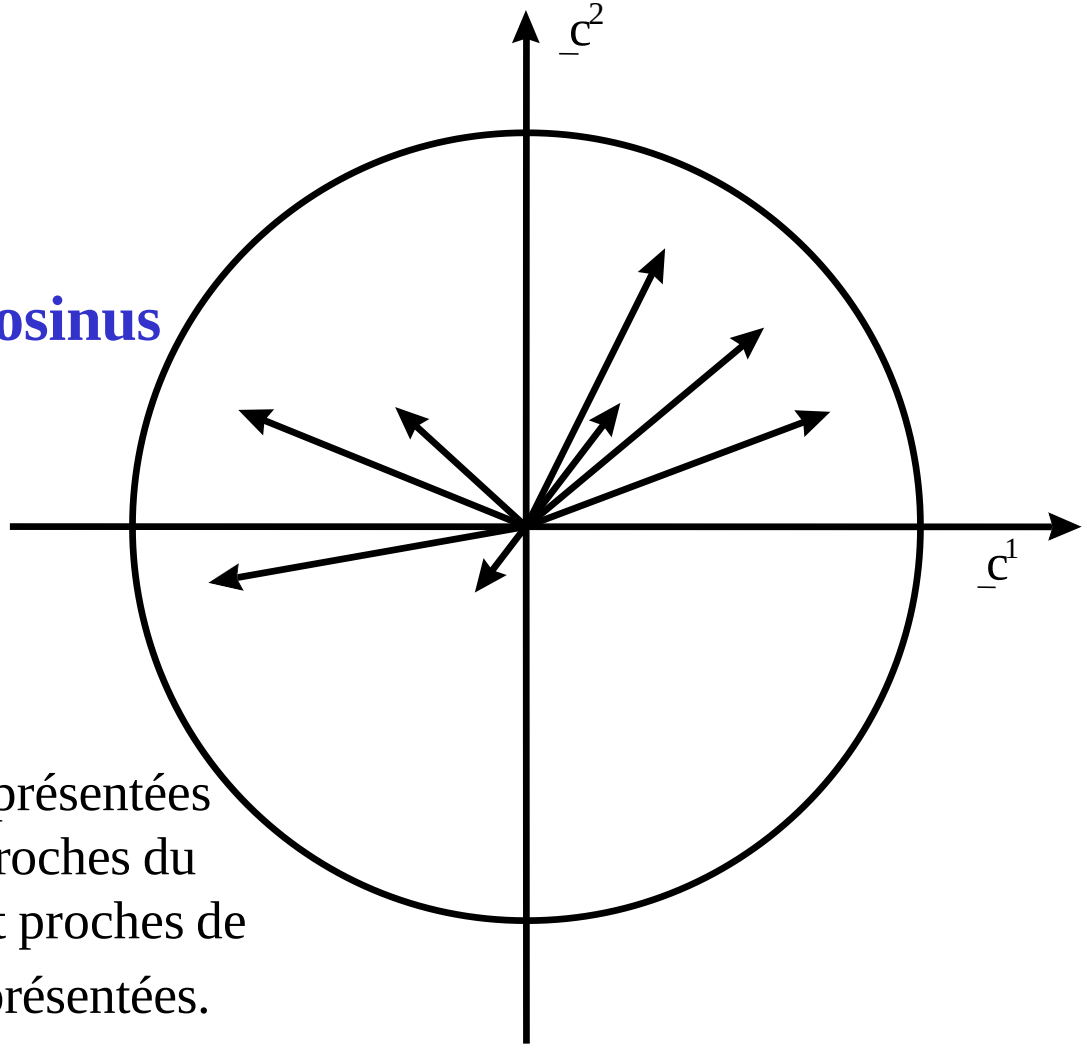
à la composante n° k est définie par

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

3. REPRÉSENTATION DES VARIABLES

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

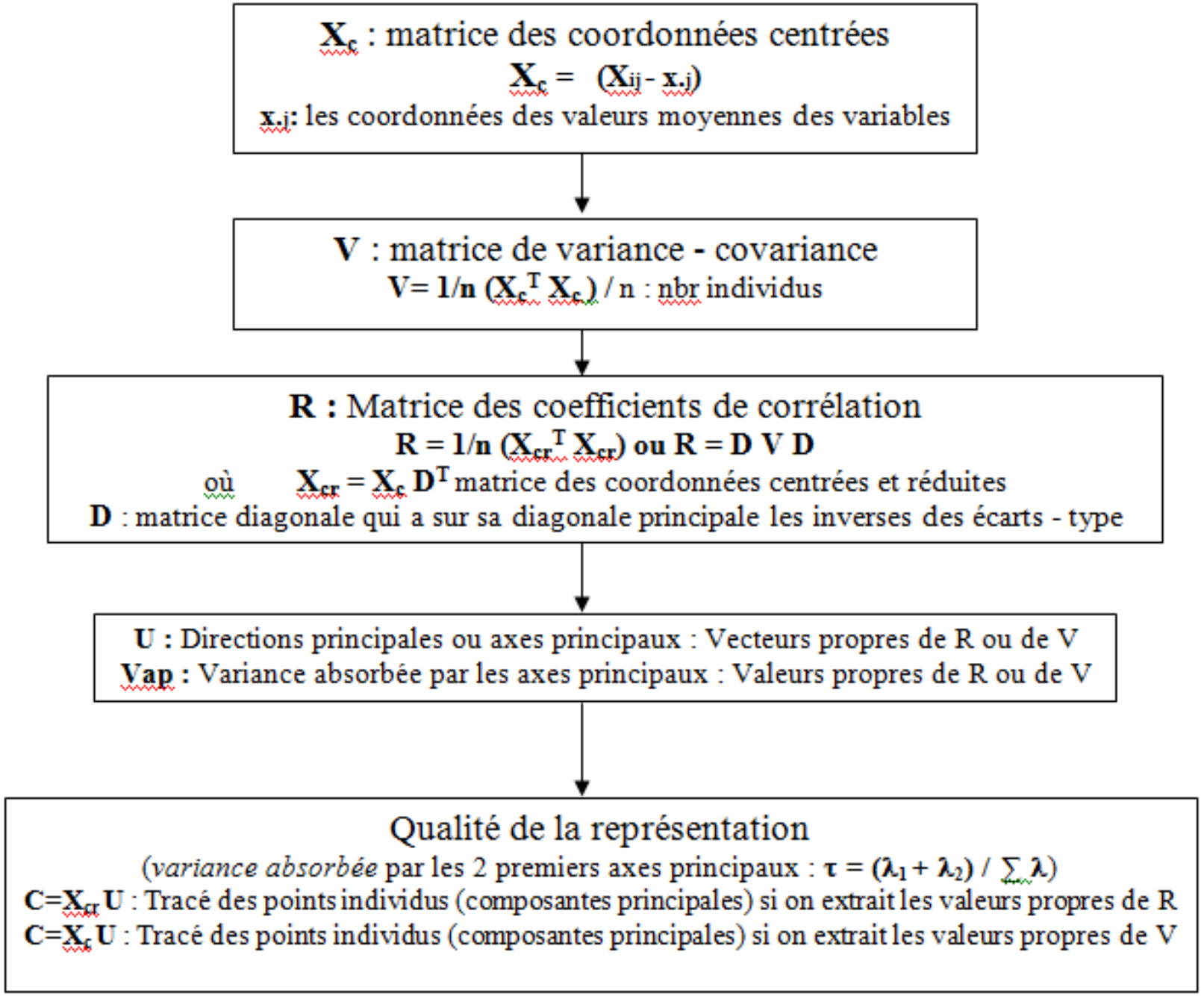
corrélation = cosinus



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

Organigramme d'ACP

Analyse dans R^P


X_c : matrice des coordonnées centrées
$X_c = (X_{ij} - x_{.j})$
$x_{.j}$: les coordonnées des valeurs moyennes des variables
V : matrice de variance - covariance
$V = 1/n\ (X_c^T X_c)$ / n : nbr individus
R : Matrice des coefficients de corrélation
$R = 1/n\ (X_{cr}^T X_{cr})$ ou $R = D\ V\ D$
où $X_{cr} = X_c\ D^T$ matrice des coordonnées centrées et réduites
D : matrice diagonale qui a sur sa diagonale principale les inverses des écarts - type
U : Directions principales ou axes principaux : Vecteurs propres de R ou de V
Vap : Variance absorbée par les axes principaux : Valeurs propres de R ou de V
Qualité de la représentation
(*variance absorbée* par les 2 premiers axes principaux : $\tau = (\lambda_1 + \lambda_2) / \sum \lambda$)
$C = X_{cr}\ U$: Tracé des points individus (composantes principales) si on extrait les valeurs propres de R
$C = X_c\ U$: Tracé des points individus (composantes principales) si on extrait les valeurs propres de V

Analyse dans R^N

$$\mathbf{Y}_i = \sqrt{\lambda_i}\,\mathbf{U}_i$$

Y : Matrice des points-variables

↓

Tracé du cercle des corrélations