Ministère de l'enseignement supérieur et de la recherche scientifique
Université de JIJEL
Faculté des Sciences Exactes et Informatique
Département d'Informatique

ACP

Analyse en Composantes Principales

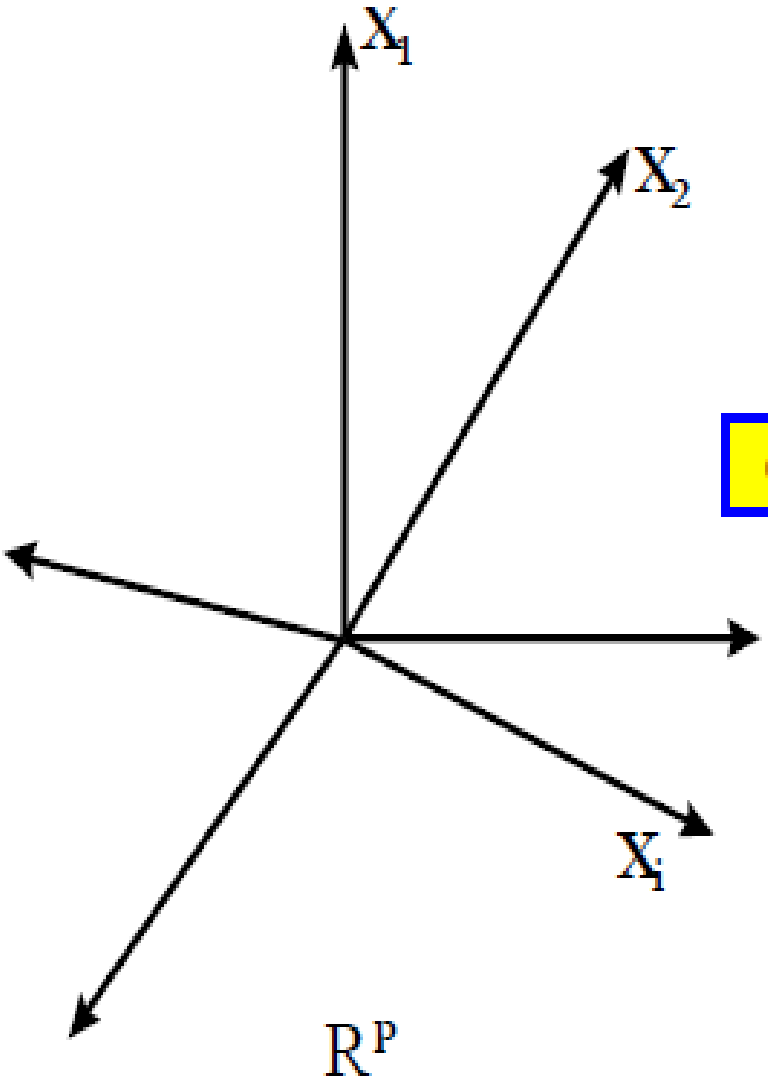
Introduction

- Les méthodes factorielles

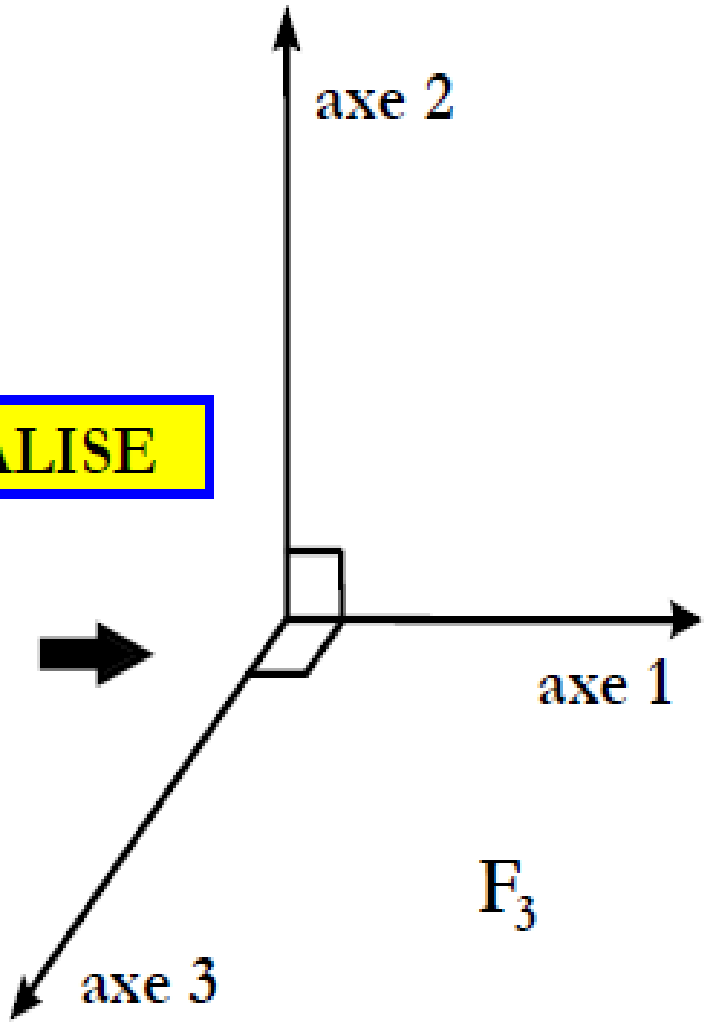
  - la projection sur un espace de dimension inférieur,

  - Une visualisation de l'ensemble des liaisons entre variables ,

  - Réduire le nombre de variables, tout en minimisant la perte de l'information.



ON VISUALISE



axes principaux

Introduction

- L'ACP (Hotelling, 1933) a pour objectif de réduire le nombre de données, souvent très élevé, d'un tableau de données :

  - Algébriquement: **matrice**,

  - Géométriquement : **nuage de points**.

Introduction

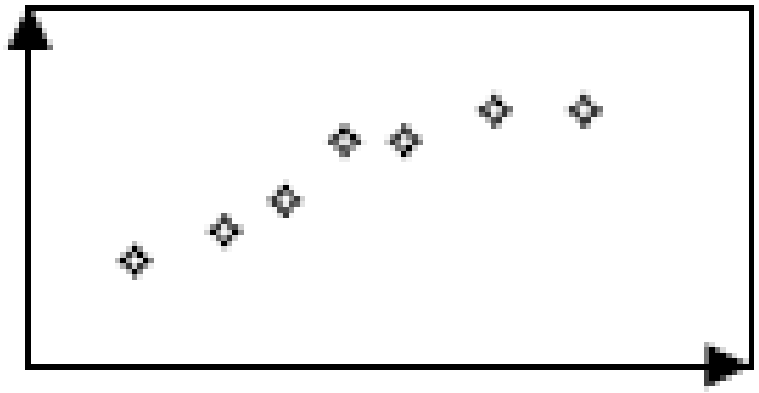
- L'ACP consiste en l'étude des **projections** des **points** de ce **nuage** sur:

  - **un axe**, **un plan** ou **un hyperplan**

(Mathématiquement: des sous-espaces vectoriels).

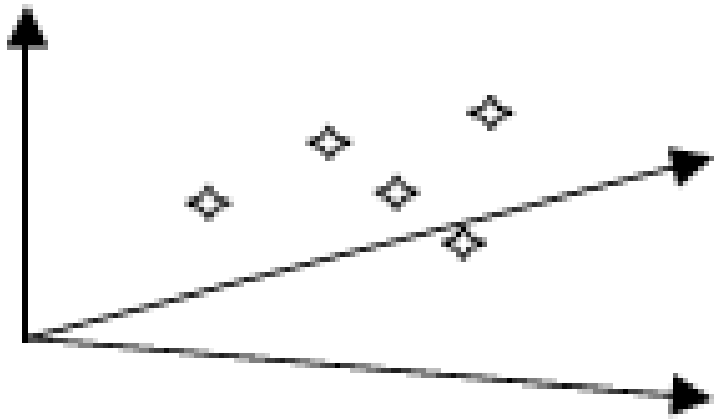
La représentation graphique

- Lorsqu'il n'y a que **deux dimensions** (exemple: largeur et longueur), il est facile de représenter les données sur un plan :



La représentation graphique

- Avec **trois dimensions** (largeur, hauteur et profondeur par ex.), c'est déjà plus difficile :



La représentation graphique

- Mais au delà de **3 dimensions**, il est impossible de représenter les données sur un plan ou même de les visualiser mentalement.

Projeter la réalité sur un plan

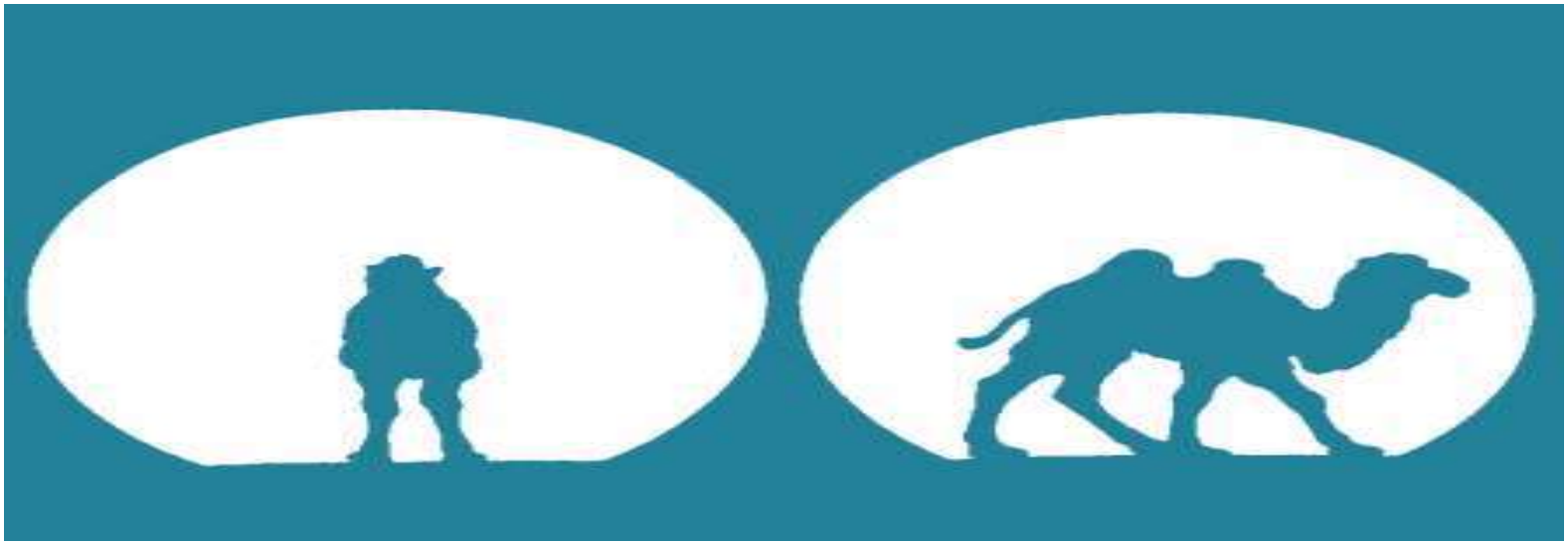


Figure de J.P.Fenelon

- Selon le point de vue, l'information retenue ne sera pas la même.

- L'ACP nous propose un point de vue permettant de voir au mieux les individus d'un tableau.

Résumer les données

- Lorsqu'on projette les données sur un plan, on obtient un graphique déformé de la réalité.

- Le rôle de l'ACP est de trouver des espaces de dimensions plus petites minimisant ces déformations.

Données et leurs caractéristiques

- **Tableau des données**

  Chaque tableau contient des lignes qui représentent les individus et des colonnes qui représentent les variables.

  Ce tableau rectangulaire (**matrice**) qu'on note par **X** possède des observations à n individus et p variables.

  Il a la forme suivante :

Données et leurs caractéristiques

- **Tableau des données**

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{ij} & & \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \in M_R(n, p),$$

où **xij** est la valeur prise par la **variable j** sur l'**individu i**.

Individu=Élément de R^p

Variable = Élément de R^n

Données et leurs caractéristiques

Individus et variables

- **Individu**: Le ieme individu est un vecteur à p composantes réelles qu'on le note par ei tel que

$$\mathbf{ei = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in R^p; \text{ pour } i = 1, n}$$

Données et leurs caractéristiques

Individus et variables

- **Variable**: La j eme variable est la liste des n valeurs qu'elle prend sur n individus, on la note par xj tel que:

$$\mathbf{x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^t \in R^n; \text{ pour } j = 1, p:}$$

Données et leurs caractéristiques

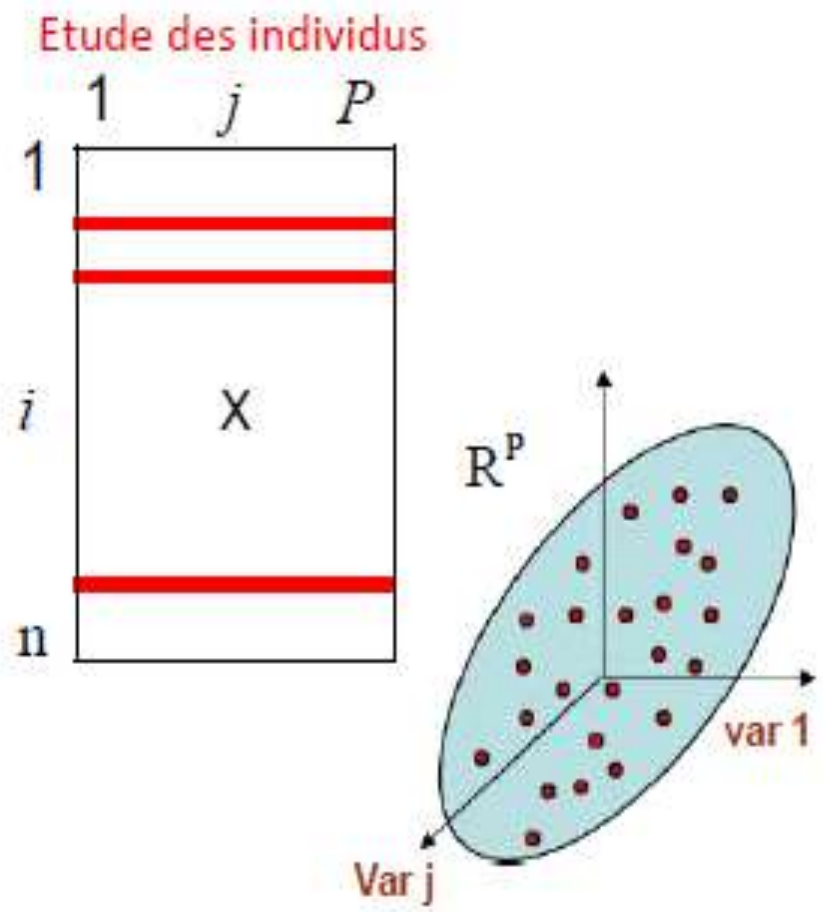
- L'A.C.P: permet d'explorer les liaisons entre variables et les ressemblances entre individus.
- **Résultats:**

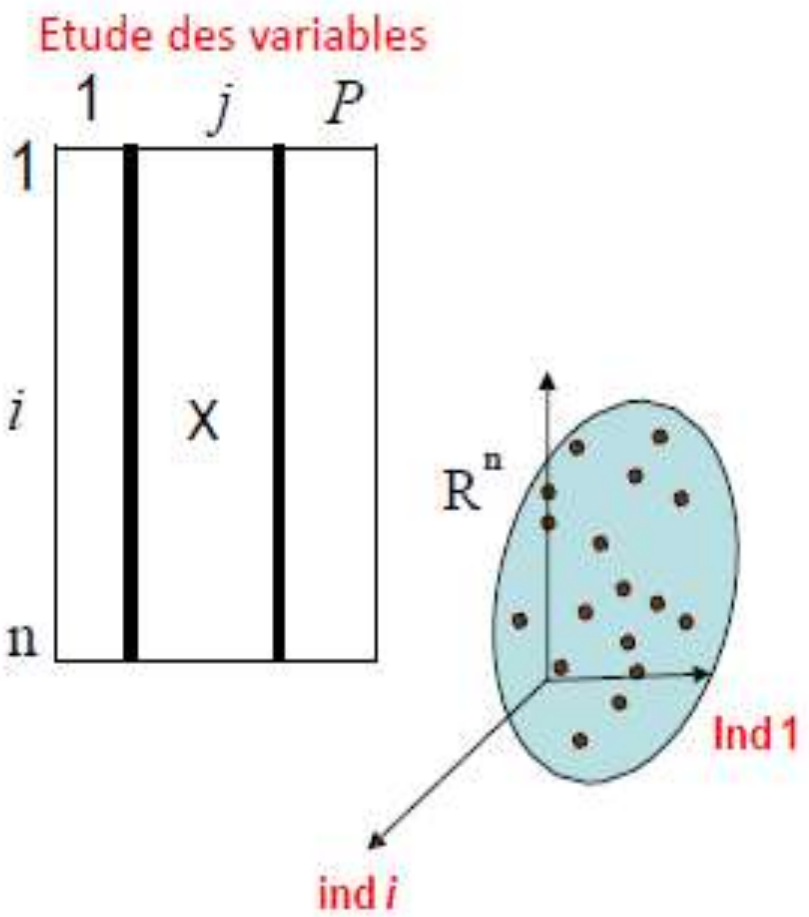
  - **Visualisation des individus** (Notion de distances entre individus)

  - **Visualisation des variables** (en fonction de leurs corrélations)

Deux nuages de points

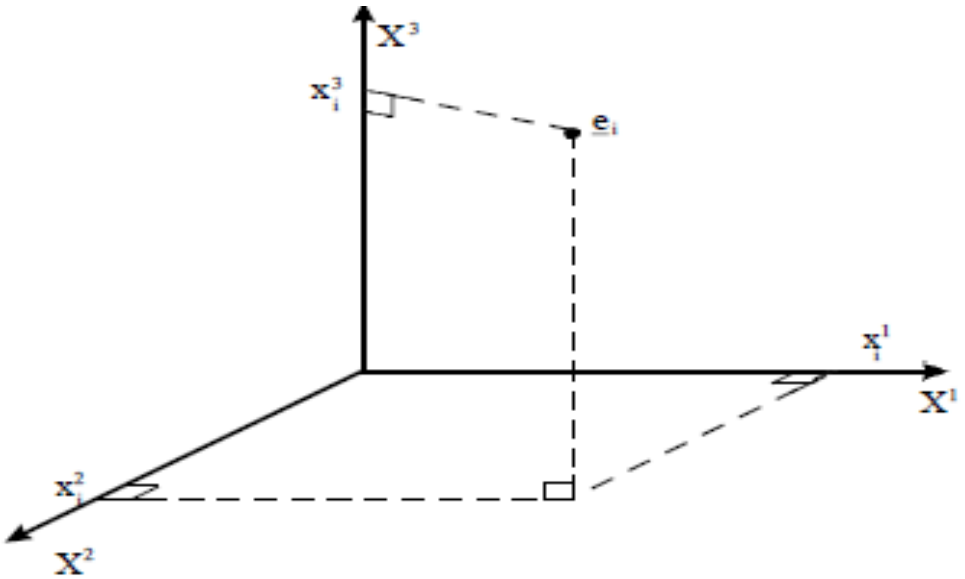
- Le tableau peut être vu comme un ensemble de lignes ou un ensemble de colonnes.





Nuage des individus

- A chaque individu noté ei, on peut associer un point dans Rp= espace des individus.
- A chaque variable du tableau X est associé un axe de Rp.



Impossible à visualiser dès que p > 3.

Principe de l'ACP

- On cherche une représentation des n individus, dans un sous-espace $\mathbf{F_k}$ de $\mathbf{R^p}$ de dimension **k** (k petit 2, 3) (k<p)

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

Principe de l'ACP

- Ces variables seront appelées «***composantes principales »***
- les axes qu'elles déterminent : « ***axes principaux »***
- les formes linéaires associées : « ***facteurs principaux »***

Perdre le moins d'information possible:

1

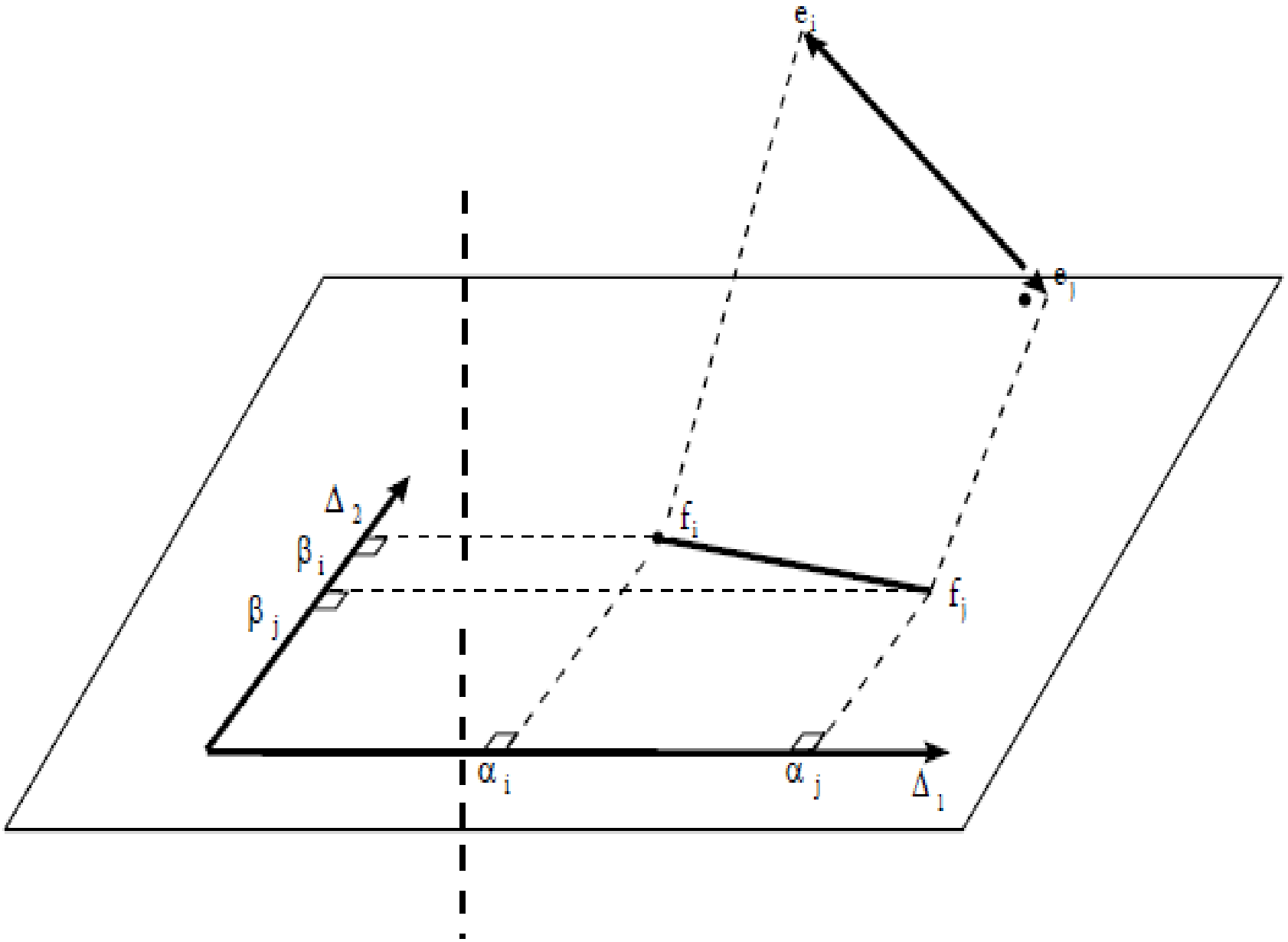
F_k devra être « **ajusté** » le mieux possible au nuage des individus: la somme des carrés des **distances** des individus à F_k doit être **minimale**

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**

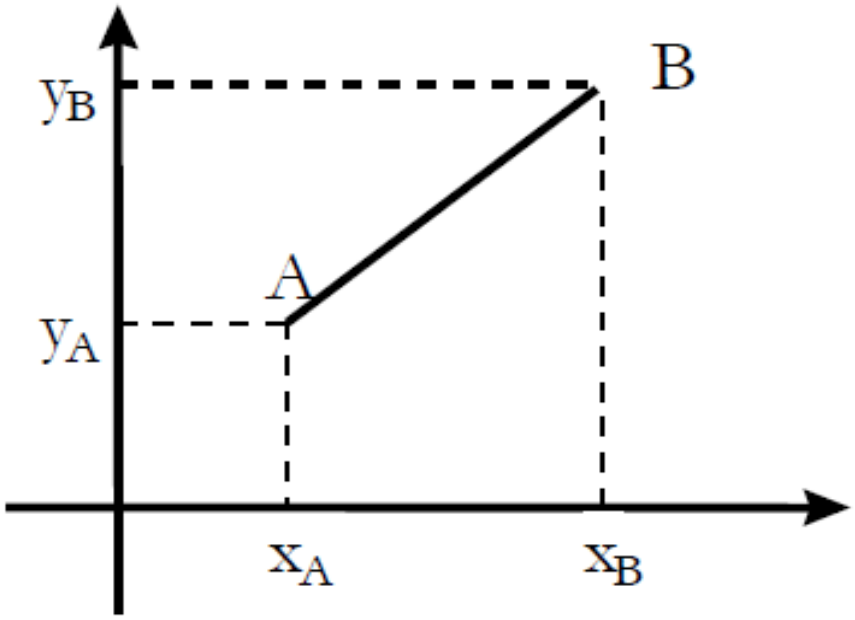


La distance entre fi et fj est inférieure ou égale à celle entre ei et ej

LE CHOIX DE LA DISTANCE ENTRE INDIVIDUS

- Dans le plan:

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



LE CHOIX DE LA DISTANCE ENTRE INDIVIDUS

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

$$e_i = \left(x_i^1\ x_i^2\ \ldots\ x_i^p\right) \qquad e_j = \left(x_j^1\ x_j^2\ \ldots\ x_j^p\right)$$

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

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

Inertie 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.**

- Si les j variables sont centrés-réduits, l'**inertie sera égale à j.**

Inertie totale du nuage de points

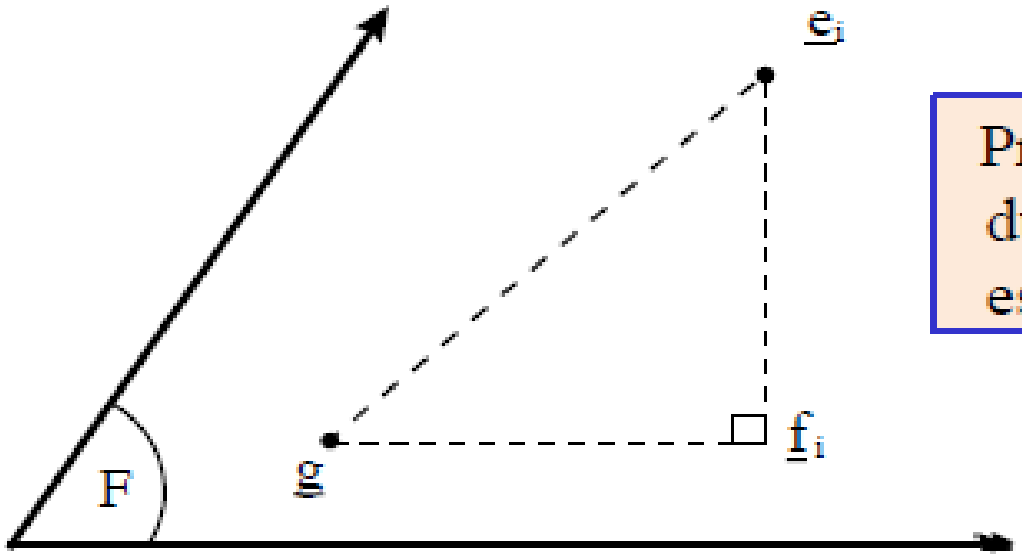
- On note l'inertie totale du nuage de *points Ig=mesure de* ***dispersion des points au sein du nuage*** = somme pondérée des **carrés des distances par rapport au centre de gravité *G* *du nuage***

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

ou de façon plus générale

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

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 e_i sur F

$$\|e_i - g\|^2 = \|e_i - f_i\|^2 + \|f_i - g\|^2 \qquad \forall i = 1 \ldots n$$

Équivalence des deux critères concernant la perte d'information

On va chercher F tel que :

$$\sum_{i=1}^{n} p_i \left\| e_i - f_i \right\|^2 \text{ soit minimal}$$

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

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

Équivalence des deux critères concernant la perte d'information

$$\|e_i - g\|^2 = \|e_i - f_i\|^2 + \|f_i - g\|^2 \qquad \forall i = 1 \dots n$$

Donc : $\underbrace{\sum_{i=1}^{n} p_i \|e_i - g\|^2}_{\text{Inertie totale}} - \underbrace{\sum_{i=1}^{n} p_i \|e_i - f_i\|^2}_{\text{minimiser cette quantité (carrés des distances entre points individus et leurs projections)}} = \underbrace{\sum_{i=1}^{n} p_i \|f_i - g\|^2}_{\text{maximiser l'inertie du nuage projeté}}$

$\Leftrightarrow$

Les étapes pour déterminer la composante principale :

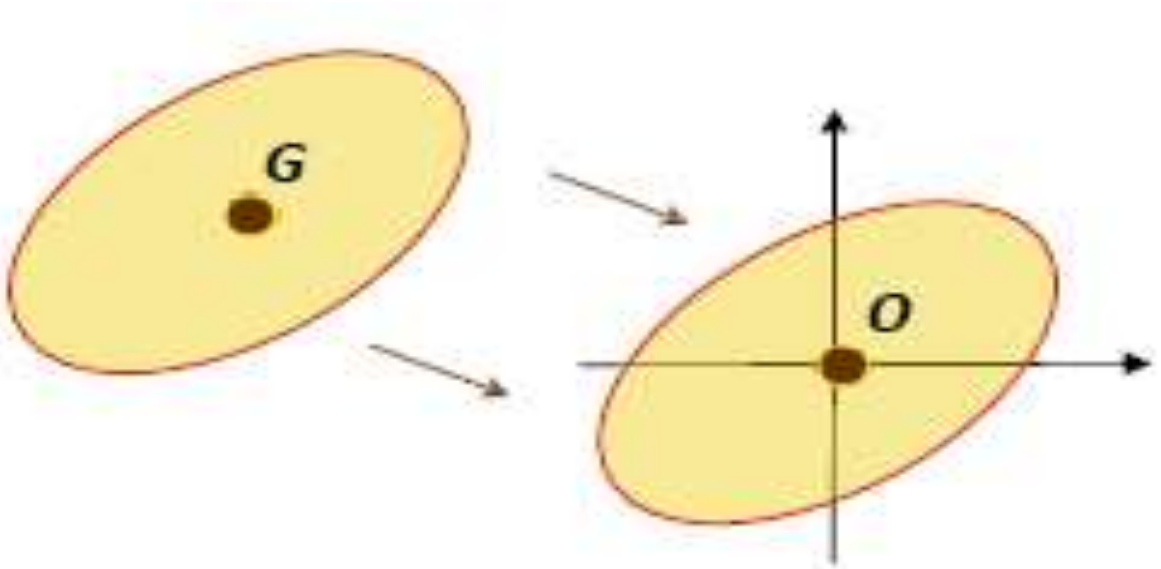
- **Centrage et réduction des données**

- **Déterminer les valeurs propres et les vecteurs propres sur la base de la matrice de corrélation entre les variables**

- **Déterminer les axes factoriels**
- **Sélectionner les composantes principales**

Centrage des données

- Le centrage est réalisé de façon **systématique en ACP**
- Translation du centre de gravité du nuage sur l'origine



- Centrer les données ne modifie pas la forme du nuage ⇒**toujours centrer**

Réduction des données

Exemple:

Echantillon1			Echantillon2	
Poids (g)	Diamètre (mm)		Poids (g)	Diamètre (cm)
100	70		100	7
95	65		95	6,5
6.25	6.25	Variance (inertie)	6.25	0.0065

Dans le premier cas, quand on va chercher le premier axe principale d'inertie, les variables poids et diamètre influencent de manière égale le calcul de l'axe (elles ont toutes deux une variance de 6.25)

Réduction des données

Exemple:

Mais dans le second cas, la variable poids « **pèsera beaucoup plus lourd** » que la variable diamètre dans le calcul, car 6.25 est bien plus grand que 0.0065.

C'est problématique, car le premier et le second cas représentent exactement les mêmes pommes

→**Réduire les données**

Réduction des données

- Plus la variable a un **écart-type élevé, plus elle apporte de l'inertie en projection et plus elle «attire les axes».**
- Or, l'écart type dépend directement de l'unité de mesure...

- Pour éviter d'accorder une plus grande importance aux variables exprimées arbitrairement avec de plus grandes valeurs, **on réduit les variables**

Réduction des données

- Transformer nos variables de telle manière que leur **moyenne soit égale à 0 (centrage)** et que **leur variance soit égale à 1 (la réduction)**

- Après avoir centré les données, si on les divise par leur écart type, alors on obtient des valeurs dont **la variance vaut 1**

Réduction des données

- Lorsque les variables sont exprimées dans des **unités de mesure différentes, → réduction systématique des données.**

- En cas **d'unités de mesure identiques** ?
- **Réduction : consiste à accorder une même importance à chaque variable**
- **Non réduction : accorde plus d'importance aux variables de forte dispersion**

Centrage et réduction des données

- Matrice Centrée Réduite est obtenue par la formule suivante :

$$x_{ij} = \frac{x_{ij} - \overline{X}}{\sigma_i}$$

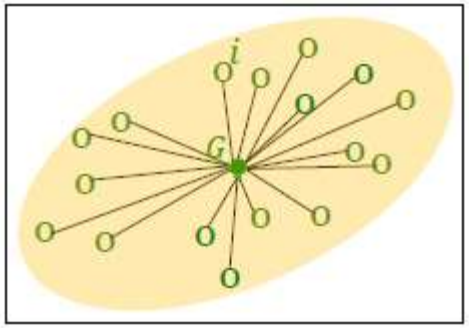
- **La moyenne** est un outil de calcul permet de résumer une liste de valeurs numériques en un seul nombre réel sans tenir compte de l'ordre de la liste.

$$\overline{X} = \frac{1}{n} \sum x_i$$

Centrage et réduction des données

- On appelle le point moyen ou centre de gravité le vecteur G des moyennes arithmétiques de chaque variable:

$$G = \left(\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_q\right)$$



- Lorsqu'on analyse des variables centrées, ce point moyen G sera le centre du repère considéré:

$$G \equiv O$$

Centrage et réduction des données

- **L'écart type** est un outil de calcul permet de mesurer **la dispersion des valeurs d'un échantillon.** C'est la racine carrée de la variance :

$$\sigma = \sqrt{variance}$$

- Avec la variance est la moyenne des carrées des écarts à la moyenne :

$$V = \frac{1}{n} \sum (x_i - \overline{X})^2$$

x_i : les valeurs de la variable

$\overline{X}$: la moyenne de la variable

Centrage et réduction des données

- Réduire ou normer donne la même dispersion, une même importance, à chaque variable (dans l'espace, elles ont même longueur:1)

→On dit que l'on réalise une **ACP normée**

- Ne pas réduire ou ne pas normer laisse à chaque variable son écart-type initial ce qui conduit à accorder à chaque variable une importance proportionnelle à son écart-type.

→On dit que l'on réalise une **ACP non normée (simple)**

Recherche des axes factoriels

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

Recherche des axes factoriels

- En d'autres termes, on effectue un changement de repère dans Rp 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.

Recherche des axes factoriels

- Cette réorganisation s'appuie sur la **diagonalisation de la matrice de variances-covariances (matrice de corrélations pour des données centrées-réduites).**

- Les **axes principaux sont ceux maximisant la variance projetée:** ce sont les **vecteurs propres normés** associés aux plus **grandes valeurs propres** de la matrice de **covariance/corrélation**.

Recherche des axes factoriels

- les **vecteurs propres normés à 1**(axes de direction ou axes factoriels)

- les **valeurs propres** (inerties associées aux axes)

Recherche des axes factoriels

- Le premier axe est celui associé à la plus grande valeur propre . On le note u1

- Le deuxième axe est celui associé à la deuxième valeur propre . On le note u2
- …

Matrice des variances covariances

- La Matrice des variances covariances permet de mesure la liaison linéaire qui peut exister entre un couple de variables statistiques

Var X1	**Cov(X1,X2)**	**Cov(X1,X3)**
Cov(X2,X1)	**Var X2**	**Cov(X2,X3)**
Cov(X3,X1)	Cov(X3,X2)	**Var X3**

- Si Cov(X2,X1) =0→les variables X1 et X2 sont indépendantes
- Si Cov(X2,X1) ≠ 0→les variables X1 et X2 sont dépendantes (existe une relation linéaire entre les variable)

Matrice des variances covariances

- Obtenue par la formule suivante :

$$V = 1/n * {}^{t}Mc * Mc$$

- Mc : Matrice centrée
- ${}^{t}Mc$: Matrice centrée transposée

Matrice des corrélations

- Matrice des corrélations entre variables permet d'analyser les relations bilatérales entre les variables :
- Obtenue par la formule suivante :

$$\Gamma = 1/n \; {}^{t}Mcr * Mcr$$

Mcr : Matrice centrée réduite

^{t}Mcr : Matrice centrée réduite transposée

Calculer les valeurs propres

- Déterminer le polynôme caractéristique :

Det $|X-\lambda I|$

- Calculer les **valeurs propres λ**

- Déterminer **les vecteurs propres** orthogonaux associés aux valeurs propres

Caractères des composantes principales

- **Il n'y a pas de redondance** d'information entre **deux composantes principales**.

- Les composantes principales sont **centrées**.

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

Caractères des composantes principales

- Soit u1 est le vecteur propre associé à la première grande valeur propre λ1. Soit donc Δ_{u1} ce premier axe principal.
- Si on veut chercher un deuxième axe Δ_{u2} , où u2 est son vecteur unitaire orthogonal à u1 (c-à-d <u1, u2>=0)

Caractères des composantes principales

- C-à-d le vecteur unitaire u2 de la droite Δ_{u2} est le vecteur propre associé à la deuxième plus grande valeur propre λ2, il est orthogonal à u1.

- Ainsi de suite on cherche le troisième axe et …… jusqu'au q ieme axe, q<p

Choix du nombre de facteur à retenir

Le critère qui permet de choisir le nombre d'axes principaux à retenir utilisé est celui de pourcentage inertie totale expliquée:

Caractères des composantes principales

- Soient $\Delta_{u1}\, \Delta_{u2}\, \ldots\, \Delta_{uq}$ les q premiers axes principaux de vecteurs unitaires u1, u2, …, uq.

- On appelle pourcentage d'inertie expliquée par l'axe Δ_{uj} la quantité suivante définie par :

$$\frac{\lambda j}{I_N(O)} = \frac{\lambda j}{Tr(V)} \qquad j=1..q$$

Représentation des individus

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

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

$$I_g = Tr(V)$$

- **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).**

Représentation des individus

- Supposons que nous avons retenu q axes principaux, q<=p. Alors on doit effectuer la projection des individus xi ϵR^p dans l-hyper plan H formés par les q axes principaux.

- La valeur de la projection de xi sur l'axe Δ_{ul} notéé c(i,l) est donnée par

$$\mathbf{c(i,l)= xi * u_l}$$

Représentation des individus

- La jème composante principale fournit les coordonnées des n individus sur le **jème axe principal.**

$$\underline{c}^j = \begin{pmatrix} c_1^j \\ c_2^j \\ \vdots \\ c_n^j \end{pmatrix}$$

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

Représentation d'individus supplémentaires

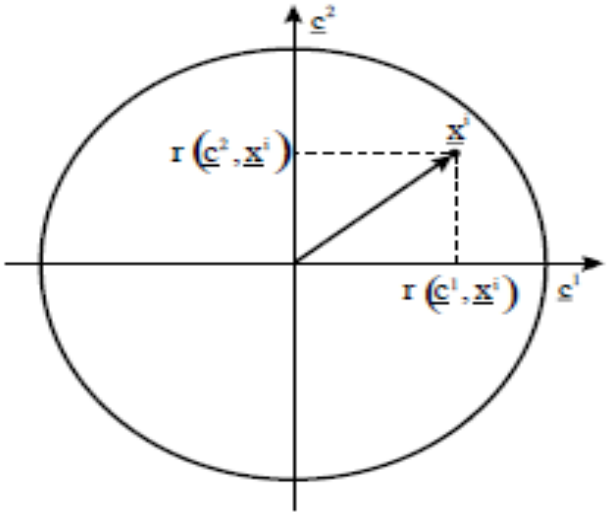
- On désire savoir où placer sur un graphique un nouveau individu x_k dont on connait ses coordonnés dans R^p

$$x_k = \begin{pmatrix} x^1_k \\ x^2_k \\ .. \\ x^p_k \end{pmatrix} \in R^p$$

Pour cela on calcule les coordonnées de xk dans le système des axes principaux, c-à-d on calcule les valeurs ${}^t x_k$ a1, ${}^t x_k$ a2, …, ${}^t x_k$ ap où les al sont les facteurs principaux, lk=1,p, al=Mul

Représentation des variables supplémentaires

- Les proximités entre les composantes principales et les variables initiales sont mesurées par les covariances, et sur tout **les corrélations.**
- R(cj,xi) est le **coefficient de corrélation linéaire entre cj et xi**

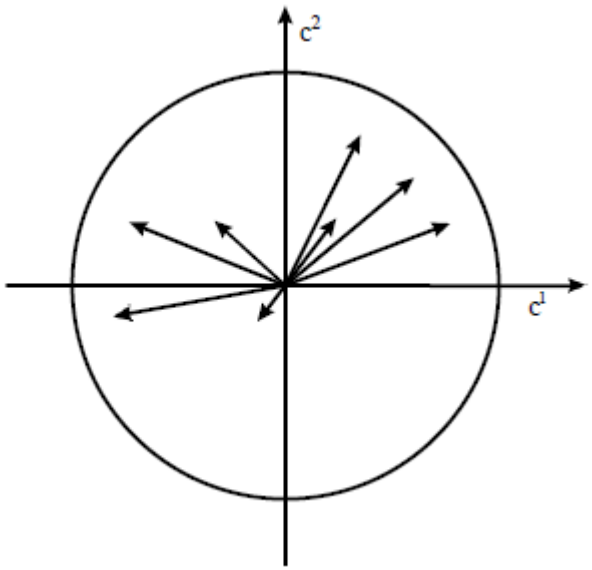


Cercle des corrélations

Représentation des variables supplémentaires

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

corrélation =cosinus

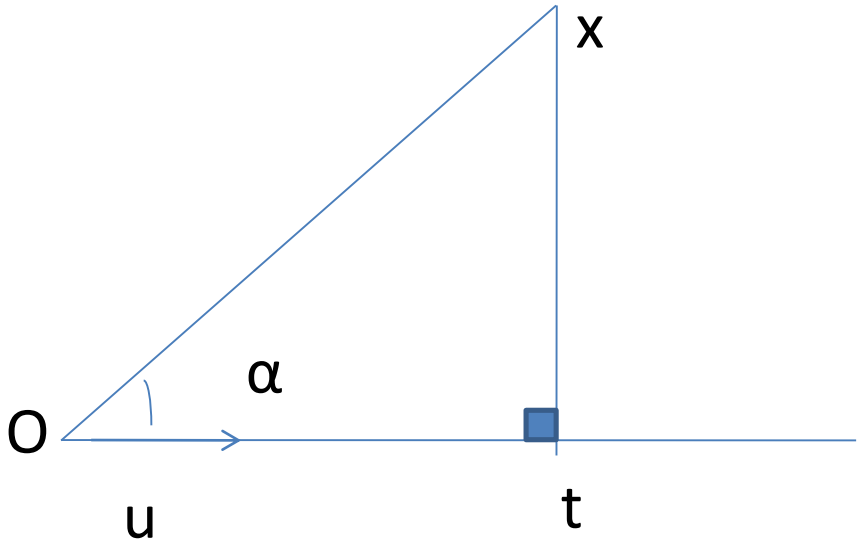


- Les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représentées.

La qualité de la représentation des individus

Pour avoir une idée sur la qualité de la représentation de chacun de n individus dans le sous espace constitué par les axes principaux, on calcul les cos^2 des angles compris entre l'individu xi et leurs projections dans les différents sous espaces.

La qualité de la représentation des individus



- Nous dirons qu'un individu est mieux représenté lorsque le **$\cos^2$** est **proche de 1**.

  (La valeur 1 nous l'obtenons si l'on retenait p axes)