Cours
Réduction de dimensionnalité (ACP)

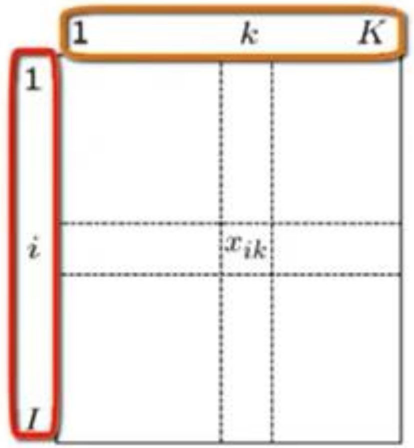
K. BELATTAR,

Département Informatique - Université d’Alger 1

11/12/2023

Etude de données via ACP

- L’analyse en composante principales (ou ACP) s’intéresse à des **tableaux de données** rectangulaires avec des individus en ligne et des variables quantitatives en colonnes.



Exemples d’application de l’ACP

- Ecologie: concentration du polluant k dans la rivière i
- Economie: valeur de l’indicateurs k pour l’année i
- Génétique: expression du gène k pour le patient i
- Biologie: mesure k pour l’animal i
- Marketing: valeur d’indice de satisfaction k pour la marque i
- Diagnostic médical: caractéristique k pour une image d’un patient i
- Etc.

Etude d'un tableau de données via ACP

Exemple: Tableau X de notes **(de 1 à 7**) attribuées **à P=7** mots, par **n=12** répondant

Mots / Répondants	arbre	cadeau	danger	morale	orage	politesse	sensuel
R01	7	4	2	2	7	1	6
R02	6	2	1	2	5	1	7
R03	4	3	2	2	3	4	4
R04	5	3	1	5	2	7	1
R05	4	5	2	7	1	4	2
R06	5	7	1	5	2	4	5
R07	4	2	1	3	5	3	6
R08	4	1	3	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	5	7	7	6	7
R12	2	2	1	2	1	3	4

<u>Objectif de l'ACP</u>: étudier ce tableau de données

Etude d’un tableau de données via ACP

Etude des individus

- Tableau: ensemble de *n* lignes

- Recherche des ressemblances entre les individus (proches, différents)

- Partition des individus: construire des groupes d’individus homogènes de point de vue de variables.

- Caractériser les différents groupes d’individus.

Etude d’un tableau de données via ACP

Etude de variables

- Tableau: ensemble de *P* colonnes
- Recherche des liaisons entre les variables
- Relations linéaires (simples et très fréquentes)

Objectif de l’étude :

- Mesurer, visualiser les corrélations entre les variables.
- Rechercher des indicateurs qui résument les variables (par exemple: la moyenne des valeurs d’une variable)

Etude d'un tableau de données via ACP

Tableau de données

Soit la table de données X_n^p suivantes avec $x^j(j$ = 1, … , p) variables et e_i individus (i = 1, … , n)

$$X_n^p = \begin{bmatrix} x_1^1, x_1^2, \dots \dots x_1^P \\ x_2^1, x_2^2, \dots \dots x_2^P \\ \dots \dots \dots \dots \dots .. \\ x_n^1, x_n^2, \dots \dots x_n^P \end{bmatrix}$$

n individus

P variables

Etude d’un tableau de données via ACP

Tableau de données

$$\underset{(n,p)}{X} : \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix} \qquad [x^1, x^2, \dots, x^j, \dots, x^p]$$

Variable

$$x^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{bmatrix}$$

Individu

$$e'_i = \begin{bmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^p \end{bmatrix}$$

Etude d’un tableau de données via ACP

Tableau de données

$$
\underset{(12,7)}{X} = \begin{pmatrix}
7 & 4 & 2 & 2 & 7 & 1 & 6 \\
6 & 2 & 1 & 2 & 5 & 1 & 7 \\
4 & 3 & 2 & 2 & 3 & 4 & 4 \\
5 & 3 & 1 & 5 & 2 & 7 & 1 \\
4 & 5 & 2 & 7 & 1 & 4 & 2 \\
5 & 7 & 1 & 5 & 2 & 4 & 5 \\
4 & 2 & 1 & 3 & 5 & 3 & 6 \\
4 & 1 & 3 & 4 & 5 & 4 & 7 \\
6 & 6 & 2 & 4 & 7 & 5 & 5 \\
6 & 6 & 3 & 5 & 3 & 6 & 6 \\
7 & 7 & 5 & 7 & 7 & 6 & 7 \\
2 & 2 & 1 & 2 & 1 & 3 & 4
\end{pmatrix}
$$

$$
\begin{aligned}
e_1 &= (7,\ 6,\ 2,\ 2,\ 7,\ 1,\ 6) \\
e_2 &= (6,\ 2,\ 1,\ 2\ \ 7\ \ 1\ \ 7) \\
e_3 &= (4,\ 3,\ 2,\ 2,\ 3,\ 4,\ 4) \\
e_4 &= (5,\ 3,\ 1,\ 5\ \ 2\ \ 7,\ 1) \\
e_5 &= (4,\ 5,\ 2,\ 7,\ 1,\ 4,\ 2) \\
e_6 &= (5,\ 7,\ 1,\ 5,\ 2,\ 4,\ 5) \\
e_7 &= (4,\ 2,\ 1,\ 3,\ 5,\ 3,\ 6) \\
e_8 &= (4,\ 1,\ 3,\ 4,\ 5,\ 4,\ 7) \\
e_9 &= (6,\ 6,\ 2,\ 4,\ 7,\ 5,\ 5) \\
e_{10} &= (6,\ 6,\ 3,\ 5,\ 3,\ 6,\ 6) \\
e_{11} &= (7,\ 7,\ 5,\ 7,\ 7,\ 6,\ 7) \\
e_{12} &= (2,\ 2,\ 1,\ 2,\ 1,\ 3,\ 4)
\end{aligned}
$$

$$
x^1 = \begin{pmatrix}7\\6\\4\\5\\4\\5\\4\\4\\6\\6\\7\\2\end{pmatrix},\
x^2 = \begin{pmatrix}4\\2\\3\\3\\5\\7\\2\\1\\6\\6\\7\\2\end{pmatrix},\
x^3 = \begin{pmatrix}2\\1\\2\\1\\2\\1\\1\\3\\2\\3\\5\\1\end{pmatrix},\
x^4 = \begin{pmatrix}2\\2\\2\\5\\7\\5\\3\\4\\4\\5\\7\\2\end{pmatrix},\
x^5 = \begin{pmatrix}5\\7\\3\\2\\1\\2\\5\\5\\7\\3\\7\\1\end{pmatrix},\
x^6 = \begin{pmatrix}1\\1\\4\\7\\4\\4\\3\\4\\5\\6\\6\\3\end{pmatrix},\
x^7 = \begin{pmatrix}6\\7\\4\\1\\2\\5\\6\\7\\5\\6\\7\\4\end{pmatrix}
$$

Etude d’un tableau de données via ACP

- **Centre de gravité du nuage de points (point moyen)**: est le vecteur **g** des moyennes arithmétiques/pondérées de chaque variable :

$$\mathbf{g} = \sum_{i=1}^{n} p_i\,(e_i) = \begin{bmatrix} \sum_{i=1}^{n} p_i\,(x_i^1) \\ \vdots \\ \sum_{i=1}^{n} p_i\,(x_i^p) \end{bmatrix} = \begin{bmatrix} \bar{x}^1 \\ \vdots \\ \bar{x}^p \end{bmatrix}$$

$$\mathbf{g}' = (\bar{x}^1, \dots, \bar{x}^p)$$

- La moyenne arithmétique: $\bar{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$

- La moyenne pondérée: $\bar{x}^j = \sum_{i=1}^{n} p_i\, x_i^j$ sachant que p_i est le poids de chaque individu au sein de la population

Etude d’un tableau de données via ACP

Le centre de gravité g est écrit sous forme matricielle:

$$g = X'\, D_p 1_n$$

D_p: matrice de poids $\quad D_p = \begin{pmatrix} P_1 & & 0 \\ & P_2 & \\ 0 & & P_n \end{pmatrix}$ tel que P_i = $\boldsymbol{1/n}$ et $\sum_{i=1}^{n} P_i$ =1

1_n: matrice d'identité de taille n*n, $\quad 1_n = \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{pmatrix}$

Etude d'un tableau de données via ACP

$$x^1 = \begin{pmatrix} 7\\6\\4\\5\\4\\5\\4\\4\\6\\6\\7\\2 \end{pmatrix},\ x^2 = \begin{pmatrix} 4\\2\\3\\3\\5\\7\\2\\1\\6\\6\\7\\2 \end{pmatrix},\ x^3 = \begin{pmatrix} 2\\1\\2\\1\\2\\1\\1\\3\\2\\3\\5\\1 \end{pmatrix},\ x^4 = \begin{pmatrix} 2\\2\\2\\5\\7\\5\\3\\4\\4\\5\\7\\2 \end{pmatrix},\ x^5 = \begin{pmatrix} 5\\7\\3\\2\\1\\2\\5\\5\\7\\3\\7\\1 \end{pmatrix},\ x^6 = \begin{pmatrix} 1\\1\\4\\7\\4\\4\\3\\4\\5\\6\\6\\3 \end{pmatrix},\ x^7 = \begin{pmatrix} 6\\7\\4\\1\\2\\5\\6\\7\\5\\6\\7\\4 \end{pmatrix}$$

$$\bar{x}^1 = \sum_{i=1}^{12} p_i\, x_i^1 = \frac{1}{12}\sum_{i=1}^{12} x_i^1 = 5$$

$$\bar{x}^2 = 4,\quad \bar{x}^3 = 2,\quad \bar{x}^4 = 4,\quad \bar{x}^5 = 4,\quad \bar{x}^6 = 4,\quad \bar{x}^7 = 5$$

$$\mathbf{g} = \begin{bmatrix} 5\\4\\2\\4\\4\\4\\5 \end{bmatrix}$$

$$\mathbf{g}' = (5,\ 4,\ 2,\ 4,\ 4,\ 4,\ 5)$$

Etude d’un tableau de données via ACP

- La variance de X est définie par:

$$var(x) = \sigma_x^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2 \text{ ou } \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- La covariance observée entre deux variables x et y est :

$$\text{cov(x,y)} = \sigma_{x,y} = \sum_{i=1}^{n} p_i (x_i - \bar{x}) (y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{x}\,\bar{y}$$

- Le coefficient de corrélation cor(x,y) ou r est donnée par :

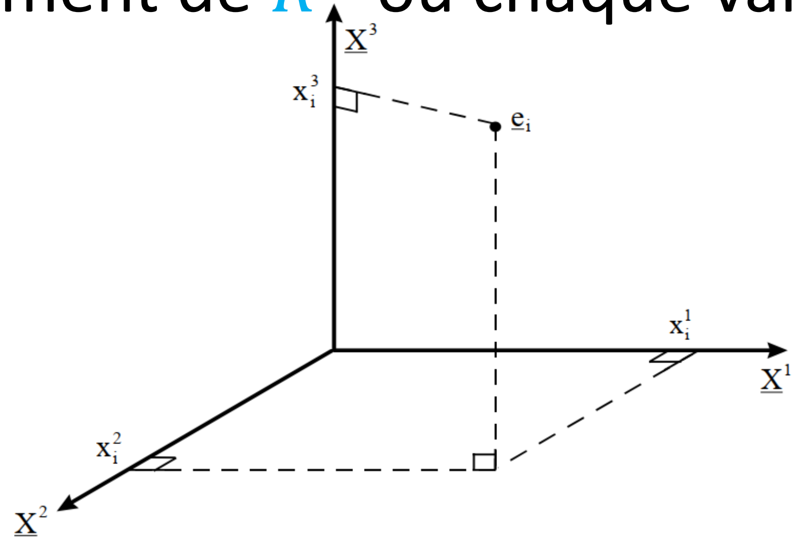
$$\text{cor}(x, y) = r_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{\text{cov(x,y)}}{\sqrt{var(x)}\,\sqrt{var(y)}}$$

Espaces des individus

- D’un point de vue géométrique, on cherche à représenter le nuage de points
- Un individu e_i est un élément de l’espace des individus R^P

$$e_i(x_i^1, x_i^2, \dots \dots x_i^P)$$

- Une variable est un élément de R^n où chaque variable du tableau est associé un axe de R^n



Espace des variables

Pour mesurer la proximité entre les variables (centrées ou centrées et réduites), on peut utiliser un produit scalaire entre les variables.

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

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

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

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

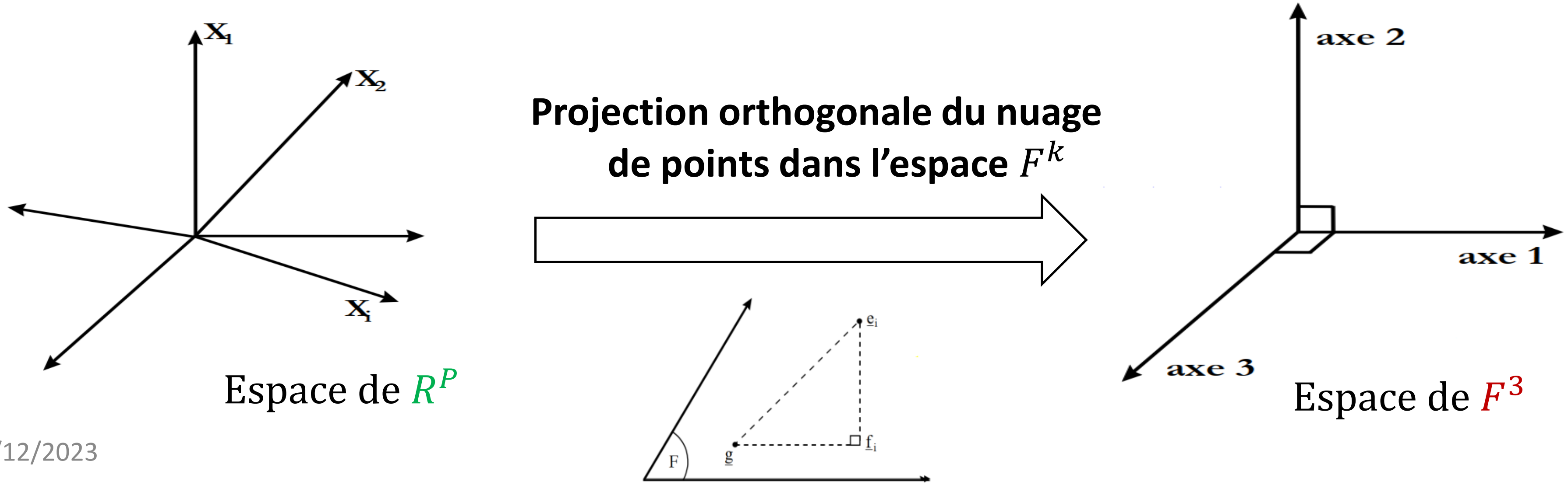
$\boxed{\left\|\underline{x}^{i}\right\| = s_{i}}$ Écart-type de $\underline{x}^{i}$

$$\boxed{\mathrm{Cos}\left(\widehat{\underline{X}^{i}, \underline{X}^{j}}\right) = \frac{\left\langle \underline{x}^{i}, \underline{x}^{j} \right\rangle}{\left\|\underline{X}^{i}\right\| \; \left\|\underline{X}^{j}\right\|} = \frac{\mathrm{Cov}\left(\underline{X}^{i}, \underline{X}^{j}\right)}{s_{i} \; s_{j}} = r\left(\underline{X}^{i}, \underline{X}^{j}\right)}$$

Analyse en composantes principales

<u>Principe</u>:

- Obtenir une représentation approchée du nuage de points (individus) dans une sous-espace F^k de dimension faible (soit k).
- Chercher à définir k nouvelle variables (combinaisons linéaires des variables initiales) avec un minimum de perte d’information possible.



Analyse en composantes principales

Une perte minimal d'information revient à:

- Minimiser la quantité entre les points e_i et leur projetés
- Ou bien maximiser inertie du nuage de points projetés sur le sous espace F^k

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

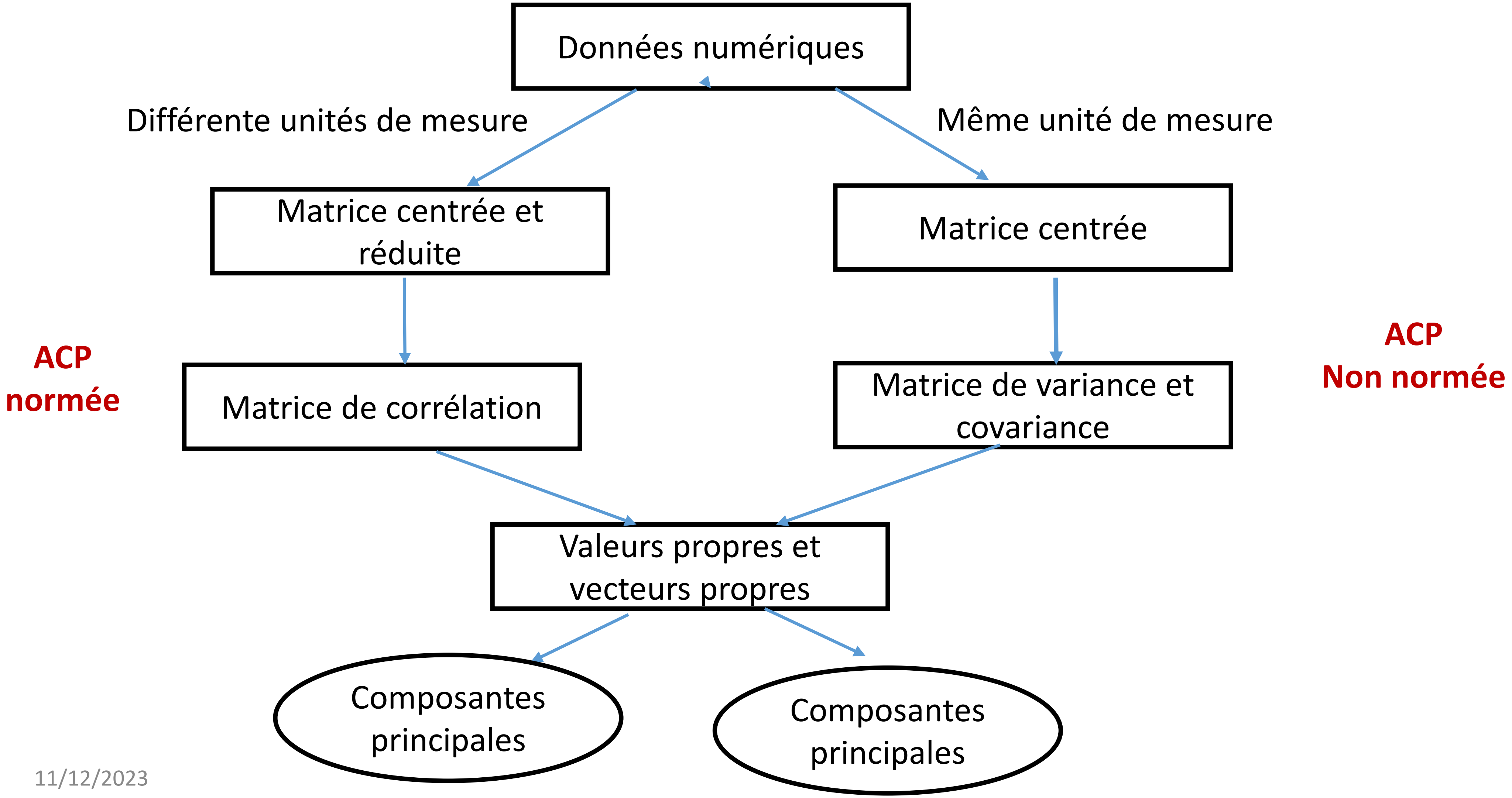
Inertie totale

Minimiser cette quantité (carrées des distances entre les points et leurs projections) $\Leftrightarrow$ **Maximiser** l'inertie du nuage projeté

f_i est une projection orthogonale de e_i sur F

L'inertie totale (I_g) mesure la dispersion du nuage de points

Analyse en composantes principales



ACP non normée

(1) Centrer le tableau X autour de leur moyenne

$$y_i^j = x_i^j - \bar{x}^j$$

En notation matricielle: $Y = X - 1_n g'$

1_n: matrice d'identité $1_n = \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{pmatrix}$

Centre de gravité g=$\begin{bmatrix} \bar{x}^1 \\ \dots \\ \bar{x}^P \end{bmatrix}$, $g' = [\bar{x}^1, \ \dots, \bar{x}^P]$

Matrice centrée $Y = \begin{bmatrix} x_1^1 - \bar{x}^1, x_1^2 - \bar{x}^2, \dots \dots x_1^P - \bar{x}^P \\ x_2^1 - \bar{x}^1, x_2^2 - \bar{x}^2, \dots \dots x_2^P - \bar{x}^P \\ \dots \dots \dots \dots \dots . . \\ x_n^1 - \bar{x}^1, x_n^2 - \bar{x}^2, \dots \dots x_n^P - \bar{x}^P \end{bmatrix}$

ACP non normée

Centrer le tableau X de notes de l’exemple précédent, avec:

$$\mathbf{g}' = (5,\ 4,\ 2,\ 4,\ 4,\ 4,\ 5)$$

$$\underset{(12,7)}{Y} = \begin{pmatrix} 7 & 4 & 2 & 2 & 7 & 1 & 6 \\ 6 & 2 & 1 & 2 & 5 & 1 & 7 \\ 4 & 3 & 2 & 2 & 3 & 4 & 4 \\ 5 & 3 & 1 & 5 & 2 & 7 & 1 \\ 4 & 5 & 2 & 7 & 1 & 4 & 2 \\ 5 & 7 & 1 & 5 & 2 & 4 & 5 \\ 4 & 2 & 1 & 3 & 5 & 3 & 6 \\ 4 & 1 & 3 & 4 & 5 & 4 & 7 \\ 6 & 6 & 2 & 4 & 7 & 5 & 5 \\ 6 & 6 & 3 & 5 & 3 & 6 & 6 \\ 7 & 7 & 5 & 7 & 7 & 6 & 7 \\ 2 & 2 & 1 & 2 & 1 & 3 & 4 \end{pmatrix} - \underbrace{\begin{pmatrix} 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 4 & 4 & 5 \\ 5 & 4 & 2 & 4 & 1 & 4 & 5 \end{pmatrix}}_{\mathbf{1_n g'}}$$

$$\underset{(12,7)}{Y} = \begin{pmatrix} 7-5 & 4-4 & 2-2 & 2-4 & 7-4 & 1-4 & 6-5 \\ 6-5 & 2-4 & 1-2 & 2-4 & 5-4 & 1-4 & 7-5 \\ 4-5 & 3-4 & 2-2 & 2-4 & 3-4 & 4-4 & 4-5 \\ 5-5 & 3-4 & 1-2 & 5-4 & 2-4 & 7-4 & 1-5 \\ 4-5 & 5-4 & 2-2 & 7-4 & 1-4 & 4-4 & 2-5 \\ 5-5 & 7-4 & 1-2 & 5-4 & 2-4 & 4-4 & 5-5 \\ 4-5 & 2-4 & 1-2 & 3-4 & 5-4 & 3-4 & 6-5 \\ 4-5 & 1-4 & 3-2 & 4-4 & 5-4 & 4-4 & 7-5 \\ 6-5 & 6-4 & 2-2 & 4-4 & 7-4 & 5-4 & 5-5 \\ 6-5 & 6-4 & 3-2 & 5-4 & 3-4 & 6-4 & 6-5 \\ 7-5 & 7-4 & 5-2 & 7-4 & 7-4 & 6-4 & 7-5 \\ 2-5 & 2-4 & 1-2 & 2-4 & 1-4 & 3-4 & 4-5 \end{pmatrix}$$

$$\underset{(12,7)}{Y} = \begin{pmatrix} 2 & 0 & 0 & -2 & 3 & -3 & 1 \\ 1 & -2 & -1 & -2 & 1 & -3 & 2 \\ -1 & -1 & 0 & -2 & -1 & 0 & -1 \\ 0 & -1 & -1 & 1 & -2 & 3 & -4 \\ -1 & 1 & 0 & 3 & -3 & 0 & -3 \\ 0 & 3 & -1 & 1 & -2 & 0 & 0 \\ -1 & -2 & -1 & -1 & 1 & -1 & 1 \\ -1 & -3 & 1 & 0 & 1 & 0 & 2 \\ 1 & 2 & 0 & 0 & 3 & 1 & 0 \\ 1 & 2 & 1 & 1 & -1 & 2 & 1 \\ 2 & 3 & 3 & 3 & 3 & 2 & 2 \\ -3 & -2 & -1 & -2 & -3 & -1 & -1 \end{pmatrix}$$

ACP non normée

(2) Calculer la matrice de variance-covariance V (matrice carrée de dimension p) d'une table centrée.

$$\sigma_j^2 = V(x^j) = \begin{bmatrix} \sigma_1^2, \sigma_{12}, \dots \dots \sigma_{1P} \\ \sigma_{21}, \sigma_2^2, \dots \dots \sigma_{2P} \\ \dots \dots \dots \dots \dots .. \\ \sigma_{P1}, \sigma_{P2}, \dots \dots \sigma_P^2 \end{bmatrix}$$

Où : $\boldsymbol{\sigma_{j\ell}}$ **=cov(** $\boldsymbol{x^j}$ **,** $\boldsymbol{x^l}$ **)** et $\boldsymbol{\sigma_j^2}$**=var(**$\boldsymbol{x^j}$)

Formule matricielle: $\sigma_j^2 = Y'D_pY$

$Y\ : matrice\ centrée\ de\ X, D_p$: matrice de poids

ACP non normée

$$\underbrace{V}_{(7,7)} = \underbrace{\begin{pmatrix} 2 & 1 & -1 & 0 & -1 & 0 & -1 & -1 & 1 & 1 & 2 & -3 \\ 0 & -2 & -1 & -1 & 1 & 3 & -2 & -3 & 2 & 2 & 3 & -2 \\ 0 & -1 & 0 & -1 & 0 & -1 & -1 & 1 & 0 & 1 & 3 & -1 \\ -2 & -2 & -2 & 1 & 3 & 1 & -1 & 0 & 0 & 1 & 3 & -2 \\ 3 & 1 & -1 & -2 & -3 & -2 & 1 & 1 & 3 & -1 & 3 & -3 \\ -3 & -3 & 0 & 3 & 0 & 0 & -1 & 0 & 1 & 2 & 2 & -1 \\ 1 & 2 & -1 & -4 & -3 & 0 & 1 & 2 & 0 & 1 & 2 & -1 \end{pmatrix}}_{Y'} \underbrace{D_p}_{(12,12)} \underbrace{\begin{pmatrix} 2 & 0 & 0 & -2 & 3 & -3 & 1 \\ 1 & -2 & -1 & -2 & 1 & -3 & 2 \\ -1 & -1 & 0 & -2 & -1 & 0 & -1 \\ 0 & -1 & -1 & 1 & -2 & 3 & -4 \\ -1 & 1 & 0 & 3 & -3 & 0 & -3 \\ 0 & 3 & -1 & 1 & -2 & 0 & 0 \\ -1 & -2 & -1 & -1 & 1 & -1 & 1 \\ -1 & -3 & 1 & 0 & 1 & 0 & 2 \\ 1 & 2 & 0 & 0 & 3 & 1 & 0 \\ 1 & 2 & 1 & 1 & -1 & 2 & 1 \\ 2 & 3 & 3 & 3 & 3 & 2 & 2 \\ -3 & -2 & -1 & -2 & -3 & -1 & -1 \end{pmatrix}}_{Y}$$

ACP non normée

(3) Diagonaliser la matrice de variances-covariance V

La diagonalisation revient à chercher les valeurs propres avec $M = I_P$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \text{ et } I_g = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \mathrm{P}$$

La métrique M est une matrice de taille $\boldsymbol{p}$ symétrique et définie positive.

$$\mathrm{M} = D_{\frac{1}{\sigma^2}} = \begin{bmatrix} 1 & & 0 \\ & & \\ & \dots & \\ 0 & & 1 \end{bmatrix}$$

M: métrique diagonale des inverses des variances

ACP non normée

Exemple: Valeurs et vecteurs propres

Vérifier que : $Av_1 = 2v_1, Av_2 = 4v_2 \; et \; Av_3 = 6v_3$

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{bmatrix}$$

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \; \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \; \mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

On dit que **v1, v2** et **v3** sont vecteurs propres de **A** associés aux valeurs propres : λ_1 = 2, λ_2= 4 et λ_3= 6.

On a : **Tr(A)**=5+4+3=12=λ_1+ λ_2 +λ_3 .

ACP non normée

Eléments de l'ACP non normée

- **Axes principaux d'inertie** a_k (axes de direction): les vecteurs propres de da la matrice **VM** normés à 1.

$$VM\, a_k = \lambda_k\, a_k$$

V est la matrice variances-covariances

M: métrique diagonale des inverses des variances

- Le premier axe : est celui qui a la plus grande valeur propre, noté a^1

- **Le deuxième axe** : est celui associé à la deuxième valeur propre , noté a^2
- .....

ACP non normée

Eléments de l'ACP non normée

- **Facteurs principaux** u_k: sont les vecteurs propres de la matrice **MV**

$$MVu_k = \lambda_k u_k$$

En pratique, on calcule les **u** par diagonalisation de **MV**

- **Composante principale** c_k: est l'axe engendré par le vecteur propre u_k et passant par l'origine

$$c_k = Y\, u_k \text{et } u_k = Ma_k$$

$$\boxed{\begin{array}{l} VAR\,(c_k) = \lambda_k \\ \sigma_{c_k} = \sqrt{VAR(C_k)} = \sqrt{\lambda_k} \end{array}}$$

Exemple: $c^1 = u_1^1\ x^1 + u_2^1\ x^2 + \ldots + u_P^1\ x^P$ et $var(c^1) = \lambda_1$

c^1 est le vecteur renfermant les cordonnées des projections des individus sur l'axe 1.

ACP non normée

Eléments de l'ACP non normée
Une composante principale contient les coordonnées des projections M-orthogonales des individus centrés sur l'axe défini par a_k

- La composante c^1 est le vecteur renfermant les cordonnées des projections des individus sur l'axe 1.

- La composante c^2 est le vecteur renfermant les cordonnées des projections des individus sur l'axe 2

Les composantes principales sont non corrélées deux à deux. En effet, les axes associés sont orthogonaux

ACP normée

(1) Centrer et réduire la table de données par Z tel que: $z_i^j = \frac{x_i^j - \bar{x}^j}{\sigma_j}$

En notation matricielle: $Z = Y\ D_{\frac{1}{\sigma}} = \begin{bmatrix} \frac{x_1^1 - \bar{x}^1}{\sigma_1}, \frac{x_1^2 - \bar{x}^2}{\sigma_2}, \dots \dots \frac{x_1^P - \bar{x}^P}{\sigma_p} \\ \frac{x_2^1 - \bar{x}^1}{\sigma_1}, \frac{x_2^2 - \bar{x}^2}{\sigma_2}, \dots \dots \frac{x_2^P - \bar{x}^P}{\sigma_p} \\ \dots \dots \dots \dots \dots . . \\ \frac{x_n^1 - \bar{x}^1}{\sigma_1}, \frac{x_n^2 - \bar{x}^2}{\sigma_2}, \dots \dots \frac{x_n^P - \bar{x}^P}{\sigma_p} \end{bmatrix}$

$D_{\frac{1}{\sigma}} = \begin{bmatrix} \frac{1}{\sigma_1} & & 0 \\ & \dots & \\ 0 & & \frac{1}{\sigma_P} \end{bmatrix}$ / $D_{\frac{1}{\sigma}}$: matrice diagonale des inverses des écart types

ACP normée

(2) Calculer la matrice de corrélation R de dimension P

$$R = \begin{bmatrix} 1, r_{12}, \dots \dots r_{1P} \\ r_{21}, 1, \dots \dots r_{2P} \\ \dots \dots \dots \dots \dots . . \\ r_{P1}, r_{P2}, \dots . 1 \end{bmatrix}$$

En notation matricielle: $R = Z' \, D_P Z$

ACP normée

(3) Diagonaliser la matrice de corrélation

La diagonalisation revient à chercher les valeurs propres

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ et $I_g = \lambda_1 + \lambda_2 + \cdots + \lambda_p$ = P

I_g : est l'inertie u nuage de points projetés

Métrique réduite (ou métrique diagonale des inverses des variances)

$$M = D_{\frac{1}{\sigma^2}} = \begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \cdots & \\ 0 & & \frac{1}{\sigma_P^2} \end{bmatrix}$$

ACP normée

Eléments de l'ACP normée

- **Axes principaux** a_k: $R\, a_k = \lambda_k\, a_k$

- **Facteurs principaux** u_k: $Ru_k = \lambda_k u_k$, $u_k = Ma_k$

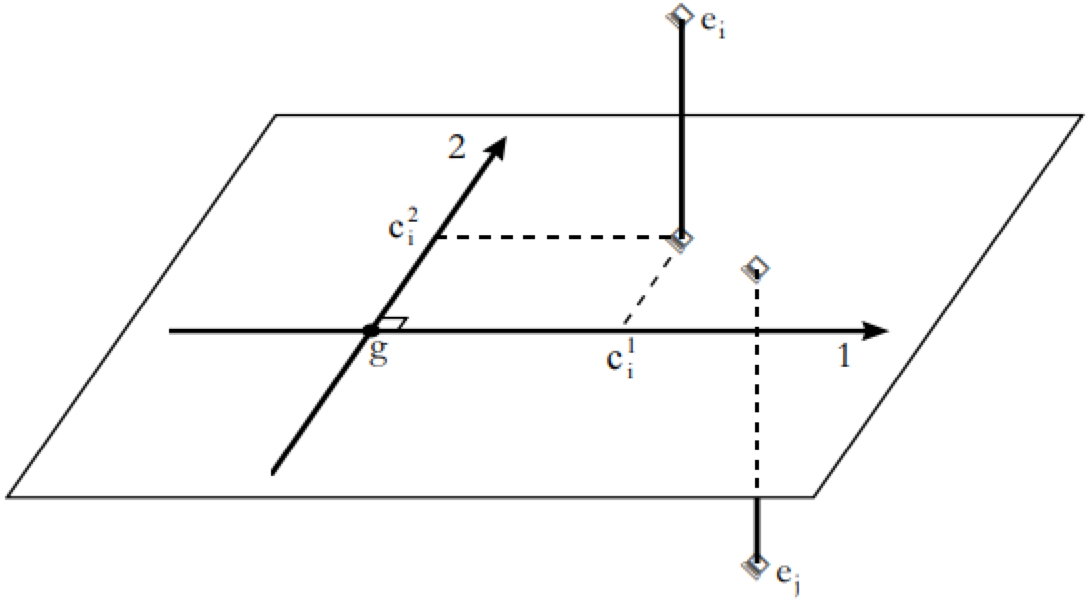
- **Composantes principales** c_k: $c_k = Z\, u_k$

$c = Z\, u$ est une combinaison linéaire des centrés réduites ayant une variance maximale

Exemple: $c^1 = u_1^1\, Z^1 + u_2^1 Z^2$+…+$u_P^1\, Z^P$ et $var(c^1) = \lambda_1$

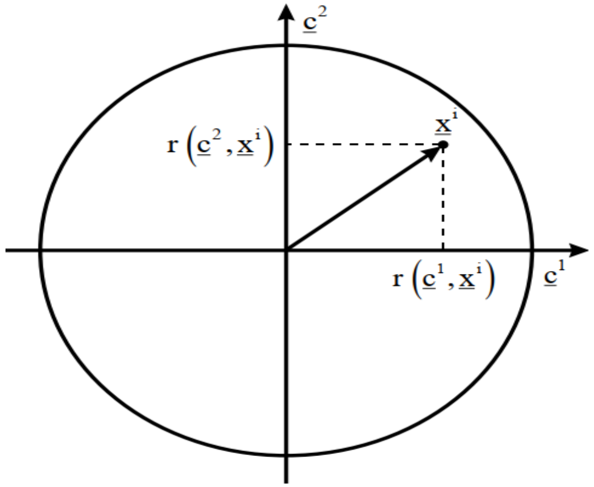
Représentation des individus

- Dans un plan principal à deux composantes principales c_1 et c_2, on représente chaque individu e_i par un point d'abscisse c_i^1 et d'ordonnée c_i^2.



Représentation des variables

Représenter la projection du nuage des variables sur le plan des composantes principales par un cercle des corrélations.



$r(c^i, x^i)$ est le **coefficient de corrélation linéaire** entre la composante principale c^i et la variable initiale x^i

Ce coefficient définit le cosinus de l'angle formé par les vecteurs correspondants des variables.

Qualité de représentation sur les plans principaux

Critère du pourcentage d'inertie totale expliquée (souvent exprimé par un pourcentage) permet de déterminer nombre d'axes retenus.

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{I_g} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \cdots + \lambda_p}$$

$\frac{\lambda_i}{I_g} =$ Mesure la part d'inertie expliquée par l'axe i

$\frac{\lambda_1+\lambda_2}{I_g} =$ La part d'inertie expliquée par le **premier plan** principale

Qualité de représentation sur les plans principaux

Contribution apportée par les individus : permet de déterminer

la contribution par les divers individus pour chaque axe.

Nous considérons la composante principale c^k, soit c_i^k la valeur de la composante pour l'individu i.

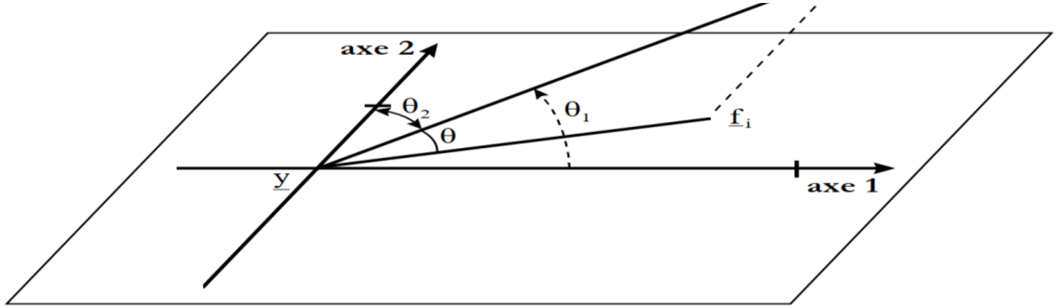
La contribution de l'individu i à c^k est :

$$ctr(i) = \frac{P_i\,(c_i^k)^2}{\lambda_k} \,/\, \sum_{i=1}^{n} ctr(i) = 1$$

Qualité de représentation

Cosinus carrés: définit par le carré du cosinus de l'angle entre l'axe de projection et le vecteur pour chaque individu.

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



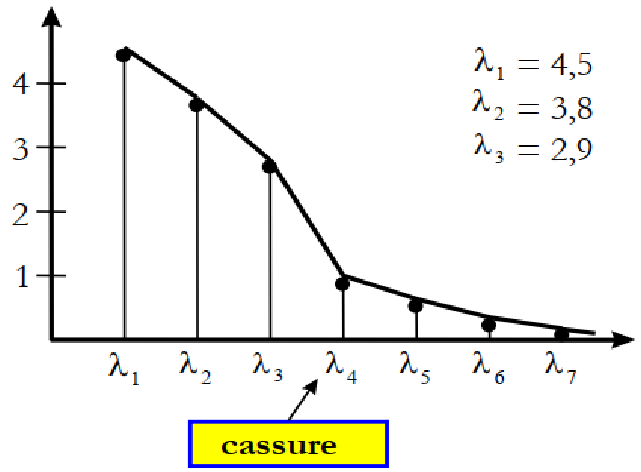
Plus la valeur est proche de 1, meilleure est la qualité de représentation des individus.

Choix de la dimension de l'espace des individus

Critère de Kaiser

Retenir que les axes dont l'inertie I_p est supérieure à l'inertie moyenne $\frac{I_p}{P}$

Dans le cas d'une ACP normée, on ne retiendra que les axes associés à des valeurs propres supérieures à 1.



Interprétation par le cercle de corrélation

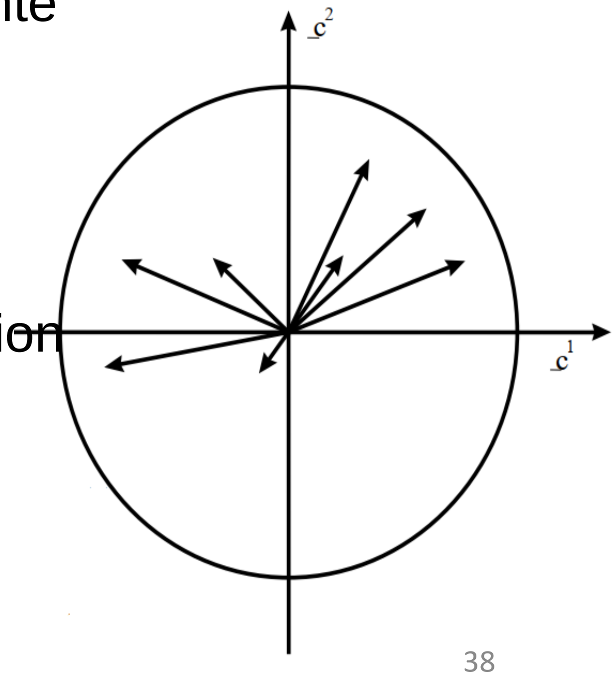
(1) Les positions des variables les uns par rapport à l'origine: les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représenté

(2) Les positions des variables les uns par rapport aux autres:

- Deux variables qui sont proches ou confondus par rapport une composante sont corrélées positivement (coefficient de corrélation proche de 1),

-Deux variables opposées (formant un angle de π) sont corrélées négativement par rapport une composante(coefficient de corrélation proche de -1),

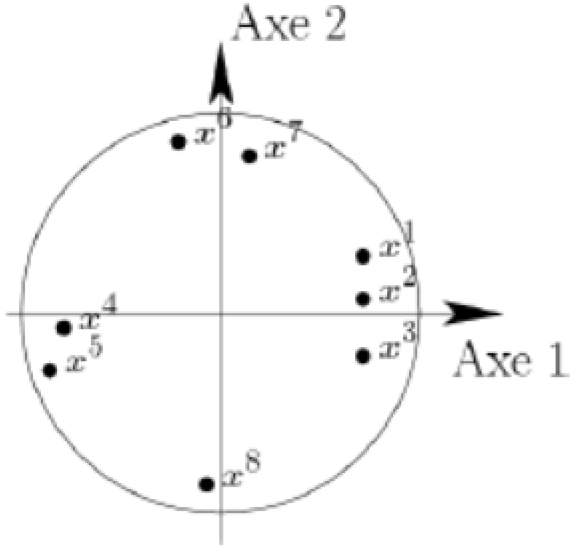
- Deux variables positionnées (formant un angle de $\pi/2$) ne sont pas corrélées (coefficient de corrélation égale à 0).



Interprétation par le cercle de corrélation

x1; x2; x3 sont corrélées positivement avec C1 ,

x4; x5 sont anticorrélées (corrélés négativement) de cet axe et x6; x7; x8 sont non corrélées avec C1.



Résumé

- L'ACP est une méthode puissante pour synthétiser et résumer de vastes populations décrites par plusieurs variables quantitatives.

- Elle permet d’étudier à la fois les variables (corrélations) et les individus (ressemblance).

- L'ACP peut être une première analyse pour l'étude d'une population dont les résultats seront enrichis par une autre analyse factorielle ou encore une classification automatique des données.