Analyse des données
Filière: IAGI

Année universitaire: 2021-2022

AZMI Mohamed

Mohamed.azmi@ensam-casa.ma

Introduction:

Qu'est ce que c'est ADD?

- Un ensemble de **méthodes statistiques** dont les caractéristiques principales doivent être **multidimensionnelles** et **descriptives**
- Le terme multidimensionnel couvre deux aspects majeurs:
    - **Les observations** (ou en d'autre terme les individus) sont décrites par plusieurs variables.
    - l'étude de ces variables se fait d'une manière simultanée (**approche globale**)
- L'intérêt de **l'étude globale** des variables réside dans le fait que ces **variables sont liées**.
- L'étude des liens entre les variables deux par deux ne constitue pas une approche multidimensionnelle dans le vrais sens du terme si ces liens ne sont pas étudiés **simultanément**.
- Nous faisons souvent appel à ces méthodes à chaque fois la notion de profil est pertinente dans l'étude des observations (individus), par exemple, les profils des consommateurs, les profils biométriques, les profils des entreprises, et ainsi de suite.

Chapitre I

Analyse en Composantes Principales

Introduction:

Données-Notation-Exemples

- L'analyse des données s'applique aux tables de données ou:
  - Les lignes représentent les individus: $Id_i=x_{i1},x_{i2},x_{i3},...,x_{iK}$. Avec i=1,2,...,I
  - Les colonnes représentent des variables quantitatives: $X_k=x_{1k},x_{2k},x_{3k},...,x_{Ik}$. Avec k=1,2,...,K.
- x_{ik} est la valeur prise par l'individus i pour la variable X_k

Id	X_1	X_2	X_3	...	X_K
Id_1	x_{11}	x_{12}	x_{13}	...	x_{1K}
Id_2	x_{21}	x_{22}	x_{23}	...	x_{2K}
Id_3	x_{31}	x_{32}	x_{33}	...	x_{3K}
Id_4	x_{41}	x_{42}	x_{43}	...	x_{4K}
Id_5	x_{51}	x_{52}	x_{53}	...	x_{5K}
...	...	...	...	...	...
...	...	...	...	...	...
Id_I	x_{I1}	x_{I2}	x_{I3}	...	x_{IK}

$$\bar{x_k} = \frac{1}{I}\sum_{i=1}^{I} x_{ik}$$

$$s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik}-\bar{x_k})^2}$$

$$Cov_{jk} = \frac{1}{I}\sum_{i=1}^{I}(x_{ij}-\bar{x_j})(x_{ik}-\bar{x_k})$$

Introduction:

Données-Notation-Exemples

Temperatures moyennes relevées dans 35 grandes villes Européennes.

	Janvier <dbl>	Février <dbl>	Mars <dbl>	Avril <dbl>	Mai <dbl>	Juin <dbl>	Juillet <dbl>	Août <dbl>	Septembre <dbl>
Amsterdam	2.9	2.5	5.7	8.2	12.5	14.8	17.1	17.1	14.5
Athènes	9.1	9.7	11.7	15.4	20.1	24.5	27.4	27.2	23.8
Berlin	-0.2	0.1	4.4	8.2	13.8	16.0	18.3	18.0	14.4
Bruxelles	3.3	3.3	6.7	8.9	12.8	15.6	17.8	17.8	15.0
Budapest	-1.1	0.8	5.5	11.6	17.0	20.2	22.0	21.3	16.9
Copenhague	-0.4	-0.4	1.3	5.8	11.1	15.4	17.1	16.6	13.3

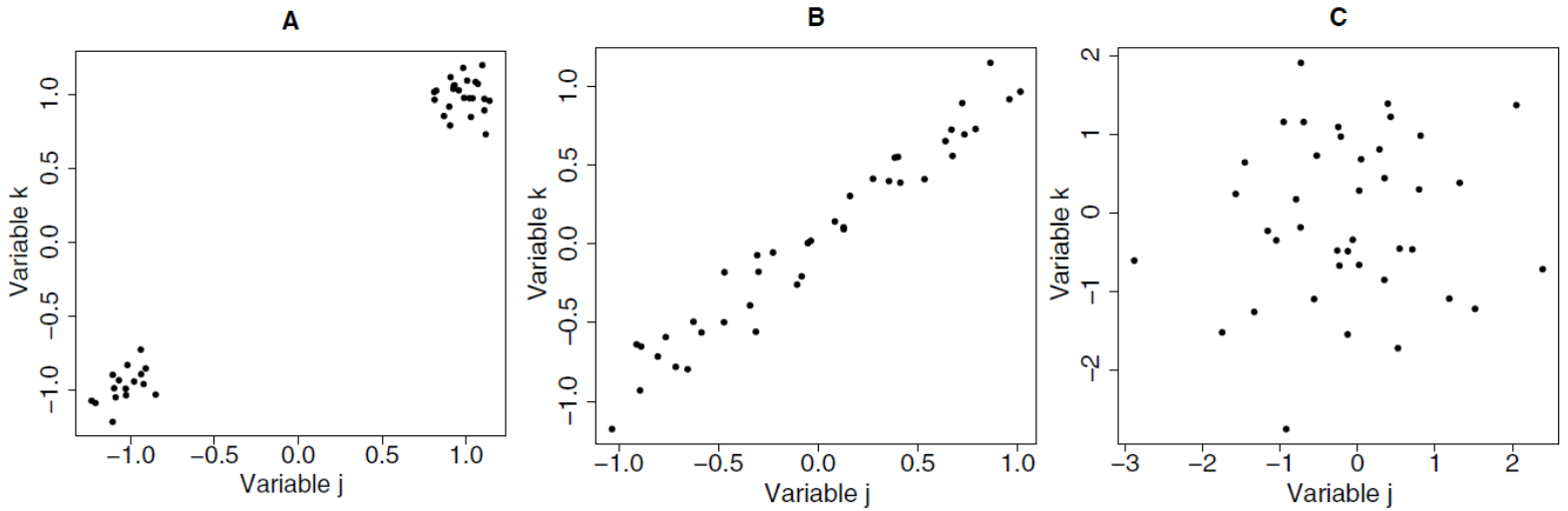
6 rows | 1-10 of 18 columns

Jus d'orange – évaluation par experts.

Jus	intensite odeur	typicite odeur	Pulpe	intensite gout	acidite	amertume	douceur
Pampryl amb .	2.82	2.53	1.66	3.46	3.15	2.97	2.60
Tropicana amb.	2.76	2.82	1.91	3.23	2.55	2.08	3.32
Fruvita fr.	2.83	2.88	4.00	3.45	2.42	1.76	3.38
Joker amb.	2.76	2.59	1.66	3.37	3.05	2.56	2.80
Tropicana fr.	3.20	3.02	3.69	3.12	2.33	1.97	3.34
Pampryl fr.	3.07	2.73	3.34	3.54	3.31	2.63	2.90

Introduction:

Données-Etude des individus



deux classes disjointes

Tendance linéaire

Distribution uniforme

- L'étude des individus implique l'identification des **similarités** entre eux (profiles/typologies),
- Peut-on former des groupes d'individus proches les uns des autres et qui seraient éloignés des autres individus ? **Quelles sont les variables qui expliquent le plus la variabilité inter-individus** ?
- De plus, ça peut nécessiter un passage par **identification des dimensions de variabilité qui met en lumière les groupement d'individus.**

Introduction:

Données-Etude des variables

Quelles sont les variables qui expliquent le plus ou le moins la variabilité inter-individus ?

Remarques sur les graphes:

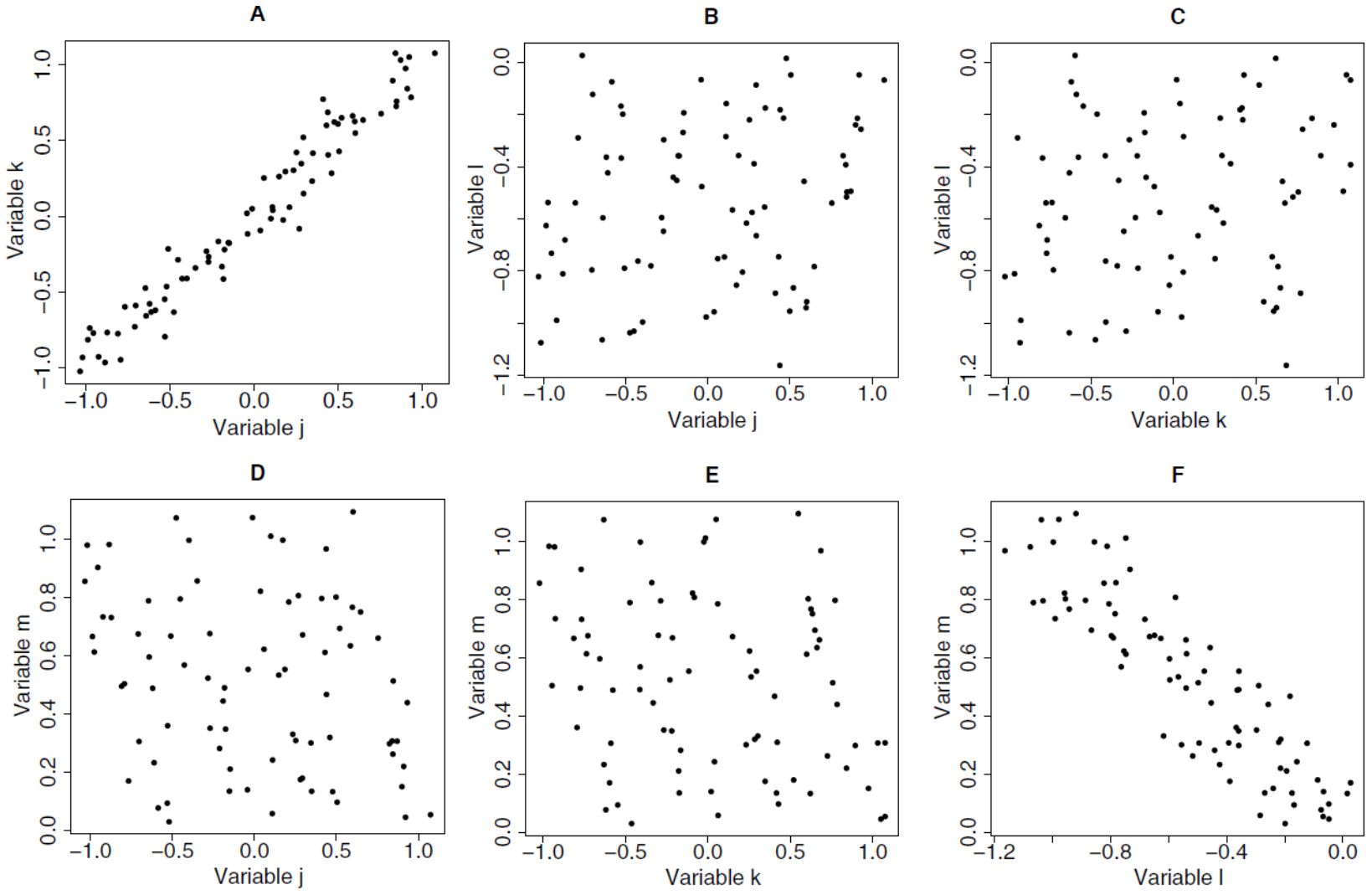
- Forte corrélation positive entre j et k (A)
- Forte corrélation négative entre l et m (F)
- absence de signe de relation entre les autres variable.

⇒ les quatre variables peuvent être regrouper en deux ensembles composés de deux variables chacun; (j, k) et (l, m)

⇒ 2 Variables synthétiques

⇒ Fastidieux dans le cas de beaucoup de variables $(\frac{K^2-k}{2})$

⇒l'Analyse en Composantes Principales (ACP)



Introduction:

Données-Individus & variables

- L'étude des individus et l'étude des variables sont **interdépendantes** puisqu'elles sont réalisées sur une même table de données, les étudier conjointement ne peut que **renforcer leurs interprétations** respectives.

- Si l'étude des individus a conduit à distinguer des **groupes d'individus**, il semble plus pertinent de les **caractériser directement par les variables** en jeu.

- De même, lorsqu'il y a des **groupes de variables**, il peut être difficile d'interpréter les relations entre elles. Dans ce cas on peut utiliser des individus spécifiques, c'est à dire des **individus extrêmes** du point de vue de ces relations.

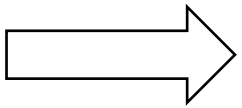
Introduction:

Données -Variables centrées réduites

Pour des raisons mathématiques de simplification, mais aussi parce que les variables dans ces tableaux peuvent être de natures différentes, on transforme la matrice X en une matrice Z de variables centrées réduites qui conserve la même structure de l'information :

$$X_k \text{ devient } Z_k = \frac{X_k - \bar{x_k}}{s_k}$$

Id	X_1	X_2	X_3	$\dots$	X_K
Id_1	x_{11}	x_{12}	x_{13}	$\dots$	x_{1K}
Id_2	x_{21}	x_{22}	x_{23}	$\dots$	x_{2K}
Id_3	x_{31}	x_{32}	x_{33}	$\dots$	x_{3K}
Id_4	x_{41}	x_{42}	x_{43}	$\dots$	x_{4K}
Id_5	x_{51}	x_{52}	x_{53}	$\dots$	x_{5K}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
Id_I	x_{I1}	x_{I2}	x_{I3}	$\dots$	x_{IK}



Id	$\frac{X_1-\bar{x_1}}{s_1}$	$\frac{X_2-\bar{x_2}}{s_2}$	$\frac{X_3-\bar{x_3}}{s_3}$	$\dots$	$\frac{X_k-\bar{x_k}}{s_k}$
Id_1	$\frac{X_1-\bar{x_1}}{s_1}$	$\frac{X_{12}-\bar{x_2}}{s_2}$	$\dots$	$\dots$	$\frac{X_{1k}-\bar{x_k}}{s_k}$
Id_2	$\frac{X_{21}-\bar{x_1}}{s_1}$	$\frac{X_{22}-\bar{x_2}}{s_2}$	$\dots$	$\dots$	$\frac{X_{2k}-\bar{x_k}}{s_k}$
Id_3	$\frac{X_{31}-\bar{x_1}}{s_1}$	$\frac{X_{32}-\bar{x_2}}{s_2}$	$\dots$	$\dots$	$\frac{X_{3k}-\bar{x_k}}{s_k}$
Id_4	$\frac{X_{41}-\bar{x_1}}{s_1}$	$\frac{X_{42}-\bar{x_2}}{s_2}$	$\dots$	$\dots$	$\frac{X_{4k}-\bar{x_k}}{s_k}$
Id_5	$\frac{X_{51}-\bar{x_1}}{s_1}$	$\frac{X_{52}-\bar{x_2}}{s_2}$	$\dots$	$\dots$	$\frac{X_{5k}-\bar{x_k}}{s_k}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
Id_I	$\frac{X_{I1}-\bar{x_1}}{s_1}$	$\frac{X_{I2}-\bar{x_2}}{s_2}$	$\dots$	$\dots$	$\frac{X_{Ik}-\bar{x_k}}{s_k}$

$$\bar{x_k} = \frac{1}{I}\sum_{i=1}^{I} x_{ik}$$

$$s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x_k})^2}$$

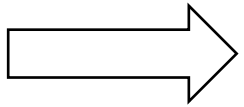
$$Cov_{jk} = \frac{1}{I}\sum_{i=1}^{I}(x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})$$

Introduction:

Données -Variables centrées réduites

- Le centrage n'a pas d'influence sur la **ressemblance entre individus**
- La réduction supprime **l'arbitrage des unités** et toutes les variables ont la même influence dans le calcul des distances entre individus
- Particularités de ces nouvelles variables :
  - les moyennes sont toutes nulles
  - les écart types sont égaux à 1

Id	X_1	X_2	X_3	…	X_K
Id_1	x_{11}	x_{12}	x_{13}	…	x_{1K}
Id_2	x_{21}	x_{22}	x_{23}	…	x_{2K}
Id_3	x_{31}	x_{32}	x_{33}	…	x_{3K}
Id_4	x_{41}	x_{42}	x_{43}	…	x_{4K}
Id_5	x_{51}	x_{52}	x_{53}	…	x_{5K}
…	…	…	…	…	…
…	…	…	…	…	…
Id_I	x_{I1}	x_{I2}	x_{I3}	…	x_{IK}



Id	$\frac{X_1-\bar{x_1}}{s_1}$	$\frac{X_2-\bar{x_2}}{s_2}$	$\frac{X_3-\bar{x_3}}{s_3}$	…	$\frac{X_k-\bar{x_k}}{s_k}$
Id_1	$\frac{X_1-\bar{x_1}}{s_1}$	$\frac{X_{12}-\bar{x_2}}{s_2}$	…	…	$\frac{X_{1k}-\bar{x_k}}{s_k}$
Id_2	$\frac{X_{21}-\bar{x_1}}{s_1}$	$\frac{X_{22}-\bar{x_2}}{s_2}$	…	…	$\frac{X_{2k}-\bar{x_k}}{s_k}$
Id_3	$\frac{X_{31}-\bar{x_1}}{s_1}$	$\frac{X_{32}-\bar{x_2}}{s_2}$	…	…	$\frac{X_{3k}-\bar{x_k}}{s_k}$
Id_4	$\frac{X_{41}-\bar{x_1}}{s_1}$	$\frac{X_{42}-\bar{x_2}}{s_2}$	…	…	$\frac{X_{4k}-\bar{x_k}}{s_k}$
Id_5	$\frac{X_{51}-\bar{x_1}}{s_1}$	$\frac{X_{52}-\bar{x_2}}{s_2}$	…	…	$\frac{X_{5k}-\bar{x_k}}{s_k}$
…	…	…	…	…	…
…	…	…	…	…	…
Id_I	$\frac{X_{I1}-\bar{x_1}}{s_1}$	$\frac{X_{I2}-\bar{x_2}}{s_2}$	…	…	$\frac{X_{Ik}-\bar{x_k}}{s_k}$

Introduction:

Données -Variables centrées réduites

	Janvier <dbl>	Février <dbl>	Mars <dbl>	Avril <dbl>	Mai <dbl>	Juin <dbl>	Juillet <dbl>	Août <dbl>	Septembre <dbl>
Amsterdam	2.9	2.5	5.7	8.2	12.5	14.8	17.1	17.1	14.5
Athènes	9.1	9.7	11.7	15.4	20.1	24.5	27.4	27.2	23.8
Berlin	-0.2	0.1	4.4	8.2	13.8	16.0	18.3	18.0	14.4
Bruxelles	3.3	3.3	6.7	8.9	12.8	15.6	17.8	17.8	15.0
Budapest	-1.1	0.8	5.5	11.6	17.0	20.2	22.0	21.3	16.9
Copenhague	-0.4	-0.4	1.3	5.8	11.1	15.4	17.1	16.6	13.3

6 rows | 1-10 of 18 columns

Les moyennes

Janvier	Février	Mars	Avril	Mai
1.3	2.2	5.2	9.3	13.9

Juin	Juillet	Août	Septembre
17.4	19.6	19.0	15.6

Variables centrées réduites

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août
Amsterdam	0.28	0.051	0.097	-0.28	-0.431	-0.79	-0.71	-0.50
Athènes	1.41	1.361	1.331	1.61	1.890	2.13	2.18	2.20
Berlin	-0.28	-0.385	-0.170	-0.28	-0.034	-0.43	-0.37	-0.26
Bruxelles	0.36	0.197	0.303	-0.10	-0.340	-0.55	-0.51	-0.32
Budapest	-0.44	-0.258	0.056	0.61	0.943	0.84	0.66	0.62
Copenhague	-0.32	-0.476	-0.808	-0.91	-0.859	-0.61	-0.71	-0.64

Valeurs positives => supérieures aux moyennes

Valeurs proches de 0 => proches des moyennes

Valeurs négatives=> inférieures aux moyennes

Introduction:

Données-Nuage variables

- Une variable est assimilée à un vecteur par le n-uplet $(X_j=x_{1j},x_{2j},x_{3j},\ldots,x_{Nj})$
- Le nuage des variables peut donc être considéré comme un ensemble de p vecteurs représentés dans un espace de dimension N dont on cherche à étudier les corrélations.
- Rappels : produit scalaire dans $\mathbb{R}^3$:

$$\langle \vec{u}, \vec{v} \rangle = ||\vec{u}|| \times ||\vec{v}|| \cos(\vec{u}, \vec{v})$$

⇒**Deux vecteurs formant un angle aigu donneront un produit scalaire positif alors que pour deux vecteurs formant un angle obtus, le produit scalaire sera négatif. Pour l'ongle droit il sera nul.**

- Autre expression :

$$\langle \vec{u}, \vec{v} \rangle = u_1 v_1 + u_2 v_2 + u_3 v_3$$

- Si les vecteurs sont de norme 1:

$$\langle \vec{u}, \vec{v} \rangle = \cos(\vec{u}, \vec{v}) = \sum_{i=1}^{3} u_i v_i$$

Introduction:

Données-Nuage variables

le coefficient de corrélation linéaire :

Soient j et j' deux variables :

$$r_{j,j'} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)\left(\frac{x_{ij'} - \bar{x}_{j'}}{s_{j'}}\right)$$

On voit par ailleurs, qu'au coefficient 1/n près, $r_{j,j'}$ correspond au produit scalaire entre deux vecteurs colonnes centrées reduites.

Déduire que $\cos(j,j') = r_{j,j'}$

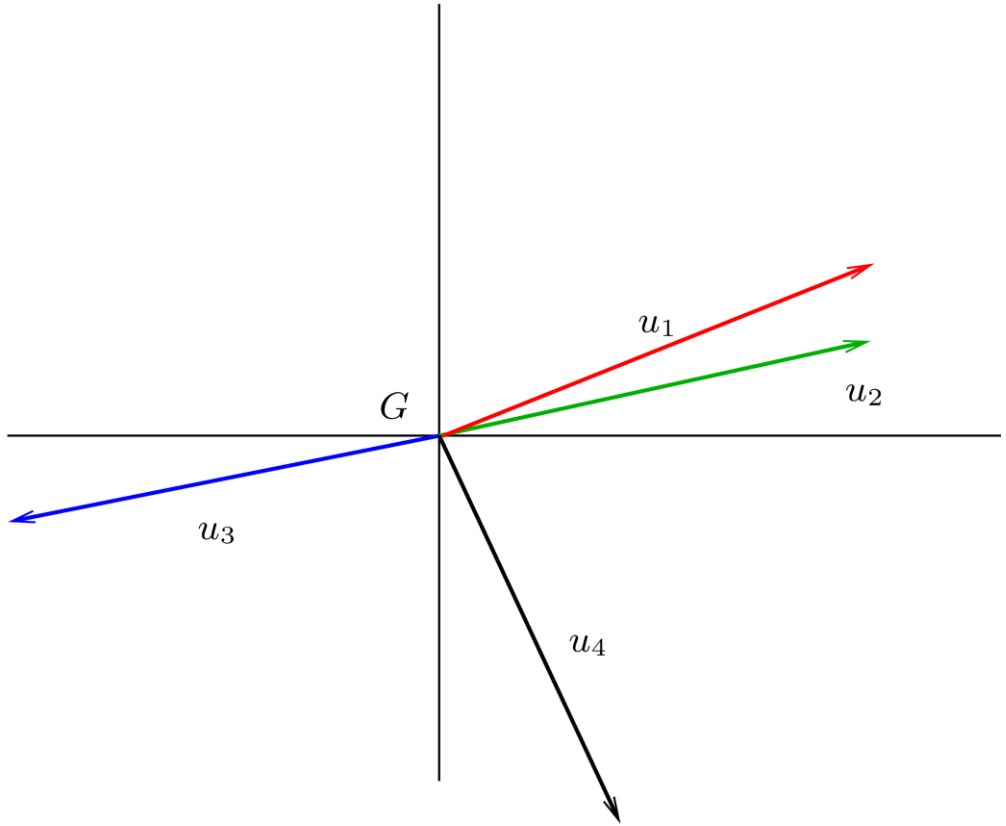
Introduction:

Données-Nuage variables

le coef de corrélation linéaire : Interprétation

- Deux variables **fortement corrélées** pourront être représentées par des vecteurs presque colinéaires et de même sens comme les vecteurs u1 et u2. L'angle entre les deux vecteurs étant de mesure presque nulle, le cosinus vaut presque 1.
- Si deux variables sont **corrélées négativement**, ça correspondrait à un angle presque plat : cos(j; j') ~-1. C'est le cas pour u1 et u3 ou u2 et u3.
- Lorsque les vecteurs sont **presque orthogonaux**, la connaissance des coordonnées d'un vecteur ne donne pas d'information particulière sur les coordonnées de l'autre : c'est le cas entre u1 et u4 par exemple ou :

$$\cos(j, j') = r_{j,j'} \sim 0.$$



Introduction:

Données-Nuage variables

L'inertie : l'information à expliquer ou l'information portée par les données.

$$I = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2$$

Interprétation en lien avec le nuage des individus

$$I = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2$$

l'inertie peut être vue comme la somme (au coefficient 1/n près) des carrés des distances au centre de gravite pour tous les individus.
En cela, **l'inertie renseigne sur la « forme» du nuage des individus.**

Interprétation en lien avec le nuage des variables

$$I = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2$$

$$= \frac{1}{n} \sum_{j=1}^{p} n$$

$$= np/n$$

$$= p.$$

L'inertie (pour une ACP normée) est donc toujours **égale au nombre de variables.**

Introduction:

Données-Nuage variables

L'inertie : l'information à expliquer ou l'information portée par les données.

$$I = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2$$

Interprétation en lien avec le nuage des individus

l'inertie peut être vue comme la somme (au coefficient 1/n pres) des carres des distances au centre de gravite pour tous les individus.
En cela, l'inertie renseigne sur la « forme» du nuage des individus.

Interprétation en lien avec le nuage des variables

L'inertie (pour une ACP normée) est donc toujours égale au nombre de variables.

L'ACP consiste en fait en une décomposition de cette inertie dans des directions privilégiées des espaces propres aux représentations des individus et des variables.

Introduction:

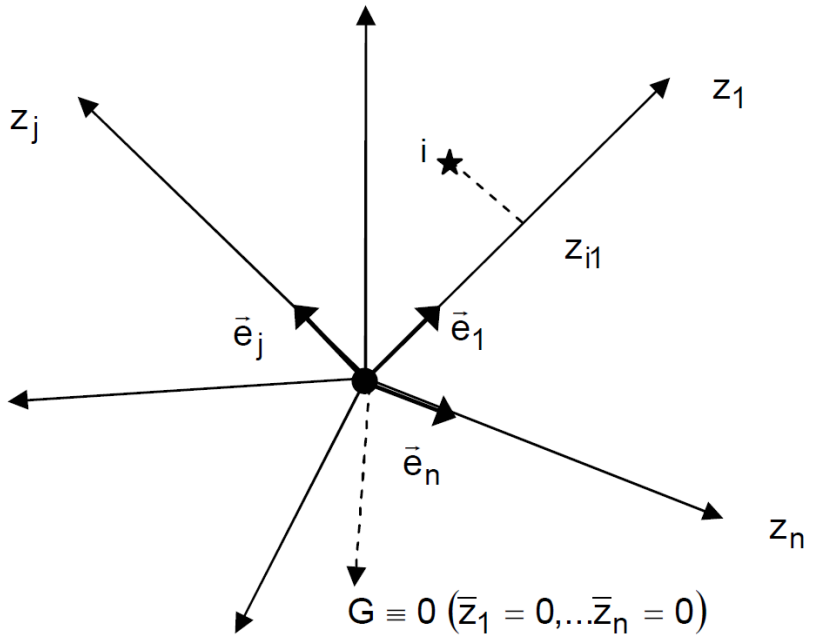
Données-Notation-Exemples

- La distance entre deux points est calculée par la distance euclidienne (théorème de Pythagores):

$$d^2(i,i') = \sum_{j=1}^{n} (z_{ij} - z_{i'j})^2$$

- Les projections orthogonales (les coordonnées) des N points sur un axe centré réduit z_j sont de moyennes nulles et de variance égale à un => Le centre de gravité G est donc l'origine des axes.

- La variance totale du nuage multidimensionnel dans un espace centré réduit est égale au nombre de variable **p** => Chaque axe porte donc 1/p* 100 de la variance total.



L'information contenue dans ces espaces est illisible du fait du nombre d'axes.

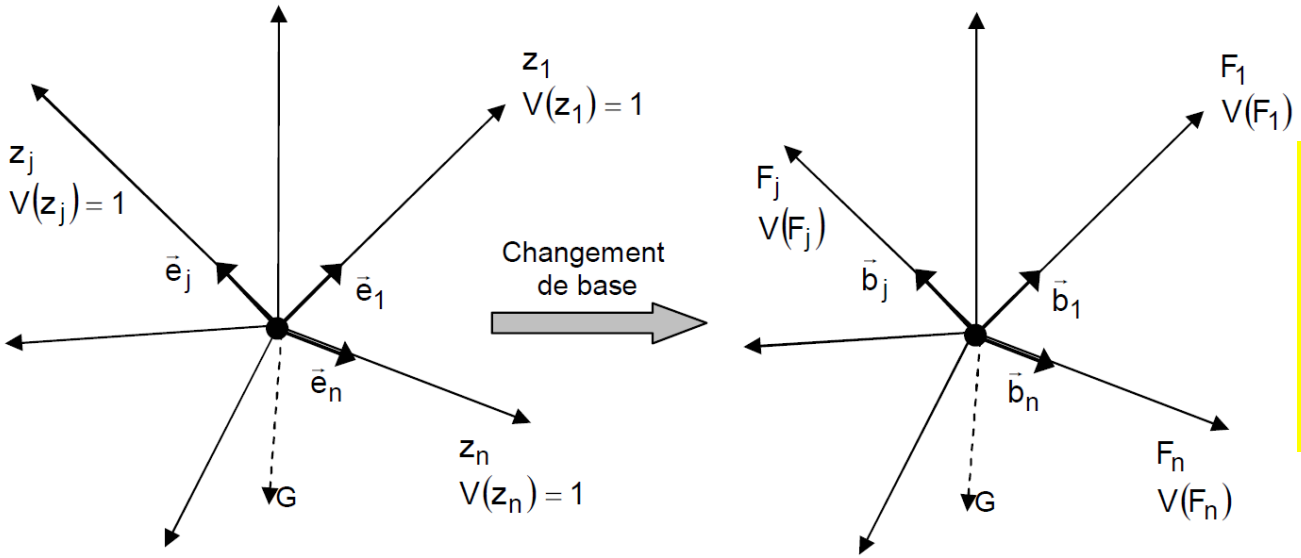
Analyse en Composantes Principales : Principe général

- L'ACP a pour but de substituer à ces espaces, **des espaces de même dimension mais de tel sorte qu'une grande part de l'information soit lisible à partir d'un nombre réduit d'axes (idéalement 2 ou 3).**
- Le principe de l'ACP consiste donc à effectuer un changement de base de tel sorte (lorsque cela est possible) que **les variances des projections orthogonales (les coordonnées) sur les nouveaux axes (appelés axes principaux) rassemblent une part significative de la variance totale à partir des deux ou trois premiers axes.**
- On peut schématiser ce principe de la façon suivante.

$Var(Z_1) = Var(Z_1) = \ldots = 1$

Variables de même importance

Difficilement analysé



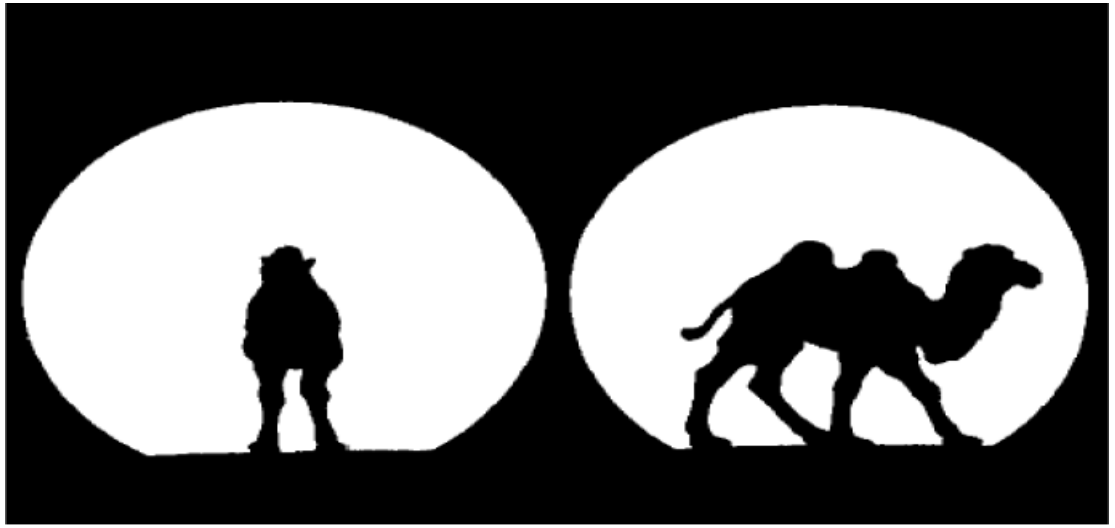
$Var(F_1) > Var(F_2) > \ldots > Var(F_n)$

Analysable en grande partie (70 à 80%)à partir des premiers facteurs (2 ou 3)

Analyse en Composantes Principales : Principe général

Les propriétés géométriques des nuages doivent répondre aux questions posées :

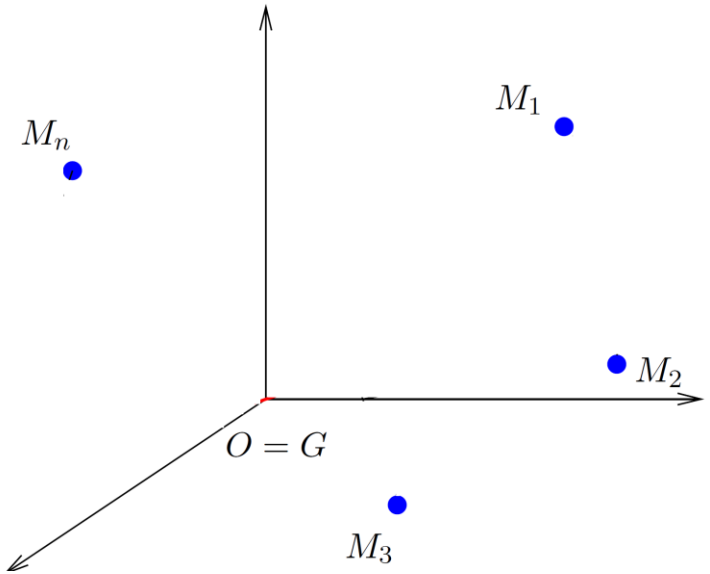
- variabilité des individus via les distances inter-individus
- liaisons entre variables via les angles inter-variables .



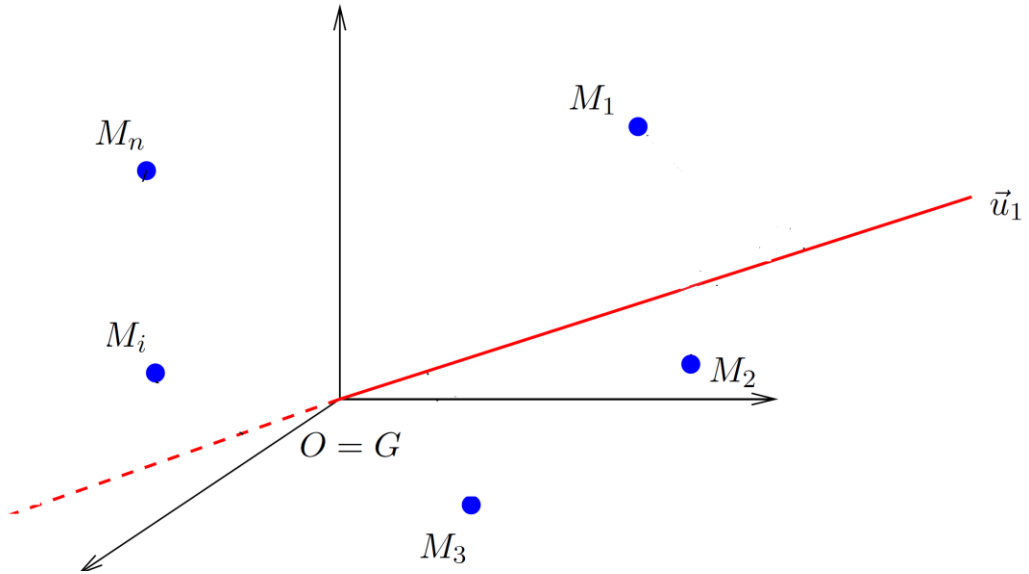
Quelle représentation choisir pour le chameau ?

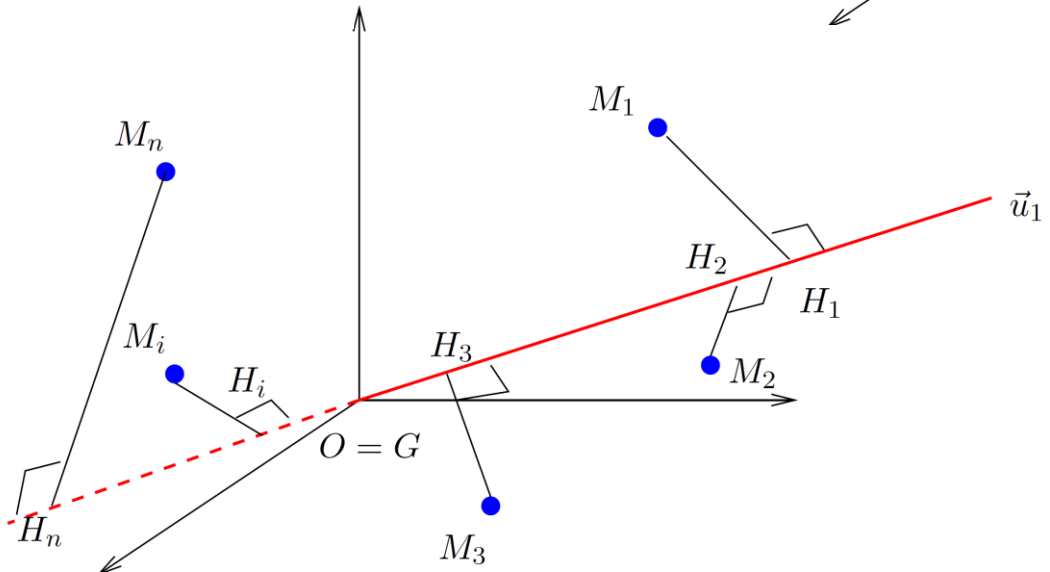
Réduire les dimensions pour obtenir une représentation plus simple du nuage des points tout en conservant le plus possible de variabilité est le principe appliqué en ACP.

Analyse en Composantes Principales : Principe général



Recherche du meilleur axe de projection u





maximiser

$$\sum_{i=1}^{n} OH_i^2$$

Analyse en Composantes Principales :
Meilleur plan de projection

Meilleur axe de projection :

On cherche un espace P tel que $\sum_{i=1}^{n} OH_i^2$ soit maximum, les H_i désignant les projetées orthogonaux de tous les individus M_i sur P.

Meilleure representation plane P :
On construit ainsi de manière itérative une suite d'axes de directions $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$, … $\vec{u}_p$ telle que:

- $\vec{u}_1$ donne la direction qui maximise l'inertie projetée.
- $\vec{u}_2$ donne la direction du reste de l'espace qui maximise l'inertie projetée.
- …

A l'issue de cette opération, on dispose donc de p vecteurs orthogonaux deux à deux qui permettent donc de reconstituer l'espace des individus.

Analyse en Composantes Principales : Formulation mathématique

Formulation mathématique de l'ACP:

1. Considérer le tableau D(N,P) de données à N lignes et P colonnes
2. Transformer la matrice D(N,P) en une matrice Z(N,P) centrées réduites
3. Calculer la matrice R des coefficients de corrélation linéaire entre les variables.
4. Calculer les P valeurs propres $(\lambda_1,...,\lambda_P)$ de R et les vecteurs propres correspondant $V=(v_1,...,v_P)$.
5. Calculer les Composantes principales C=Z×V qui sont les projections orthogonales du nuage des points individus sur les nouveaux axes $C=(C_1,...,C_P)$. Elles sont donc centrées.(Le pourcentage de variance expliquée par une composante principale C_s est donné par la quantité $\frac{\lambda_s}{P} \times 100$)
6. Sélectionner le nombre de composantes principales qui assure un pourcentage de variance expliquée satisfaisant.

Analyse en Composantes Principales :

Exercices

Exercice 1 : Considérons la matrice suivante $X = \begin{pmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}$

1- Calculer le produit $X^T X$ est s'assurer que c'est une matrice carrée symétrique

2- Calculer les valeurs propres de $X^T X$ ainsi que les vecteurs propres correspondants

Exercice 2 : Considérons la matrice suivante X= $\begin{pmatrix} 4 & 5 \\ 6 & 7 \\ 8 & 0 \end{pmatrix}$

1- Centrer et normer la matrice X

2- Calculer la matrice variances-covariances et la matrice des corrélations relatives à la matrice centrée réduite

3- Calculer les vecteurs constituant la base du meilleur plan de projection

4- calculer les composantes principales

Analyse en Composantes Principales :

Exercices

Exercice 3 :

réaliser l'ACP de la matrice de données suivante

$$X=\begin{pmatrix} \boldsymbol{x_1} & \boldsymbol{x_2} \\ 2 & 2 \\ 6 & 2 \\ 6 & 4 \\ 10 & 4 \end{pmatrix}$$

Analyse en Composantes Principales :

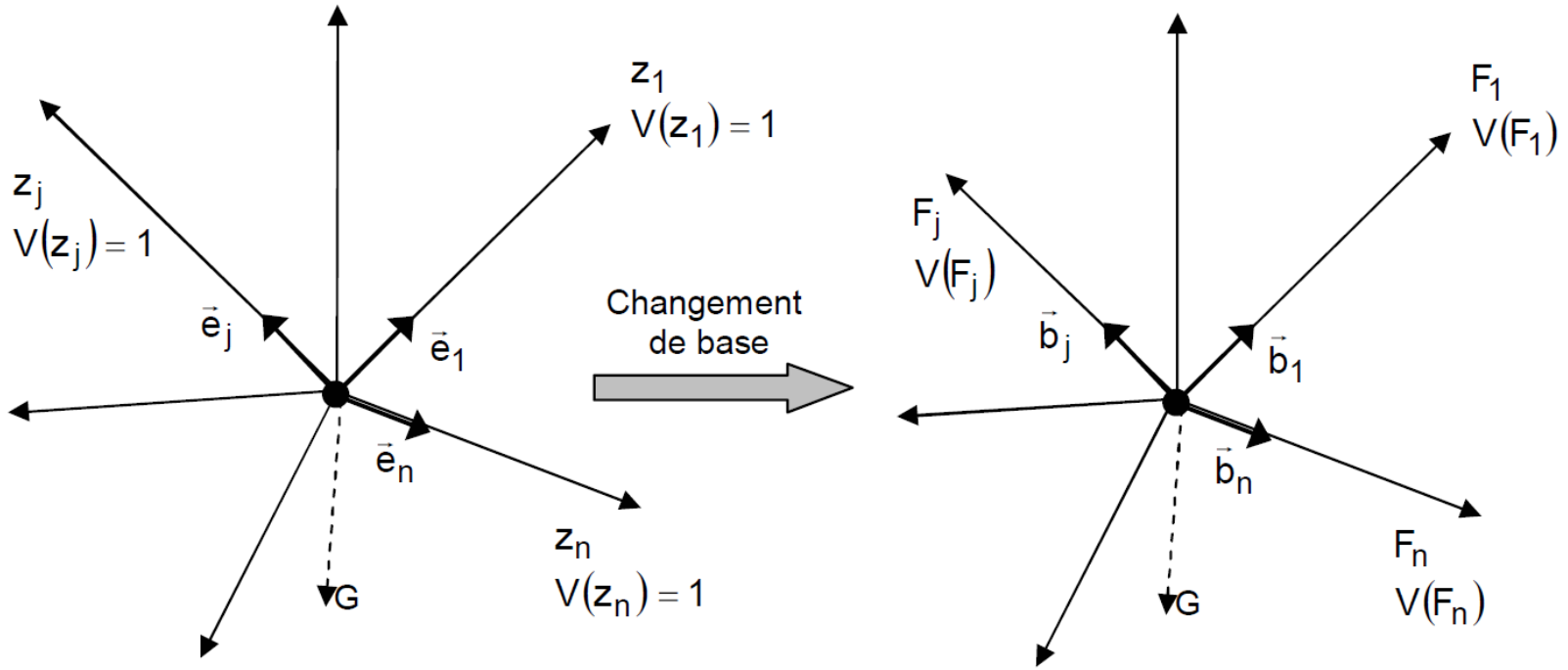
Qualité des représentations sur les plans principaux

Variables originales centrées réduites

- $Var(Z_1) = Var(Z_2) = \ldots = 1$
- $\sum_{j=1}^{p} var(Z_j) = p$
- Variables de même importance difficilement analysées

Espace construit par l'ACP

- $Var(F_1) > Var(F_2) > \ldots > Var(F_p)$
- $\sum_{j=1}^{p} var(F_j) = p$
- Grande partie (70 à 80%) de l'information retenue par les premiers facteurs (2 ou 3)



Analyse en Composantes Principales :

Qualité des représentations sur les plans principaux

Le pourcentage d'inertie

Le but de l'ACP étant de représenter les individus dans un espace de dimension plus faible que P (le nombre de variables), **la question qui se pose est d'apprécier la perte d'information subi et de savoir combien de facteurs retenir**.

- Le critère habituellement utilisé est le pourcentage d'inertie totale expliquée :

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

- Le choix du nombre d'axe à retenir est un point essentiel qui **n'a pas de solution rigoureuse**.
- Remarquons aussi que **la réduction des dimensions n'est possible que s'il y a redondance/dépendance** entre les variables de base.

Analyse en Composantes Principales :
Qualité des représentations sur les plans principaux

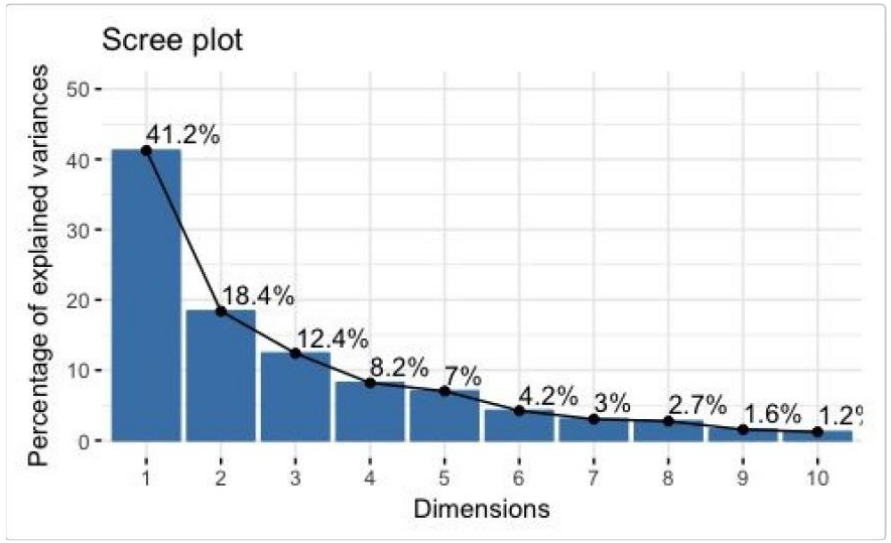
Choix de la dimension: Exemples de critères empiriques

- Retenir les valeurs propres telles que:

$$\lambda > 1+2\sqrt{\frac{P-1}{n-1}}$$

- Diagramme de décroissance des valeurs propres:

Chercher le coude séparant les valeurs propres utiles de celles qui sont peu différentes entre elles et qui n'apportent pas d'information



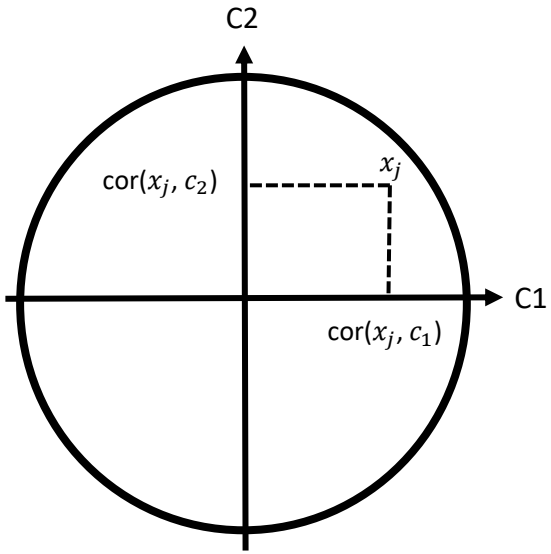
Aucun critère n'est absolu, l'interprétation des résultats d'une analyse relève du métier du statisticien qui doit tenir compte, entre autre, de la taille du tableau de données et des corrélations entre les variables

Analyse en Composantes Principales : corrélations «variables - facteurs»

La méthode triviale pour donner sens à une composante principale c'est de la relier aux variables initiales $X_1 + X_2 + \cdots + X_P$ via les coefficients de corrélation linéaire.

$$C_k = Z \times V_k \quad \Leftrightarrow \quad C_j = Z \times \begin{bmatrix} v_{k1} \\ v_{k2} \\ \cdots \\ v_{kP} \end{bmatrix} \quad \text{tel que} \quad \|V_k\| = 1$$

- **Le coefficient de corrélation de C_k avec X_j est : $cor(C_k, X_j) = \sqrt{\lambda_k}\, v_{kj}$**
- Pour un couple de composantes principales, on synthétise usuellement les corrélations sur une figure appelée **cercle de corrélation**

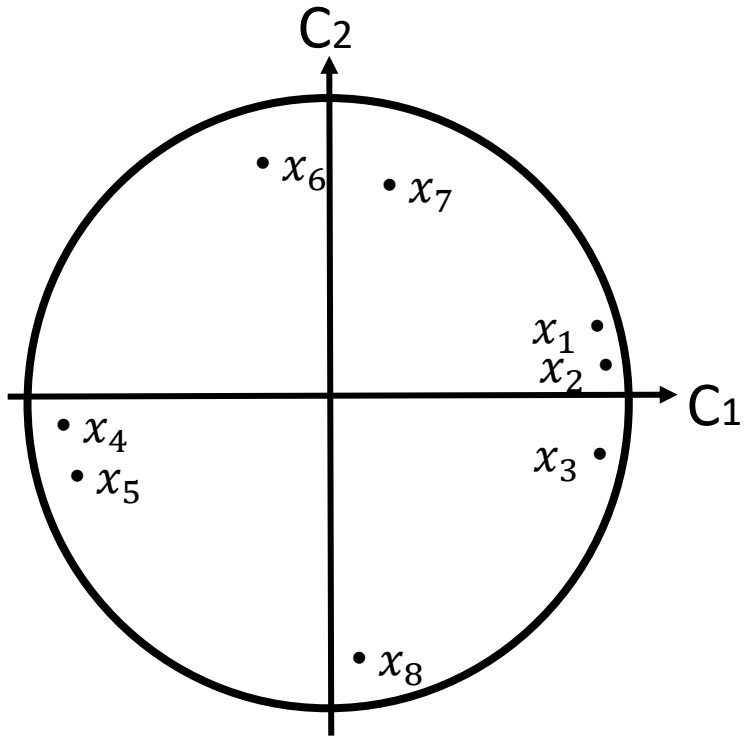


Analyse en Composantes Principales : corrélations «variables - facteurs»

Exemple:

Le cercle de corrélation projette les variables sur l'espace défini par le couple (C_1 , C_2).

- La figure montre que la composante C_1 est positivement corrélée avec les variables x_1, x_2 et x_3 , négativement corrélée avec les variables x_4 et x_5 et non linéairement corrélée avec les variables x_6, x_7 et x_8
- La composante C_2 oppose la variable x_8 aux variables $x_6, et\ x_7$
- On se gardera d'interpréter les proximités entre les points variables si ceux-ci ne sont pas proches de la circonférence.



Analyse en Composantes Principales : corrélations «variables - facteurs»

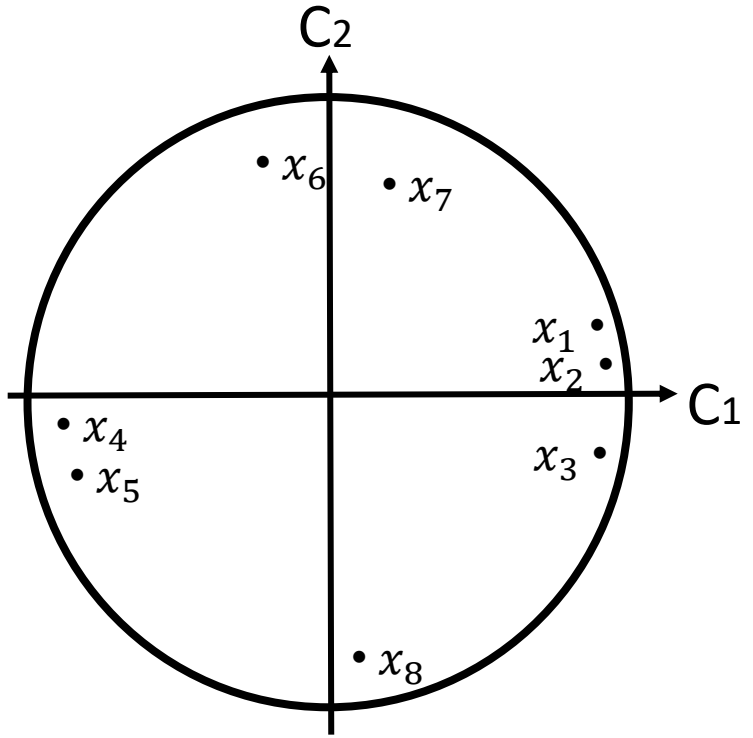
Contribution des variables aux axes

$$C_k = Z \times V_k \quad \Leftrightarrow \quad C_k = Z \times \begin{bmatrix} v_{k1} \\ v_{k2} \\ \dots \\ v_{kP} \end{bmatrix} \quad \text{tel que} \quad \|V_k\| = 1$$

Comme $\lambda_k = \sum_{j=1}^{P} cor^2(C_k, X_j)$

On appelle **contribution de la variable X_j à l'axe C_k le rapport:**

$$\frac{cor^2(C_k, X_j)}{\lambda_k} = (v_{kj})^2$$



Analyse en Composantes Principales : relations «Individus - facteurs»

Contribution des individus aux axes

Considérons la k^{ième} composante $C_k=(c_{k1}, c_{k2}, \ldots, c_{kN})$

$$Var(C_k) = \sum_{i=1}^{N} p_i \times {c_{ki}}^2 = \lambda_k \text{ avec } p_i \text{ est le poid de l'individus i}$$

On appelle **contribution de l'individus i à l'axe C_k le rapport:** $\frac{\frac{1}{N} \times {c_{ki}}^2}{\lambda_k}$

Lorsque $p_i = \frac{1}{N}$, ladite contribution devient : $\frac{\frac{1}{N} \times {c_{ki}}^2}{\lambda_k}$

Analyse en Composantes Principales : relations «Individus - facteurs»

Pour N assez grand:

$$c_{ki} \sim N(0, \sqrt{\lambda_k}) \Rightarrow \frac{{c_{ki}}^2}{\lambda_k} \sim \chi(1) \Rightarrow P\left(\frac{\frac{1}{N} \times {c_{ki}}^2}{\lambda_k} > \frac{3.84}{N}\right) = 0.05$$

- On peut considérer alors qu'un individus a une contribution significative si elle dépasse 4 fois son poids.

- Quand les individus ne sont pas anonymes, ils aident à l'interprétation des composantes principales. On cherchera par exemple les individus prépondérant en terme de contribution à l'axe ou les individus opposés le long de l'axe.

TP ACP sur R