Introduction à l'analyse des données

O. Mory

LMI-SFA-UNA

2 février 2018

Les méthodes Usuelles

- Analyse en Composantes Principales (A.C.P.) (K. Pearson, 1901)
- Analyse factorielle des correspondances simples (A.F.C.)(Hirchsfeld, 1936)
- Analyse factorielle des correspondances multiples (A.C.M.)(Guttman, 1941)
- Méthodes de classification automatique

Analyse des données

Analyser des données, c'est extraire d'une masse d'informations brutes, des éléments de réponse aux questions qui résultent des objectifs globaux poursuivis.

Analyse des données

Traitement de données en masse

- Grand nombre d'individus
- Grand nombre de variables

Analyse des données

Développement parallèle à l'informatique

- Fichiers volumineux $\Rightarrow$ demande de méthodes
- Capacité de calcul $\Rightarrow$ méthodes praticables

Analyse des données

Les outils mathématiques de l'analyse des données :

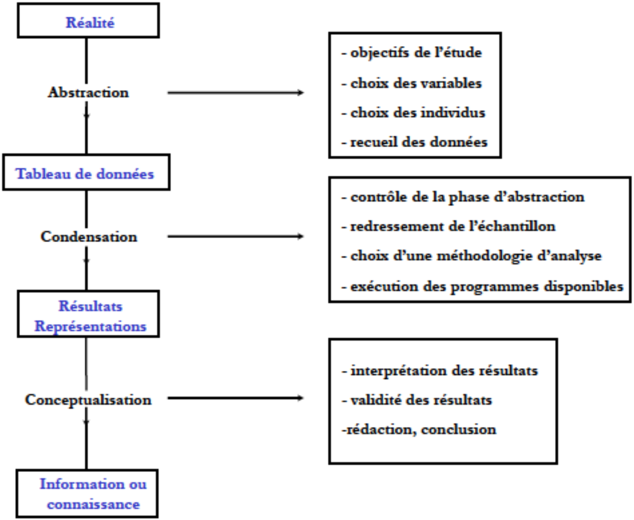
- Algèbre linéaire
- Calcul matriciel

Analyse des données

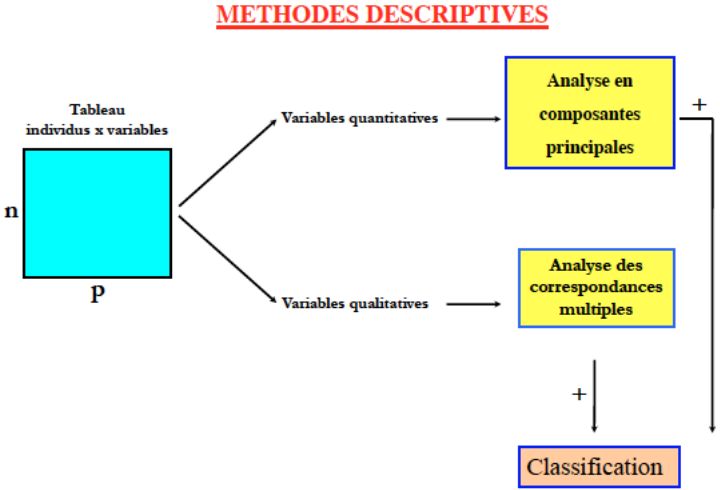
Les outils mathématiques de l'analyse des données :

- Analyse en Composantes Principales (A.C.P.)
- Analyse factorielle des correspondances
  analyse de variables qualitatives
  - Correspondances simples (A.F.C.) (Étude d'un tableau de contingence)
  - Correspondances multiples (A.C.M.) (Utile lors du dépouillement d'enquêtes)

Introduction



Introduction



ACP

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

$\Rightarrow$ Visualisation des individus (Notion de distances entre individus)

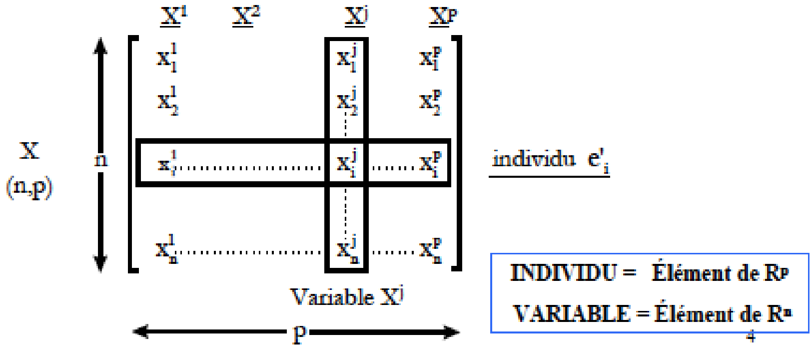
$\Rightarrow$ Visualisation des variables (en fonction de leurs corrélations)

Interprétation des axes

- Mesurer la qualité des représentations obtenues
  - critère global
  - critères individuels
- Utilisation éventuelle de variables supplémentaires

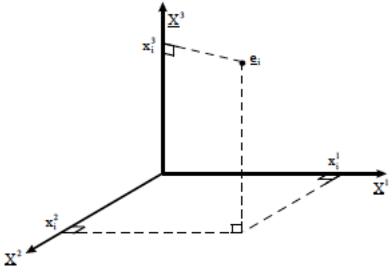
Les données

p variables quantitatives observées sur n individus.



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.



$\boxed{\text{Comment faire la représentation 4D ?}}$

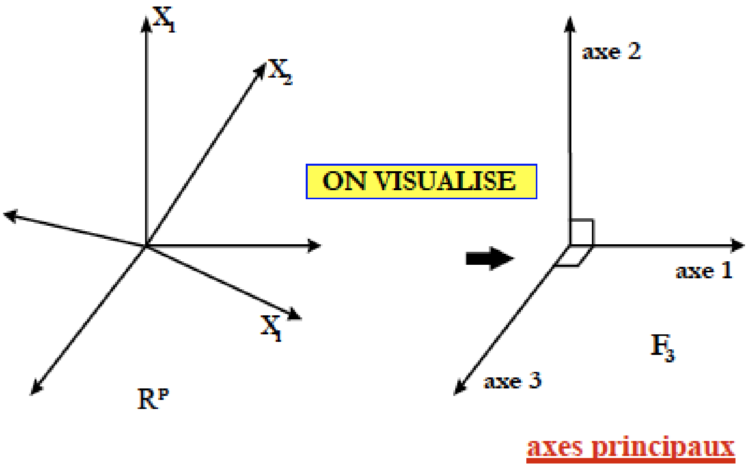
Principe de l'ACP

Principe

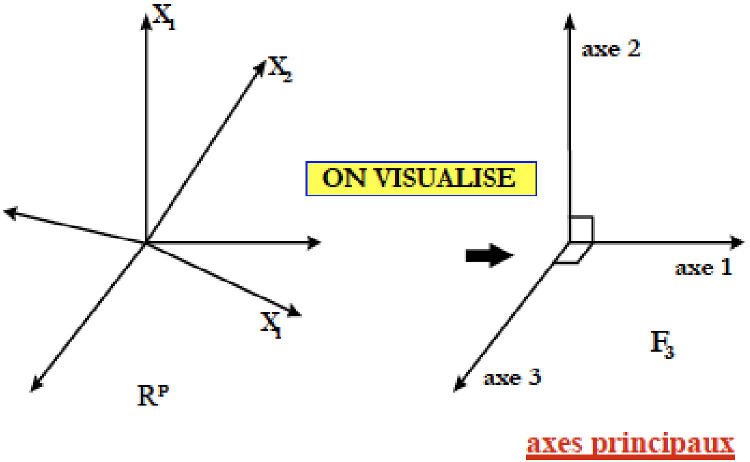
On cherche une représentation des n individus , dans un sous-espace F_k de R^p de dimension k<p (k petit 2, 3 . . . ; par exemple un plan)
Autrement dit, on cherche à définir k nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le moins d'information possible.

Vocabulaire

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



Qu'espere t'on ?



Qu'espere t'on ? $\Rightarrow$ 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.

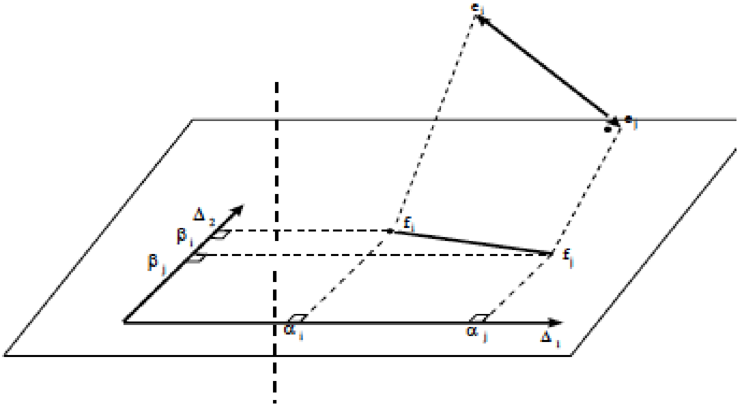
⇕

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 :

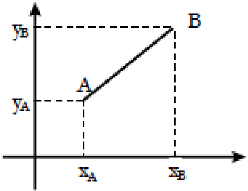
distance

projection orthogonale



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

Notion de distance euclidienne



Dans le plan:

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

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

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

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

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

Le problème des unités ?

10

Que faire ?

Normaliser

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: $\boxed{\dfrac{x_i^k - \overline{x}^k}{s_k}}$

où : x^k = moyenne de la variable X^k

s_k = écart-type de la variable X^k

Inertie

$$\boxed{I_{\underline{g}} = \sum_{i=1}^{n} \frac{1}{n} d^2\left(e_i, \underline{g}\right)}$$

ou de façon plus générale

$$\boxed{I_{\underline{g}} = \sum_{i=1}^{n} p_i \, d^2\left(e_i, \underline{g}\right)} \qquad \text{avec} \quad \sum_{i=1}^{n} p_i = 1$$

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.

Inertie

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

En notant V la matrice de variances-covariances :

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

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

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

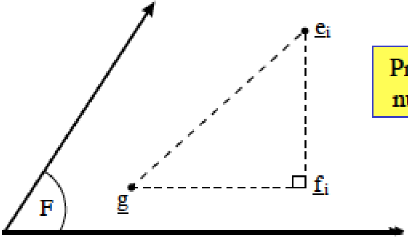
Remarque

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).

Inertie

Équivalence des deux critères concernant la perte d'information



Projection orthogonale du nuage sur un sous-espace

Soit F un sous-ensemble de R^p

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

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

14

Inertie

On va chercher F tel que :

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

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

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

Inertie

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

Donc : $\underbrace{\sum_{i=1}^{n} p_i \|\underline{e}_i - \underline{g}\|^2}_{\text{Inertie totale}} - \underbrace{\sum_{i=1}^{n} p_i \|\underline{e}_i - \underline{f}_i\|^2} = \underbrace{\sum_{i=1}^{n} p_i \|\underline{f}_i - \underline{g}\|^2}$

Inertie totale

minimiser cette quantité (carrés des distances entre points individus et leurs projections) $\Leftrightarrow$ maximiser l'inertie du nuage projeté

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 c1 est le vecteur renfermant les cordonnées des projections des individus sur l'axe 1.
- La composante c2 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.

$$c^1 = u_1^1 * X^1 + u_1^2 * X^2 + ... + u_1^p * X^p$$

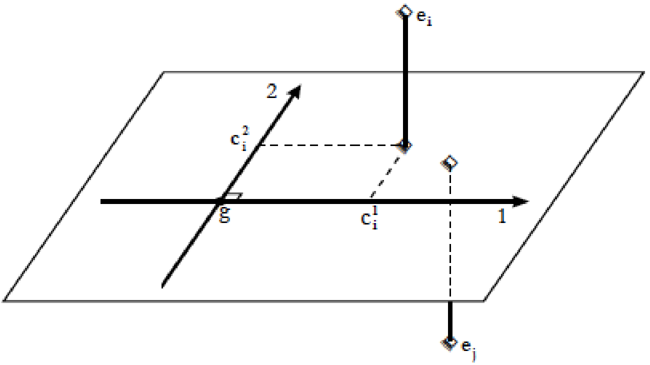
PROPRIÉTÉS DES COMPOSANTES PRINCIPALES

- La variance d'une composante principale est égale à l'inertie portée par l'axe principal qui lui est associé.
  - 1ère composante c^1 variance :
  - 2ème composante c^2 variance :
  - 3ème composante c^3 variance :
- Les composantes principales sont non corrélées deux à deux. En effet, les axes associés sont orthogonaux. λ_1

REPRÉSENTATION DES INDIVIDUS

La $j^{\text{ème}}$ 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 $j^{\text{ème}}$ axe principal.

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

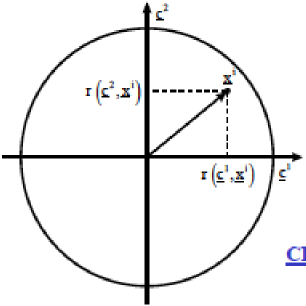


Attention à la qualité de représentation de chaque individu!

REPRÉSENTATION DES VARIABLES

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

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



CERCLE DES CORRÉLATIONS

23

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.

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

On suppose les **variables centrées**.

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

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

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

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

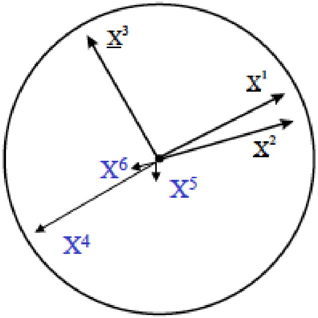
Coefficient de corrélation linéaire

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

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

VALIDITÉ DES REPRÉSENTATIONS

$$\boxed{\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ \dots\ \lambda_p}}$$

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

Exemple :

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

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.

28

Combien d'axes ?

Différentes procédures sont complémentaires:

① Pourcentage d'inertie souhaité : a priori

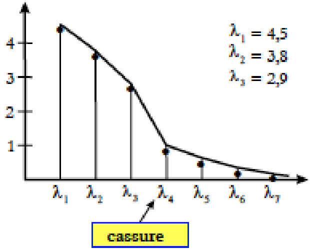
② Diviser l'inertie totale par le nombre de variables initiales

➪ 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).

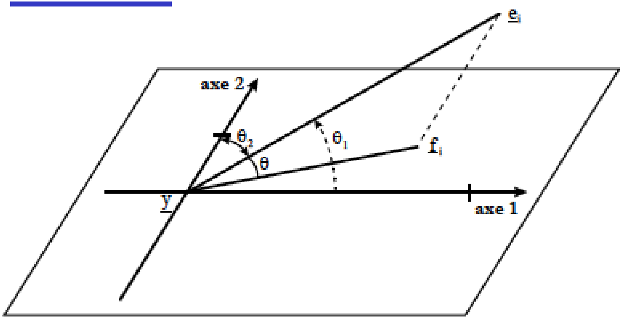
③ Histogramme

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



29

Cosinus carrés



$$\boxed{\cos^2\theta = \cos^2\theta_1 + \cos^2\theta_2}$$

30

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.

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

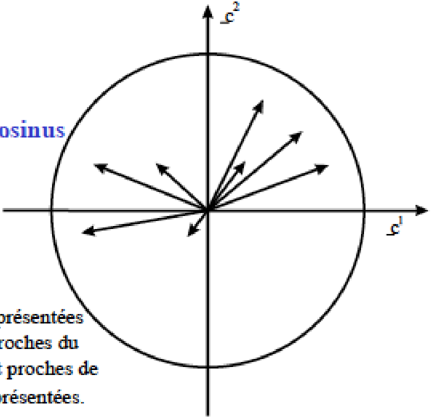
La contribution de l'individu $\underline{e}_i$ à la composante n° k est définie par

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

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.