I Objectif de l'ADD

- faciliter l'interprétation, la classification des données

3 types de méthode d'ADD :

1. Analyse en composantes principales (ACP)
2. Analyse factorielle des correspondances (AFC)
3. Analyse des composantes multiples (ACM)

· ce sont des méthodes d'analyse factorielle

Démarche :

- Traiter les données initiales afin d'éliminer les biais statistiques (ou effets d'échelle)
- Quantifier la notion d'information
  - utilisation de la variance, la covariance, corrélation
  - introduction de la notion de "distances" entre les points et un point "central".
- Réalisation de changements d'axe permettant de hiérarchiser l'information
- Choix des axes de projection en contrôlant la perte d'information
- Projection et analyse du nuage de points en validant l'analyse par des paramètres numériques

II Types de données

2. 1) Tableaux de données

Ce sont des tableaux à double-entrée

- les lignes du tableau peuvent correspondre aux individus étudiés (ACP, ACM) ou aux modalités d'une variable (AFC)
- les colonnes correspondent aux variables (ACP · ACM) ou aux modalités d'une variable (AFC)

- Ce tableau contient les modalités des variables (ACP - ACM) ou les effectifs (AFC)

2.2) Variables

- en ACP : variables quantitatives de même importance (à peu près)
- en AFC : 2 variables
- en ACM : x variables

III Calcul matriciel et ADD

3.1) Matrice symétrique

Propriété : Soit M une matrice quelconque. Alors les matrices transposées ${}^tM \times M$ et $M \times {}^tM$ sont symétriques $M = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}$ ${}^tM = \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix}$

3.2) Diagonalisation

Propriété : Toute matrice symétrique est diagonalisable dans $\mathbb{R}$

3.3) Distance

u et v 2 vecteurs colonnes dans une base orthonormée

$u = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ $v = \begin{pmatrix} -3 \\ 4 \end{pmatrix}$

$\|u\| = \sqrt{1^2 + 2^2} = \sqrt{{}^tu \times u}$

$${}^tu \times u = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 1^2 + 2^2$$

produits scalaires de u et v

$u \cdot v = 1 \times (-3) + 2 \times 4$

${}^tu \times v = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} -3 \\ 4 \end{pmatrix} = 1 \times (-3) + 2 \times 4$

IV Approche géométrique ellipsoïde d'inertie

Voir image.

Chapitre II : Paramètres d'un nuage de points (ACP)

Approche statistique

I Paramètres statistiques

- X variable statistique ou aléatoire
- n nbre d'individus
- écart type $\sigma_X = \sqrt{V(X)}$
- $\rho(X,Y) \in [-1,1]$
- $cov(X,X) = V(X)$
- $\rho(X,X) = 1$
- corrélation très forte si $\rho(X,Y) \simeq 1$
- —— — faible si —— $\simeq 0$
- anticorrélation très forte si $\rho(X,Y) \simeq -1$

Variance : $V(x) = E[(x-\bar{x})^2] = \frac{1}{n}\sum(x-\bar{X})^2$

Moyenne : $\bar{X} = E(X) = \frac{1}{n}\sum x$

Covariance : $cov(X;Y) = E[(X-\bar{X})(Y-\bar{Y})]$
$= \frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y})$

Corrélation : $\rho(X;Y) = \dfrac{cov(X;Y)}{\sigma_x \sigma_y}$

Variables (vecteurs) centrées : X^c

$$X^c = X - \bar{X} \Rightarrow \bar{X}^c = 0$$

Variables (vecteurs) centrées et réduites : X^s

$$X^s = \frac{X - \bar{X}}{\sigma_X} \Rightarrow \bar{X}^s = 0 \text{ et } V(X^s) = 1$$

II. Approche géométrique

ADD

n points (n individus)
p variables

variables centrées => centre du nuage de point = origine du repère

2.1) Matrice de variance - covariance

X Matrice des données initiales

$\dim X = (n, p)$

$\downarrow$

X^c Matrice des données centrées

$\dim X^c = (n, p)$

$$\Sigma = \frac{1}{n} \underset{(pn)\times(np)}{{}^t X^c X^c}$$

matrice carrée d'ordre p

symétrique

$$\Sigma = \begin{pmatrix} var(x_1^c) & cov(x_1^c, x_2^c) & cov(x_1^c, x_3^c) \\ cov(x_2^c, x_1^c) & var(x_2^c) & cov(x_2^c, x_3^c) \\ cov(x_3^c, x_1^c) & cov(x_3^c, x_2^c) & var(x_3^c) \end{pmatrix}$$

$\text{trace } \Sigma = var(A) + var(B) + var(C)$
$\text{trace } \Sigma = \text{var totale}$

2.2) Matrice des corrélations

$R = \frac{1}{n} {}^t X^s X^s$

X^s matrice des données centrées réduites $\dim X^s = (n, p)$

$$R = \frac{1}{n} \underset{(p,n)}{{}^t X^s} \underset{(n,p)}{X^s}$$ matrice carrée d'ordre p symétrique

diagonale de 1
termes compris entre -1 et 1

à quoi ça ressemble ? go voir campus (ou Martin)

Analyse en composantes principales (ACP)

I Aperçu général

n individus

p variables quantitatives, de même importance

$$X_1 \dots X_p$$

Si il y a des corrélations entre les variables $X_1\ X_2 \dots X_p$ → redondance

idée générale : Transformer les P variables redondantes $X_1, X_2, \dots X_p$ en P variables appellées facteurs $f_1, f_2, \dots f_p$ non redondants (non corrélés)

Remarque : les redondances sont inscrites dans Σ et R

1.1) La transformation

$$\underbrace{X_1, X_2 \dots X_p}_{\text{base}} \rightarrow \underbrace{f_1, f_2, \dots f_p}_{\text{autre base}}$$

base → autre base

Σ ou R → matrice diagonale Λ

Il faut diagonaliser Σ ou R

- On obtient une matrice diagonale Λ (matrice des valeurs propres)
- On obtient une matrice de changements de base U

→ On détermine les composantes des individus dans la nouvelle base (les facteurs)

$$F = {}^{s}XU \text{ ou } X^{c}U$$ matrice des composantes principales

F matrice de dimension (n, p), matrice centrée (Σ en colonne $= 0$) mais plus réduite

1.2) l'interprétation

Problème : donner un sens aux facteurs principaux
corrélations entre variables et facteurs

→ matrice des saturations

$$S = \frac{1}{n}\ {}^{t}X^{s} . F$$

Λ matrice des variances des facteurs

$\Lambda^{1/2}$ matrice des écarts-types des facteurs

II l'ACP, pas à pas sur un exemple

ACP : Pourquoi ?
- variables quantitatives
- variables de "même importance" (car huile lait)

quel type ?

ACP centrée réduite si :
- variables en unités différentes
- les ordres de grandeur sont trop ≠
- les dispersions sont ≠

Matrice centrée : somme colonne = 0

Diagonaliser une matrice A

1. déterminer ses valeurs propres
   ⇒ résoudre équat° $\det(A - \lambda I) = 0$
2. En déduire la matrice diagonale Δ semblable
   $\text{trace } \Delta = \text{trace } A$
3. déterminer la base de vecteurs propres

si u vecteur propre associé à λ, alors $Au = \lambda u$

Propriétés des vecteurs propres :

- vecteurs propres orthogonaux
- choix de vecteurs propres unitaires

Matrice des composantes principales
= matrice des coordonnées des individus par rapport aux nouveaux axes.

qualité de représentation d'un individu par

Matrice des saturations

= matrice des corrélations entre les variables et les facteurs
= matrice des coordonnées des variables par rapport aux axes (factoriels)

Prop. de la matrice de saturat°; cercle de corrélation

$S = [s_{ij}]_{\substack{1 \leq i \leq j \\ 1 \leq j \leq p}}$

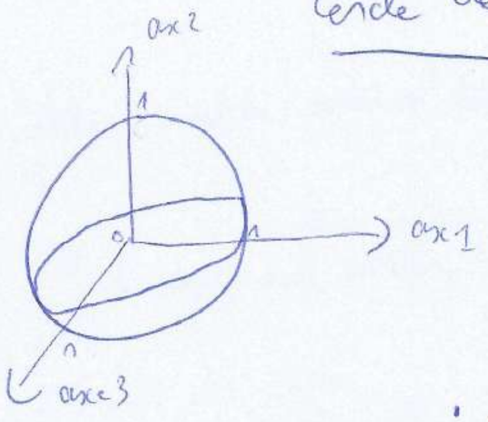
$\sum_{j=1}^{p} s^2_{ij} = 2$ ou 1? ⚠ [..]

$\sum_{i=1}^{p}$

En dim 2, $p=2$ $x^2+y^2=1$ → cercle de centre O et de rayon 1
——— 3, $p=3$ $x^2+y^2+z^2=1$ → sphère ———

$\sum_{j=1}^{p} s^2_{ij} = 1$ → les points-variables dans les axes factoriels sont situés sur une hypersphère de centre O et de rayon 1.

Cercle de corrélation



- les projetés des variables sont situés à l'intérieur du cercle de corrélat°.
- Les variables "intéressantes" d'un point de vue graphique sont celles dont les projettés "proches" du cercle

---

Analyse des notes de 34 élèves de 2nde générale

questions à se poser : Pourquoi une ACP ? → car variables quantitatives, variables "de m̂ importance"

Quel type d'ACP ?

↳ ACP centrée réduite car : ~~• unités ≠~~ (pas ds ce cas)

statistiques élém. ← • ordres de grandeur ≠

• dispersion ≠

→ Interprétat° de la matrice de corrélat°

→ Select° des axes et pages.

3 premiers axes

qge : 68,6%

→ Interprétat° de la matrice de saturat° et sa représentat°, les sphères de corrélat°

à partir du cercle :

- axe 1 : Maths Fr
  Phy

a partir de matrice :

Maths Fr

Phy SVT EPS

axe 2 : EPS

— 3 : ≈ SES

interprétat° des composantes principales

-> sélect° des individus : proche de l'axe, en situat° extrême, de bonne qualité

axe 1:

i_{31}	i_{25}
i_{12}	i_3
i_{11}	i_{23}
i_{13}	$\cancel{i_{15}}$