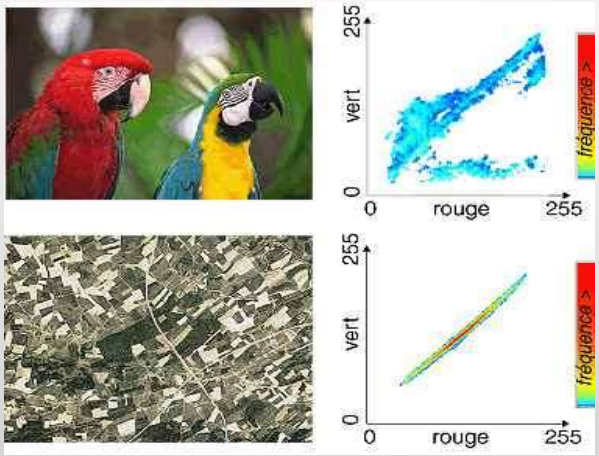
ANALYSE DE DONNÉES

INFO 4 INSI



1.5. PLAN DU COURS

I. INTRODUCTION

1. PRÉSENTATION DU COURS ET PRÉ REQUIS
2. L'ANALYSE DE DONNÉES ET LES BIG DATA
3. HISTORIQUE ET DATES CLÉS
4. MODÈLES ET PROBLÈMES
5. BIBLIOGRAPHIE

1.1. PRÉSENTATION DU COURS ET PRÉ REQUIS

- 20 HEURES DE COURS/TD & 10 HEURES DE TP
- ÉVALUATION : EXAMEN (14/20) + TP (6/20)
- CONTENU (MÉTHODES; TEXTES, SONS, IMAGES)
- SUPPORTS : POLYCOPIÉS, DIAPOSITIVES ET BIBLIOGRAPHIE
- PRÉ REQUIS : PROBABILITÉS & STATISTIQUES, INFORMATION, MATRICES
- OBJECTIF : MAITRISE DES MÉTHODES ET ALGORITHMES D'ANALYSE ET DE TRAITEMENT DE FLUX DE DONNÉES MASSIFS

1.2. ANALYSE DE DONNÉES ET BIG DATA

- FLUX DE DONNÉES DANS LE WEB MONDIAL :
  - ✓ 1992 : 100GO/JOUR
  - ✓ 2002 : 100GO/S
  - ✓ 2017 : 50000 GO/S (> 3 MILLIARDS D'UTILISATEURS)
  - ✓ 2019 : 75000 GO/S (4.2 MILLIARDS D'UTILISATEURS; 80% VIDÉO)

- TRAITEMENT DE DONNÉES (DE L'INFORMATION) :
  - ✓ - RECHERCHER, FOUILLER, EXTRAIRE
  - ✓ - CLASSER, TRIER (CLASSIFIER)
  - ✓ - ANALYSER

- LOGICIELS STATISTIQUES/DATA : SAS, R (SPLUS), XIPLUS (EXCEL)
- + MATHS : MATLAB/OCTAVE, MATHEMATICA, MAPLE

:

1.3. HISTORIQUE ET DATES CLÉS

- INFÉRENCE STATISTIQUE (NEYMAN, PEARSON, FISHER, 1900)
- ANALYSE FACTORIELLE (SPEARMAN, 1904)
- ACP (HOTELLING, 1933); KLT (LOÈVE, 1955)
- CLASSIFICATION NON SUPERVISÉE (HIÉRARCHIQUE) (HILLMAN, 1965, …)
- NUÉES DYNAMIQUES (FRIEDMAN, MACQUEEN,1967)
- RÉSEAUX DE NEURONES : PERCEPTRON (ROSENBLATT, 1958)
- PERCEPTRON MULTICOUCHES ET APPRENTISSAGE SUPERVISÉ (80'S)
- ANALYSE FACTORIELLE DES CORRESPONDANCES (BENZÉCRI, 1982)
- LOGIQUES ET ENSEMBLES FLOUS, INFORMATION & CONNAISSANCE, RÉSEAUX TYPOLOGIQUES, TYPES ABSTRAITS, …

1.4. MODÈLES ET PROBLÈMES

- APPROCHES QUALITATIVES ET QUANTITATIVES :
  - RESSEMBLANCES, CORRÉLATIONS, CORRESPONDANCES
  - ANALYSE ET VISUALISATION PAR RÉDUCTION DE DIMENSION
- REPRÉSENTATION DES DONNÉES : ANALYSES FACTORIELLES (ACP, ANALYSE DES CORRESPONDANCES, CORRESPONDANCES MULTIPLES, ANALYSE CANONIQUE)
  - ACP : DÉCOMPOSITION DANS UN REPÈRE (BASE) ET PROJECTIONS ORTHOGONALES
- CLASSIFICATION : RÉPARTITION DE POPULATION EN *N* (*N* INCONNU ?) CLASSES HOMOGÈNES
  - SUPERVISÉE (ANALYSE DISCRIMINANTE)
  - NON SUPERVISÉE (CENTRES MOBILES, HIÉRARCHIQUE)
- DONNÉES MULTIDIMENSIONNELLES : REPRÉSENTATION MATRICIELLE

1.5. BIBLIOGRAPHIE

I. G. SAPORTA : *PROBABILITÉS, ANALYSE DE DONNÉES ET STATISTIQUE* TECHNIP, 2011

II. M. AMINI, E. GAUSSIER : *RECHERCHE D'INFORMATION* EYROLLES, 2013

III. R. NISBET, G. MINER : HANDBOOK OF STATISTICAL ANALYSIS AND DATA MINING APPLICATIONS ACADEMIC PRESS, 2017

IV. F. ZAMORA MOLA : CLASSIFICATION, BIG DATA ANALYSIS AND STATISTICAL LEARNING SPRINGER, 2017

V. F. CADY : *THE DATA SCIENCE HANDBOOK* WILEY, MARS 2017

II. PROBABILITÉS ET STATISTIQUES

CALCUL DES PROBABILITÉS

- PROBABILITÉ, ÉVÉNEMENTS ET VARIABLES ALÉATOIRES
- LOI DE PROBABILITÉ DISCRÈTE, LOIS DE BERNOULLI, BINOMIALE, ÉQUIDISTRIBUÉE, HYPERGÉOMÉTRIQUE, DE POISSON
- APPLICATIONS ET PROBLÈMES
- ESPACES PROBABILISÉS
- LOIS CONTINUES, FONCTION DE RÉPARTITION ET DENSITÉ DE PROBABILITÉ
- LOIS USUELLES
- INDÉPENDANCE ET PROBABILITÉS CONDITIONNELLES
- TRANSFORMÉES ET SOMMES DE VARIABLES INDÉPENDANTES

2.1. LOIS DE PROBABILITÉ

- PROBABILITÉ, ÉVÉNEMENTS ET VARIABLES ALÉATOIRES

1. MESURE DE PROBABILITÉ : SI $A \cap B = \varnothing$, $P(A \cup B) = P(A) + P(B)$

ADDITIVITÉ DÉNOMBRABLE : (A_k), $k \in N$; $P(\cup A_k) = \sum P(A_k)$

TRIBU DES ÉVÉNEMENTS : STABLE POUR $\cap$, $\cup$, COMPLÉMENTAIRE

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, FORMULES DU CRIBLE :

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

2.1. LOIS DE PROBABILITÉ

➢ PROBABILITÉ, ÉVÉNEMENTS ET VARIABLES ALÉATOIRES

1. MESURE DE PROBABILITÉ : SI $A \cap B = \varnothing$, $P(A \cup B) = P(A) + P(B)$

ADDITIVITÉ DÉNOMBRABLE : (A_k), $k \in N$; $P(\cup A_k) = \sum P(A_k)$

TRIBU DES ÉVÉNEMENTS : STABLE POUR $\cap$, $\cup$, COMPLÉMENTAIRE

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, FORMULES DU CRIBLE :

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

2. PROBABILITÉ CONDITIONNELLE ET INDÉPENDANCE : SI $P(B) \neq 0$, $P(A/B) = P(A \cap B) / P(B)$

A ET B SONT INDÉPENDANTS : $P(A \cap B) = P(A) \cdot P(B)$

FORMULE DE BAYES : SI (B_k) EST UNE PARTITION DE L'UNIVERS, $P(A) = \sum P(A/B_k) \cdot P(B_k)$

2.1. LOIS DE PROBABILITÉ

➢ PROBABILITÉ, ÉVÉNEMENTS ET VARIABLES ALÉATOIRES

1. MESURE DE PROBABILITÉ : SI $A \cap B = \varnothing$, $P(A \cup B) = P(A) + P(B)$

ADDITIVITÉ DÉNOMBRABLE : (A_k), $k \in N$; $P(\cup A_k) = \sum P(A_k)$

TRIBU DES ÉVÉNEMENTS : STABLE POUR $\cap$, $\cup$, COMPLÉMENTAIRE

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, FORMULES DU CRIBLE :

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

2. PROBABILITÉ CONDITIONNELLE ET INDÉPENDANCE : SI $P(B) \neq 0$, $P(A/B) = P(A \cap B) / P(B)$

A ET B SONT INDÉPENDANTS : $P(A \cap B) = P(A) \cdot P(B)$

FORMULE DE BAYES : SI (B_k) EST UNE PARTITION DE L'UNIVERS, $P(A) = \sum P(A/B_k) \cdot P(B_k)$

3. VARIABLE ALÉATOIRE : X; VALEURS ET LOI DE PROBABILITÉ $P(X \in B)$, $P(X = x_k)$

CAS DISCRET : TABLEAUX À 2 COLONNES, ÉCRITURE DIRECTE OU SUITES

CAS CONTINU : FONCTION DE RÉPARTITION (F), DENSITÉ ($f = F'$); $P(X \in B) = \int_B f(x)\,dx$

2.1. LOIS DE PROBABILITÉ

- PROBABILITÉ, ÉVÉNEMENTS ET VARIABLES ALÉATOIRES

4. TABLEAU ALÉATOIRE : T = [Y1, Y2, …, Ym], OÙ LES COLONNES CONTIENNENT N VARIABLES ALÉATOIRES (INDÉPENDANTES DE MÊME LOI DE PROBABILITÉ)

SI ON OBSERVE DES RÉALISATIONS DES MÊMES VARIABLES ALÉATOIRES DISCRÈTES X1, X2, …, Xm,

ON PEUT TRADUIRE L'INDÉPENDANCE PAR DES RELATIONS SUR LES EFFECTIFS :

$N_{i,j} \approx n\ P\ (X_j = x_i)$

L'INDÉPENDANCE DE Xj ET DE Xk SE TRADUIT PAR LES RELATIONS $n\ N_{i_1 i_2} \approx N_{i_1,i}\ N_{i_2,k}$

AVEC LES EFFECTIFS $N_{i_1 i_2} \approx n\ P\ (X_j = x_{i_1}$ et $X_k = x_{i_2})$

2.2. VALEURS MOYENNES ET CALCULS

ESPÉRANCE MATHÉMATIQUE

VARIABLE DISCRÈTE $E(X) = \sum p_k x_k$: somme ou série (si convergente)

ESPÉRANCE D'UNE VARIABLE CONTINUE; PROPRIÉTÉS ET CALCULS

$E(X) = \int x\, f(x)\, dx$: intégrale (si convergente); opérateur linéaire,

ESPÉRANCE ET LOIS USUELLE

Discrète : $E(X) = p$ pour X : B (p), $E(X) = n\,p$ pour B (n, p), $E(X) = \frac{1}{p}$ pour G (p), $E(X) = \lambda$ pour $\mathcal{P}(\lambda)$

Continue : $E(X) = \mu$ pour $X : \mathcal{N}(\mu, \sigma)$, $E(X) = \frac{a+b}{2}$ pour U (a, b),), $E(X) = \frac{1}{\lambda}$ pour exp (λ)

- VARIANCE ET MOMENTS $M_m(X) = E(X^m)$; $\sigma^2 = V(X) = E([X - E(X)]^2) = E(X^2) - [E(X)]^2$

CAS DES LOIS USUELLES/ CALCULS

- CORRÉLATION ET COVARIANCE $\text{cor}(X, Y) = E(XY)$; $\text{cov}(X, Y) = E([X - E(X)]\,[Y - E(Y)])$
- LOIS DES GRANDS NOMBRES si (X_k) VA indépendantes *id* d'espérance μ, $\frac{S_n}{n} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \to \mu$ (*)
- THÉORÈME DE LIMITE CENTRALE ET APPROXIMATION DE LOIS $\frac{S_n - n\,\mu}{\sigma\sqrt{n}} \cong \mathcal{N}(0, 1)$
- CALCULS DE MOYENNES ET DE PROBABILITÉS $E[g(X)] = \int g(x) f(x)\, dx$; $P[g(X) \in B] = \int_{g^{-1}(B)} f(x)\, dx$

2.3. VECTEURS ALÉATOIRES

- VECTEUR ALÉATOIRE, RÉPARTITION ET DENSITÉ

$$F_{X,Y}(x, y) = \mathrm{P}(X \le x,\ Y \le y)$$

$X = (X_1, \ldots, X_n)$, et $x \in \mathrm{R}^n$, $F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(X_1 \le x_1, \ldots, X_n \le x_n$

$$f_X(x) = \frac{\partial^n F_X}{\partial x_1 \ldots \partial x_n}(x) \quad \text{et} \quad F_X(x) = \int_{t_1=-\infty}^{x_1} \ldots \int_{t_n=-\infty}^{x_n} f_X(t)\, dt_1 \ldots dt_n$$

$$\mathrm{P}(X \in \mathrm{B}) = \int_{\mathrm{B}} f_X(x)\, dx$$

- RETOUR SUR LE CONCEPT D'INDÉPENDANCE

X et Y sont indépendantes si $\forall (x, y) \in \mathrm{R}^2$, $F_{X,Y}(x, y) = F_X(x)\, F_Y(y)$, ou $f_{X,Y}(x, y) = f_X(x) f_Y(y)$

$X_1, \ldots, X_n$ indépendantes dans leur ensemble

- FONCTION CARACTÉRISTIQUE

$Z = X + iY$, $\mathrm{E}(Z) = \mathrm{E}(X) + i\,\mathrm{E}(Y)$

$\psi_X(u) = \mathrm{E}(e^{i u X})$ caractérise X; cas vectoriel

ψ_X est uniformément continue sur R.

$\psi_{X+Y}(u) = \psi_X(u)\, \psi_Y(u)$

si X continue, $\psi_X(u) = \int_R f_X(x)\, e^{i u x}\, dx$

2.3. VECTEURS ALÉATOIRES

- VECTEUR ALÉATOIRE, RÉPARTITION ET DENSITÉ

$$F_{X,Y}(x, y) = \mathrm{P}(X \le x,\ Y \le y)$$

$X = (X_1, \ldots, X_n)$, et $x \in \mathrm{R}^n$, $F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(X_1 \le x_1, \ldots, X_n \le x_n$

$$f_X(x) = \frac{\partial^n F_X}{\partial x_1 \ldots \partial x_n}(x) \quad \text{et} \quad F_X(x) = \int_{t_1=-\infty}^{x_1} \ldots \int_{t_n=-\infty}^{x_n} f_X(t)\ dt_1 \ldots dt_n$$

$$\mathrm{P}(X \in \mathrm{B}) = \int_{\mathrm{B}} f_X(x)\ dx$$

- RETOUR SUR LE CONCEPT D'INDÉPENDANCE

X et Y sont indépendantes si $\forall (x, y) \in \mathrm{R}^2$, $F_{X,Y}(x, y) = F_X(x)\, F_Y(y)$, ou $f_{X,Y}(x, y) = f_X(x) f_Y(y)$

$X_1, \ldots, X_n$ indépendantes dans leur ensemble

- FONCTION CARACTÉRISTIQUE

$Z = X + iY$, $\quad \mathrm{E}(Z) = \mathrm{E}(X) + i\,\mathrm{E}(Y)$

$\psi_X(u) = \mathrm{E}(e^{i u X})$ caractérise X; cas vectoriel

ψ_X est uniformément continue sur R.

$\psi_{X+Y}(u) = \psi_X(u)\ \psi_Y(u)$

si X continue, $\psi_X(u) = \int_R f_X(x)\ e^{i u x}\ dx$

2.4. VECTEURS ALÉATOIRES. 2

➢ MATRICE DE COVARIANCE

VECTEUR ALÉATOIRE $X = (X_1, \ldots, X_n) \in L^2(R^n)$; MATRICE $\Gamma = \Gamma_X$: $\Gamma_{ij} = \mathrm{COV}(X_i\ X_j)$.

* Γ EST SYMÉTRIQUE (POSITIVE); DIAGONALISABLE ET FACTORISABLE (CHOLESKY)

* Γ EST DIAGONALE SI LES VARIABLES X_i SONT INDÉPENDANTES 2 À 2, ET $\Gamma_{ii} = V(X_i)$.

* Γ CARACTÉRISE LE DEGRÉ DE CORRÉLATION / DE DÉPENDANCE DES VARIABLES

➢ VECTEURS GAUSSIENS

SI $Y = A\,X$, X FORMÉ DE VARIABLES NORMALES RÉDUITES INDÉPENDANTES.

LOI DE X CARACTÉRISÉE PAR SES 2 PREMIERS MOMENTS : $\mu = E(X)$ ET $\Gamma = \Gamma_X$.

FONCTION CARACTÉRISTIQUE $\psi_X(u) = e^{\,i <u, \mu> - \frac{1}{2} <\Gamma u, u>}$ / DENSITÉ

➢ TRANSFORMATIONS ET CALCULS

$$E(g(X)) = \int f_x(x)\ g(x)\ dx\ ;\quad P(g(X) \in B) = \int_{g(x) \in B} f_x(x)\ g(x)\ dx$$

2.5. ESTIMATION

- MODÈLES ET ECHANTILLONS
- ESTIMATEURS SANS BIAIS, VARIANCE ET CONSISTANCE
- ESTIMATEURS EMPIRIQUES ET STATISTIQUE DESCRIPTIVE
- MOINDRES CARRÉS
- ESTIMATION PARAMÉTRIQUE
- VRAISEMBLANCE ET MAXIMUM DE VRAISEMBLANCE
- LE CAS BAYÉSIEN : PARAMÈTRES ALÉATOIRES

2.6. TESTS PARAMÉTRIQUES ET NON PARAMÉTRIQUES

- PROBLÉMATIQUE DES TESTS BIHYPOTHÈSES ET MULTIHYPOTHÈSES
- TESTS NON PARAMÉTRIQUES ET TEST DE LA LOI DE L'ÉCHANTILLON
- TEST DU KHI DEUX ET TESTS DE KOLMOGOROV
- TESTS D'INDÉPENDANCE ET D'ASSOCIATION
- TEST DE WILCOXON
- TESTS PARAMÉTRIQUES : TEST DE FISHER ET TEST DE STUDENT
- CAS MULTIVARIÉ ET TEST DE HOTELLING
- ANALYSE DE VARIANCE

2.6. TESTS PARAMÉTRIQUES ET NON PARAMÉTRIQUES

➢ PROBLÉMATIQUE DES TESTS BIHYPOTHÈSES ET MULTIHYPOTHÈSES

TEST MULTIHYPOTHÈSES : H_1, H_2, …, H_m (DÉTECTION, MODÈLES MULTIPLES)

TEST BIHYPOTHÈSES = TEST BINAIRE : H_0, H_1

PROBLÈME ASYMÉTRIQUE : L'HYPOTHÈSE H_0 EST LA PLUS DISCRIMINANTE

ON DOIT CHOISIR H_0 OU H_1 AVEC UN CRITÈRE DE DÉCISION C (C_0 OU C_1); ERREURS DE DEUXIÈME ET DE PREMIÈRE ESPÈCE (PUISSANCE DU TEST) :

$\alpha = P(C_1 / H_0)$

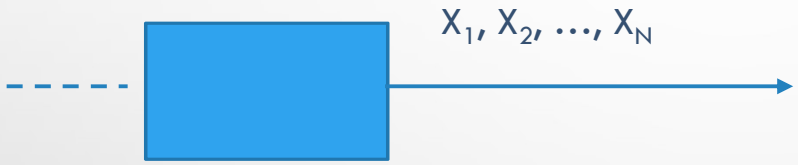
$\beta = P(C_0 / H_1)$

LE CRITÈRE C DÉPEND DE L'ÉCHANTILLON, DES DONNÉES DU MODÈLE, D'HYPOTHÈSES A PRIORI, D'UN CADRE PARAMÉTRIQUE* OU BAYÉSIEN*

2.6. TESTS PARAMÉTRIQUES ET NON PARAMÉTRIQUES

- PROBLÉMATIQUE DES TESTS BIHYPOTHÈSES ET MULTIHYPOTHÈSES
- TESTS NON PARAMÉTRIQUES ET TEST DE LA LOI DE L'ÉCHANTILLON

MODÈLE STANDARD : ECHANTILLON $X = (X_1, X_2, \ldots, X_N)$ $X_1, X_2, \ldots, X_N$ $X_1, X_2, \ldots, X_N$



F

VARIABLES ALÉATOIRES INDÉPENDANTES, DE MÊME LOI *F* INCONNUE

AUCUNE INFORMATION A PRIORI SUR F

HYPOTHÈSE $H_0 : F = F_0$

REMARQUE : CAS GÉNÉRAL; $H_0 : g(F) \in G_0$ ($h(F) = H$)

2.6. TESTS PARAMÉTRIQUES ET NON PARAMÉTRIQUES

- PROBLÉMATIQUE DES TESTS BIHYPOTHÈSES ET MULTIHYPOTHÈSES
- TESTS NON PARAMÉTRIQUES ET TEST DE LA LOI DE L'ÉCHANTILLON
- TEST DU KHI DEUX ET TESTS DE KOLMOGOROV

CRITÈRE : LOI EMPIRIQUE F_n, ET CRITÈRE D'ÉCART : D (F_n, F_0)

ON CHOISIT H_0 SI C_0 : $D(F_n, F_0) \leq \eta$

LE SEUIL η EST DÉTERMINÉ EN INVERSANT LA LOI DE PROBABILITÉ DE L'ÉCART D :

$P(D(F_n, F_0) > \eta / H_0) = \alpha$, OÙ α EST FOURNI A PRIORI (1%, 5%, 10%),

OU PLUS EXACTEMENT UNE LOI LIMITE (QUAND $n \to \infty$) : D (F_n, F_0) $\to$ Z :

$P(Z > \eta) = 1 - F_Z(\eta) = \alpha$

LA NATURE DU TEST EST DÉTERMINÉE PAR : - LA CLASSE DE F_0 (LOI CONTINUE, DISCRÈTE)

- LE CHOIX DU CRITÈRE D'ÉCART D (DICTÉ PAR - LA CONVERGENCE VERS Z)

2.6. TESTS PARAMÉTRIQUES ET NON PARAMÉTRIQUES

➢ TEST DU KHI DEUX ET TESTS DE KOLMOGOROV

LOI DU KHI DEUX À *r* DEGRÉS DE LIBERTÉ ET TEST DU χ^2

TEST D'UNE LOI DISCRÈTE.

X PEUT PRENDRE LES VALEURS $x_1, x_2, \ldots, x_m$

AVEC DES PROBABILITÉS INCONNUES $p_1, p_2, \ldots, p_m$

LOI EMPIRIQUE F_n CARACTÉRISÉE PAR LES FRÉQUENCES $N_1, N_2, \ldots, N_m$ DES x_k.

LOI THÉORIQUE F_0 : $p_1 = \pi_1,\ p_2 = \pi_2, \ldots, p_m = \pi_m$

CRITÈRE D'ÉCART : $D(F_n, F_0) = \frac{(N_1 - N\pi_1)^2}{N\pi_1} + \ldots + \frac{(N_m - N\pi_m)^2}{N\pi_m}$

CONVERGENCE : $D(F_n, F_0) \rightarrow Z\ (\chi^2_{m-1})$

$\alpha = P(D(F_n, F_0) > \eta / H_0) \cong P(Z > \eta)$ (TABLES INVERSES DU χ^2)

SI *n* GRAND : APPROXIMATION GAUSSIENNE

2.6. TESTS PARAMÉTRIQUES ET NON PARAMÉTRIQUES

➢ TEST DU KHI DEUX ET TESTS DE KOLMOGOROV

TEST DE KOLMOGOROV-SMIRNOV

TEST D'UNE LOI CONTINUE.

RÉALISATION DE L'ÉCHANTILLON : $x_1, x_2, \ldots, x_n$

LOI EMPIRIQUE F_n : FONCTION EN ESCALIER

LOI THÉORIQUE F_0

CRITÈRE D'ÉCART : $D(F_n, F_0) = \text{Sup}\ (\ |\ F_n(x) - F_0(x)\ |\ ,\ x \in \mathrm{R})$

CONVERGENCE : $P\left(D(F_n, F_0) > \frac{c}{\sqrt{n}}\right) \rightarrow \alpha(r)$

CALCUL : $D(F_n, F_0) = \text{Sup}\ (|\ F_n(x_{k+}) - F_0(x_{k+})\ |\ ,\ |\ F_n(x_{k-}) - F_0(x_{k-})\ |\ ,\ k = 1 \ldots n)$

ET $\alpha(r) = 2 \sum_{r=1}^{+\infty} (-1)^{r-1}\ e^{-2r^2c^2}$

2.6. TESTS PARAMÉTRIQUES ET NON PARAMÉTRIQUES

- TESTS D'INDÉPENDANCE ET D'ASSOCIATION
- TEST D'INDÉPENDANCE DU KHI DEUX CF. TP

III. INFÉRENCE, INFORMATION, CLASSIFICATION

1. RETOUR SUR LE CONCEPT D'INFORMATION
2. L'INFÉRENCE BAYÉSIENNE
3. DONNÉES ET INFORMATION; REPRÉSENTATION DES DONNÉES
4. CLASSIFICATION SUPERVISÉE ET NON SUPERVISÉE
5. ACP, ACI, ET ANALYSE FACTORIELLE, …

3.1. RETOUR SU LE CONCEPT D'INFORMATION

- INFORMATION CONTINUE ET DISCRÈTE (CODAGE)
- INFORMATION ET REDONDANCE; ENTROPIE
- ESPÉRANCE ET MESURE; INÉGALITÉ DE JENSEN, …
- INFORMATION MUTUELLE ET INFORMATION DE FISHER
- BORNES EN ESTIMATION
- VARIATIONS D'INFORMATION ET RÉDUCTION DE L'INCERTITUDES

3.2. L'INFÉRENCE BAYÉSIENNE

- DONNÉES (OBSERVATIONS) : $X_1, X_2, \ldots, X_N$; VARIABLES : $V_1, V_2, \ldots,$
- MODELE : VECTEURS ALÉATOIRES $X = (X_1, \ldots, X_N)$, $V = (V_1, V_2 \ldots$
- LOIS DE PROBABILITÉ : $F(V)$, $F(X/V)$, $F(V/X)$
- ESTIMATEURS : E (V/X), MAX F (V/X),
  - ✓ LE PLUS SOUVENT ON N'ÉTUDIE PAS TOUTES LES VARIABLES
  - ✓ UNE PARTIE DES DONNÉES PEUT ÊTRE IGNORÉE
- VARIABLES ET PARAMÈTRES DE NATURE TRÈS VARIÉE
  - ✓ EXEMPLE : TÉMOIGNAGES, INDICES ET PREUVES (ENQUÊTE)
  - ✓ INFORMATIONS SUR LE WEB (VRAIES OU FAUSSES)

3.3. DONNÉES ET INFORMATION

- DONNÉES INDÉPENDANTES OU CORRÉLÉES
- INFORMATIONS INTERNET : NATURE, FORME, QUANTITÉ ET DIMENSION VARIABLES ; BIG DATA
- TRI DES DONNÉES/ INFORMATIONS PERTINENTES (CRITÈRES SUBJECTIFS OU OBJECTIFS – DISPERSION, PRÉSENCE DE VALEURS ABERRANTES)
- OBJECTIF DES ANALYSES DE DONNÉES : FAIBLE RÉDUCTION DE L'INFORMATION DU POINT DE VUE UTILISATEUR
- CRITÈRES DE DÉCISION : FONDÉS SUR L'INFORMATION PRÉSENTE
- DÉCISIONS AUTOMATIQUES OU HUMAINES (TRACÉS ET GRAPHIQUES)

3.4. CLASSIFICATION SUPERVISÉE/ NON SUPERVISÉE

- QUALITATIVE/ QUANTITATIVE
- CRITÈRES DE DISTANCE/ SIMILARITÉ (INDIVIDUS)
- CAS SUPERVISÉ : NATURE ET NOMBRE DE CLASSES CONNU.
  - PROBLÈMES : APPRENTISSAGE (MACHINE LEARNING)
  - IDENTIFIER LES RELATIONS INDIVIDU ⇗ CLASSE
  - CRITÈRES : PROBABILISTES/STATISTIQUES
  - AUTRE APPROCHE : RÉSEAUX DE NEURONES
- CAS NON SUPERVISÉ : AGRÉGATION DES INDIVIDUS EN CLASSES
  - CRITÈRES DE DISTANCE
  - APPROCHE PAR MOYENNES MOBILES
  - APPROCHE HIÉRARCHIQUE (ADAPTATIVE)

3.5. ACP, ACI ET ANALYSE FACTORIELLE

- ACP : ANALYSE DES MATRICES DE DONNÉES ET DE VARIANCE/COVARIANCE ; ETUDE SPECTRALE (VALEURS PROPRES = COMPOSANTES PRINCIPALES + AXES PRINCIPAUX)
- ACI : ANALYSE EN COMPOSANTES INDÉPENDANTES / SÉPARATION ET IDENTIFICATION DE SOURCES
- ANALYSES FACTORIELLES : SUR TABLEAUX DE DISTANCES (DISSIMILARITÉS, MDS), OU NON LINÉAIRES

IV. DONNÉES, TABLEAUX ET MATRICES

1. CALCUL MATRICIEL ET NOTATIONS
2. MATRICES STRUCTURÉES; INVERSION
3. ETUDE SPECTRALE: VALEURS PROPRES ET VECTEURS PROPRES
4. TABLEAUX DE DONNÉES
5. MATRICES DE VARIANCE/COVARIANCE

4.1 CALCUL MATRICIEL ET NOTATIONS

- Matrice $A \in R^n \times R^p$ = tableau à 2 dimensions (double entrée)
- *Notations :* $A = (a_{ij}) = [c_1\ c_2\ \dots\ c_p] = \begin{bmatrix} l_1 \\ \dots \\ l_n \end{bmatrix}$; $A^t = [l_1\ l_2\ \dots\ l_n]$

  rang (A) = rang $(c_1\ c_2\ \dots\ c_p)$ = rang $(l_1\ l_2\ \dots\ l_n)$ = rang (A^t)
- Cas complexe : $A \in C^n \times C^p$; $A^* = \overline{A^t}$
- Opérations : addition (n, n) et produit (n, p) & (p, q)
- Produit de matrices blocs/sous indices : $(AB)_I^K = \sum A_I^J\ B_J^K$; exemples
- Inverse : $AB = I$; $B = A^{-1} = LU = QR$;
- Cas de matrices structurées : Symétriques, Toeplitz
- Pseudo inverse : $B = (A^*A)^{-1}\ A^*$ si rang (A) = max (n, p)

  $ABA = A$; $BAB = B$; $(AB)^* = AB$; $(BA)^* = BA$;

4.2 MATRICES STRUCTURÉES; INVERSION

- Matrice symétrique ou hermitienne, positive : Factorisation de Cholesky
- Si A (n, n), $A = C^* C$, C triangulaire inférieure ou supérieure.
- Si A (n, p), $p > n$, $AX = b \Leftrightarrow A_I X^I = b - A_{\bar{I}} X^{\bar{I}}$, où A_I est inversible
- Inversion par élimination, ou factorisation $L\ U$, L et U triangulaires Inférieure/Supérieure (équivalent méthode du «pivot »)
- Matrice identité : $I = [e_1\ e_2 \ldots e_n]$, $B = A^{-1} = [c_1\ c_2 \ldots c_n]$, $A.c_k = e_k$

4.3 VALEURS PROPRES ET VECTEURS PROPRES

Dans $E = \mathbb{R}^n$

matrice A ou application linéaire f : $f(x) = \lambda x, x \neq 0,$ ou $Ax = \lambda x$

- $Ker\,(A - \lambda I) \neq \{0\}$, ou $det\,(A - \lambda I) = 0$, ou $\psi_A(\lambda) = 0$
- $\psi_A(X) = a_n X^n + \ldots + a_1 X + a_0$, avec $a_n = (-1)^n$, $a_0 = det\,(A)$

- A diagonalisable si E a une base de vecteurs propres, ou $E = \cup\, Ker\,(A - \lambda_k I)$

  ou $n = \sum n_k$, avec $n_k = \dim Ker\,(A - \lambda_k I)$
- Si A diagonalisable, et Q la matrice de passage, $D = Q^{-1} A Q$ est diagonale

* Si A est symétrique (hermitienne), $A = U D U^t$, U orthogonale (unitaire)

Valeurs propres réelles. Si A positive, valeurs propres positives.

* Polynôme minimal : $\psi_A(A) = 0$; $Ann\,(A) = M \,.\, \mathcal{M}_n(\mathbb{R})$

Sur C, $\psi_A(X)$ est scindé, et diagonalisation suivant les racines de ψ_A

4.4 TABLEAUX DE DONNÉES

Matrices de données : $A = (a_{ij})$ (n, p), p variables et n individus.

matrice $A = [a^1\ a^2\ \dots\ a^p]$, $A^t = [a_1\ a_2\ \dots\ a_n]$

Poids : Matrice D (diagonale); matrice pondérée $C = DA$.

Cas classique : $D = \frac{1}{n} I_n$ et $C = \frac{A}{n}$

- Moyennes : vecteur μ , tel que $\mu^t = [\mu_1\ \mu_2\ \dots\ \mu_p]$, $\mu_k = \frac{a_{1k} + a_{2k} \dots + a_{nk}}{n}$

  $\mu = A^t D\ \mathbf{1}$ avec $\mathbf{1}^t = [1\ 1\ \dots\ 1]$
- Matrice centrée : $B = (b_{ij}) = A - \mathbf{1}\ \mu^t$ $(b_{ij} = a_{ij} - \mu_j)$
- $B = A - \mathbf{1}\ \mu^t = A - \mathbf{1}\ \mathbf{1}^t\ D\ A = (I_n - \mathbf{1}\ \mathbf{1}^t\ D)\ A$: $B = G\ A$
- *Exemple*

Individu	Taille(cm)	Poids (kg)	Pointure	Ceinture
Pierre	183	74	43	82
Jean	176	75	42,5	84
Marc	178	72	42	80

4.4 TABLEAUX DE DONNÉES. 2

$A = (a_{ij})$ $(3, 4)$, 4 variables et 3 individus.

matrice $A = [a^1\ a^2\ \dots\ a^4] = \begin{bmatrix} 183 & 74 & 43 & 82 \\ 176 & 75 & 42.5 & 84 \\ 178 & 72 & 42 & 80 \end{bmatrix}$, $A^t = \begin{bmatrix} 183 & 176 & 178 \\ 74 & 75 & 72 \\ 43 & 42.5 & 42 \\ 82 & 84 & 80 \end{bmatrix}$

Poids : Matrice $D = \frac{1}{3}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$; matrice pondérée $C = DA = \frac{A}{3}$

- Moyennes : $\mu^t = [\mu_1\ \mu_2\ \mu_3\ \mu_4] = [179, 73.67, 42.5, 82]$

  $\mu = A^t D\ \mathbf{1}$ avec $\mathbf{1}^t = [1\ 1\ 1]$

- Matrice centrée : $B = (b_{ij}) = A - \mathbf{1}\ \mu^t = \begin{bmatrix} 4 & 0.33 & 0.5 & 0 \\ -3 & 1.33 & 0 & 2 \\ -1 & -1.67 & -0.5 & -2 \end{bmatrix}$

- $B = A - \mathbf{1}\ \mu^t = A - \mathbf{1}\ \mathbf{1}^t\ D\ A = (I_n - \mathbf{1}\ \mathbf{1}^t\ D)\ A$: $B = G\ A$

- $G = I_n - \mathbf{1}\ \mathbf{1}^t\ D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

4.5 MATRICES DE VARIANCE/COVARIANCE

Matrice de données A (n, p), $A^t = [a_1\ a_2\ \dots\ a_n]$, $\mu^t = [\mu_1\ \mu_2\ \dots\ \mu_p]$

(a_k^t est la k-ème ligne de A)

$$A^t A = [a_1\ a_2\ \dots\ a_n] \cdot \begin{bmatrix} a_1^t \\ \dots \\ a_n^t \end{bmatrix} = a_1\, a_1^t + a_2\, a_2^t + \dots + a_n\, a_n^t$$

Pour une matrice de pondération $D = \begin{bmatrix} p_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n \end{bmatrix}$, avec $p_1 + \dots + p_n = 1$

$$A^t D A = p_1\, a_1\, a_1^t + p_2\, a_2\, a_2^t + \dots + p_n\, a_n\, a_n^t$$

La matrice de (variance/) covariance de A s'écrit : $V = B^t D B = A^t D A - \mu\, \mu^t$

(matrice carrée d'ordre p)

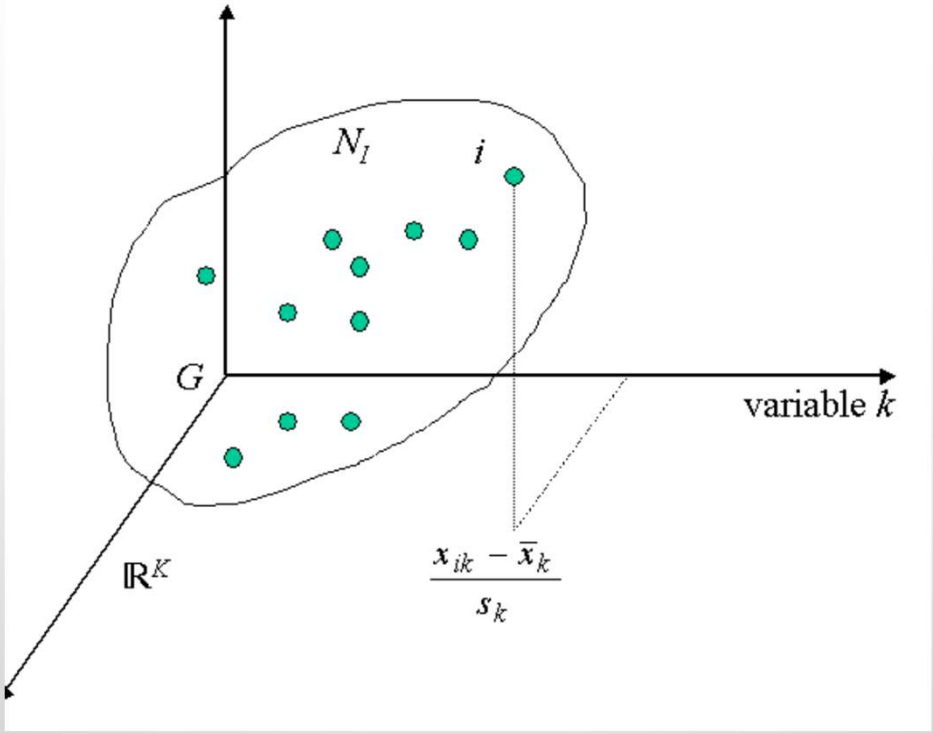
Avec $B = (I_n - \mathbf{1}\ \mathbf{1}^t D) A = G A = A - \mathbf{1}\, \mu^t$ (et $\mu = A^t D\, \mathbf{1}$)

V. L'ANALYSE EN COMPOSANTES PRINCIPALES

1. TABLEAUX DE DONNÉES ET MATRICES DE COVARIANCE
2. SOUS ESPACES SUPPLÉMENTAIRES ET PROJECTEURS
3. DÉCOMPOSITIONS SPECTRALES ET FACTORISATIONS
4. ACP. APPROCHE ALGÉBRIQUE : COMPOSANTES PRINCIPALES
5. APPROCHE GÉOMÉTRIQUE

V. L'ANALYSE EN COMPOSANTES PRINCIPALES

1.



5.1 TABLEAUX DE DONNÉES ET MATRICES

Matrice de données A (n, p), $A^t = [a_1\ a_2\ \dots\ a_n]$, $\mu^t = [\mu_1\ \mu_2\ \dots\ \mu_p]$

n individus (données uniformes), et p variables corrélées

Objectif : réduction du nombre de variables , à fin de visualisation, de détection ou de classification (individus/variables), et de détection d'artefacts.

Travail sur la matrice de variance-covariance

Pour une matrice de pondération $D = \begin{bmatrix} p_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n \end{bmatrix}$, avec $p_1 + \dots + p_n = 1$

$V = B^t D B = A^t D A - \mu \mu^t = A^t D A$ si on suppose (fait ci-dessous) que $\mu = 0$

Avec $B = (I_n - \mathbf{1}\ \mathbf{1}^t D) A = G A = A - \mathbf{1}\ \mu^t$ (et $\mu = A^t D\ \mathbf{1}$)

Les variables sont transformées en facteurs, pour un nouveau tableau.

On définit l'inertie du tableau A : $I = \frac{1}{n} \sum_{i=1}^{n} a_i\ a_i^t = \frac{1}{n} \sum_{i=1}^{n} \|a_i\|^2$

Si $p_i = \frac{1}{n}$ $V = \frac{1}{n} A^t A = \frac{1}{n} \sum_{i=1}^{n} a_i{}^t a_i$

5.2 ESPACES SUPPLÉMENTAIRES ET PROJECTEURS

Si $E = F + G$ et $E \cap F = \varnothing$, E et F sont supplémentaires.

Projecteurs sur E et F : $p = p_E$ et $q = p_F$

$pq = qp = 0$, $p + q = id$; $Im\ (p) = ker\ q = E$, $Im\ (q) = ker\ p = F$

p est un projecteur si et seulement si $p^2 = p$;

Si u et v sont des vecteurs de R^n , tels que $v^t\ u = 1$, et qu'on considère la matrice $P = u\ v^t$

P est la projection : sur la droite D de vecteur directeur u,

parallèlement à l'orthogonal G de $Im\ v$ (de dimension $n - 1$)

$P^2 = u\ v^t\ u\ v^t = u\ (v^t\ u)\ v^t = u\ v^t = P$

Si $x = \alpha\ u$, $Px = \alpha\ u = x$

Si $x \perp v$, $Px = u\ v^t\ x = u\ (v^t\ x) = 0$

Donc $Im\ P = D$ et $ker\ P = G$

5.2 ESPACES SUPPLÉMENTAIRES ET PROJECTEURS

Si $u = v$ et $u^t\, u = 1$ (donc si $||u|| = 1$), $P = u\, u^t$

Si $U = [u_1, u_2, \ldots, u_r]$ définit une base b de F $(b = \{u_1, u_2, \ldots, u_r\})$

où les vecteurs u_k sont des vecteurs colonnes, $U^t\, U = I_r$

On peut vérifier que projecteur orthogonal sur F est $P = \sum u_k\, u_k^t = U\, U^t$

Pour w dans E, on cherche z dans F tel que $w - z \perp F$, soit encore $U^t (w - z) = 0$,

donc $U^t\, w = U^t\, z$,

et si $z = U\, c$, $U^t\, z = c$, et $U\, U^t\, z = U\, c = z$, et si $t \perp F$, $U\, U^t\, t = 0$

La matrice de P est $A = U\, (U^t\, U)^{-1}\, U^t$

5.3 DÉCOMPOSITIONS SPECTRALES ET OPÉRATEURS

Si il existe une base de vecteurs propres de A, A est diagonalisable, et si on note D la matrice diagonale formée des valeurs propres, et Q la matrice de passage vers la base diagonale,
$A = Q\,D\,Q^{-1}$

Si A est symétrique, $A = Q\,D\,Q^{t}$

Si $Q = [u_1\ u_2\ \dots\ u_n]$, et $(Q^{-1})^t = [v_1\ v_2\ \dots\ v_n]$, $A = \sum_{i=1}^{n} \lambda_i\, u_i\, v_i^{t}$

Si A est rectangulaire, on étudie la matrice carrée $B = A\,A^t$, et les valeurs propres de B sont les valeurs singulières de A. rang (B) = rang (A), et les valeurs propres λ_i de B sont positives. Les valeurs singulières de A sont les valeurs $\mu_i = \sqrt{\lambda_i}$

Si rang $(A) = r$, D est la matrice diagonale des valeurs propres de B, et $S = \sqrt{D}$

$A = U\,S\,W^t$ et $A^t\,A = W\,S^2\,W^t = \sum_{i=1}^{r} \lambda_i\, v_i\, v_i^{t}$

avec $U = [u_1\ u_2\ \dots\ u_n]$ la matrice des vecteurs propres de B, et

$W = [v_1\ v_2\ \dots\ v_p]$ la matrice des vecteurs propres de $A^t\,A$

5.4 ACP; COMPOSANTES PRINCIPALES

On part de la matrice V, et on cherche les axes qui maximisent l'inertie :

On utilise une projection orthogonale sur un espace de dimension $q < p$

On peut observer le cas d'un espace de dimension 1. La projection orthogonale de x sur une droite D dirigée par u est définie par $P_u(x) = x^t u \frac{u}{\|u\|}$

On cherche l'axe v qui minimise la distance entre les éléments de R^p et leurs projections :

v minimise ($\sum_{i=1}^{n} \| a_i - P_u(a_i) \|^2$, $\| u \| = 1$)

Comme $P_u(a_i)$ et $a_i - P_u(a_i)$ sont orthogonaux, $\| a_i \|^2 = \| a_i - P_u(a_i) \|^2 + \| P_u(a_i) \|^2$

Et donc minimiser ($\sum_{i=1}^{n} \| a_i - P_u(a_i) \|^2$) = ($\sum_{i=1}^{n} [\| a_i \|^2 - \| P_u(a_i) \|^2]$) revient à

maximiser $\sum_{i=1}^{n} \| P_u(a_i) \|^2$, et si $\| u \| = 1$, $\| P_u(a_i) \|^2 = u^t a_i a_i^t u$

v maximise donc $u^t A^t A u = u^t V u$, sous la contrainte $\| u \| = 1$.

Une étude à l'aide des multiplicateurs de Lagrange conduit à l'équation : $V u = \lambda u$

Et le premier axe factoriel est l'espace propre associé à la plus grande valeur propre de V.

5.4 ACP; COMPOSANTES PRINCIPALES

En dimension $q < p$

On peut observer le cas d'un espace de dimension 1.

Si on a défini des axes factoriels $v^1, v^2, \ldots, v^m$, on cherche l'axe v^{m+1} orthogonal aux précédents et de norme 1 qui maximise $u^t A^t A u = u^t V u$.

De fait, l'espace de projection de dimension q (espace d'inertie maximale) est la somme directe des espaces propres associés aux q plus grandes valeurs propres de V.

Les axes factoriels sont associés à ces plus grandes valeurs propres

Exemple élémentaire : Si on définit la matrice de données 3×2 par $A^t = \begin{bmatrix} 2 & 3 & -1 \\ 0 & 5 & -2 \end{bmatrix}$;

$B^t = \begin{bmatrix} 1 & -1 & 0.5 \\ -1 & 1 & -0.5 \end{bmatrix}$, et $B^t B = \begin{bmatrix} 2.25 & -2.25 \\ -2.25 & 2.25 \end{bmatrix}$; Les valeurs propres sont 0 et 4.5

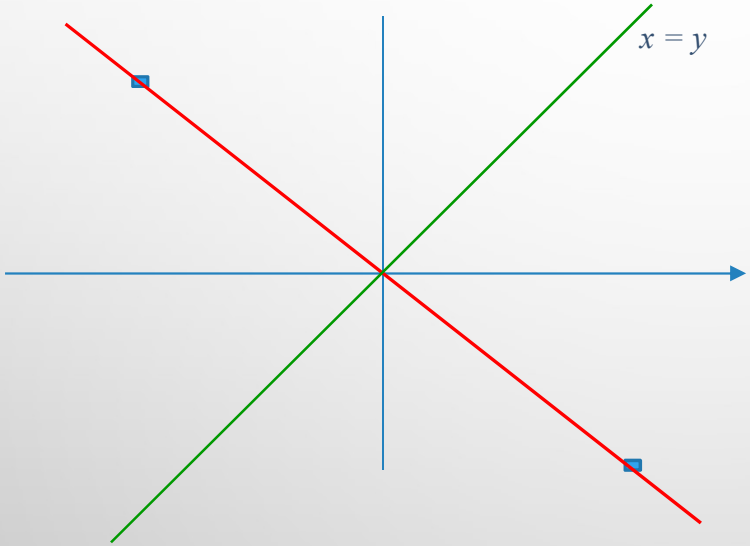
L'axe principal est associé à $\lambda = 4.5$, et a pour équation $x - y = 2x$, soit $x = -y$

5.4 ACP; COMPOSANTES PRINCIPALES

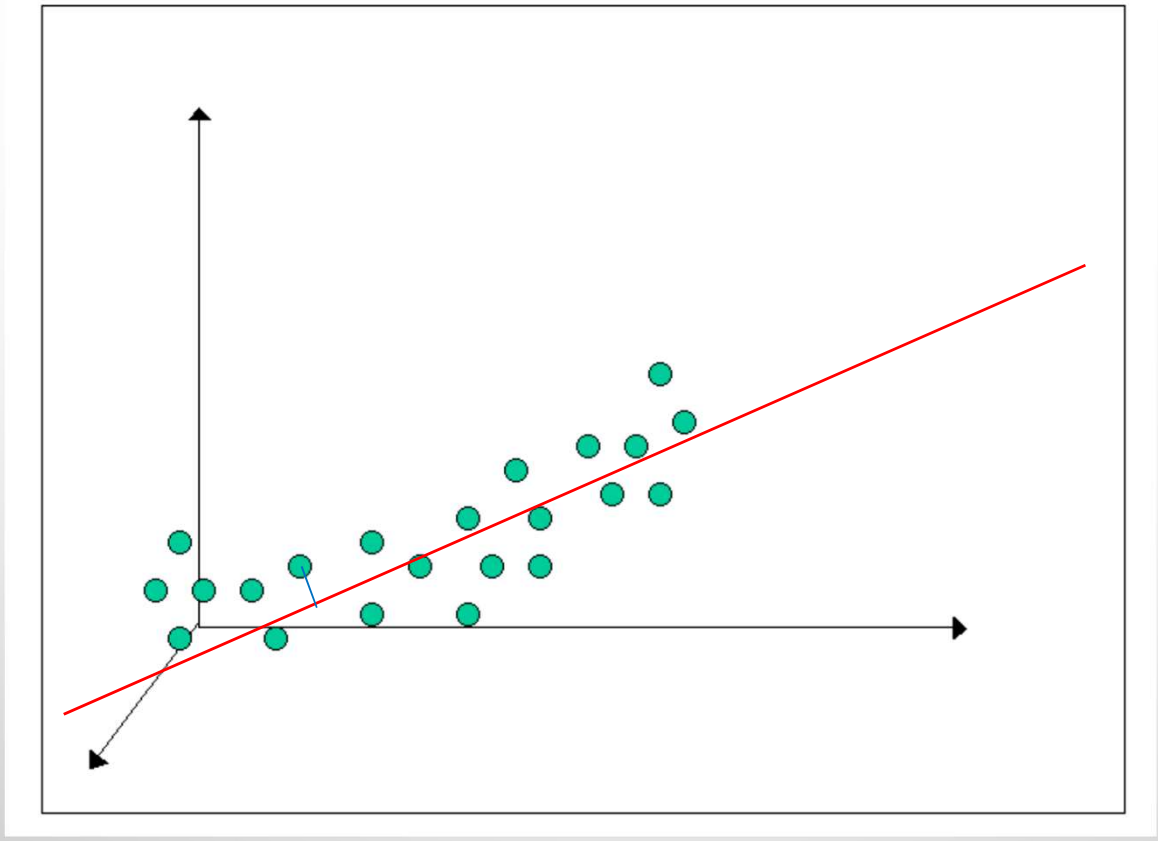
Exemple suite : L'axe principal est associé à $\lambda = 4.5$, et a pour équation $y = -x$

L'autre axe principal est le noyau de V, d'équation $x = y$

Le nuage (de 2 points) est sur l'axe principal, qui maximise la distance !

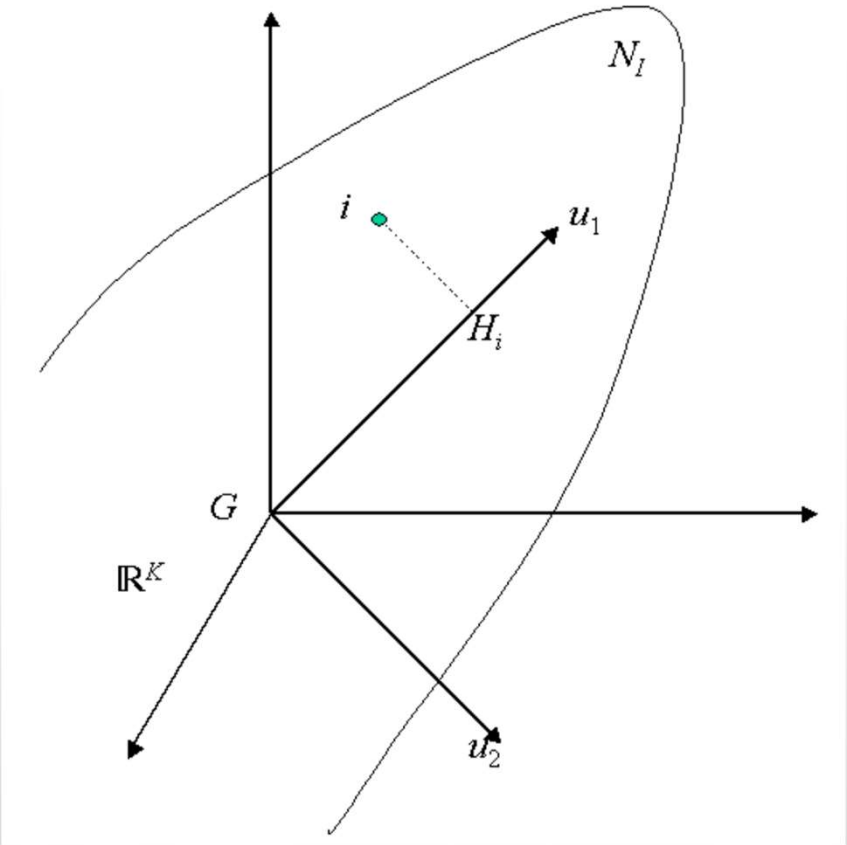


5.4 ACP; COMPOSANTES PRINCIPALES



Directions d'allongement

5.4 ACP; COMPOSANTES PRINCIPALES



Ajustement du nuage N_I des individus pour l'ACP.

V. L'ANALYSE DES CORRESPONDANCES

1. ANALYSE CANONIQUE
2. TABLEAUX DE CONTINGENCES
3. MÉTRIQUE DU KHI DEUX ET ACP
4. ANALYSE CANONIQUE QUALITATIVE
5. CORRESPONDANCES MULTIPLES
6. EXEMPLES

5.1 ANALYSE CANONIQUE

CAS DE DEUX GROUPES DE VARIABLES

TABLEAU $X = [X_1, X_2]; X_1 : n \times p ; X_2 : n \times q$

Les sous espaces de R^n :

E_1 engendré par les colonnes de X_1

E_2 engendré par les colonnes de X_2

$x \in E_k$ si $x = X_k \ a, a \in R^p \ (k = 1)$ ou $a \in R^q \ (k = 2)$

Si $E_1 = E_2$, les variables sont redondantes

Si E_1 et E_2 sont orthogonaux, les phénomènes décrits par les variables de X_1 et X_2 sont indépendants

On cherche les liens entre ces groupes de variables ou de caractères

5.1 ANALYSE CANONIQUE

RECHERCHE DE VARIABLES CANONIQUES

On considère une métrique D dans R^n :

1. On cherche des vecteurs normés $x_1 \in E_1$ et $y_1 \in E_2$ dont le produit scalaire est maximal en valeur absolue (revient à minimiser l'angle formé par les vecteurs)

2. On cherche des vecteurs normés $x_2 \in E_1$ et $y_2 \in E_2$, tels que x_1 et x_2, respectivement y_1 et y_2, sont orthogonaux :

$$< x_1 , x_2> = x_1^t \, D \, x_2 = 0.$$

$$< x_1 , x_2> = x_1^t \, D \, x_2 = 0.$$

3. On construit deux suites de m vecteurs normés (m = min (p, q)), chaque suite formant une famille orthogonale.

Si P_1 et P_2 sont les matrices des opérateurs de projection orthogonale sur E_1 et E_2

$$P_1 = X_1 \, (X_1^t \, D \, X_1)^{-1} \, X_1^t \, D$$

$$P_2 = X_2 \, (X_2^t \, D \, X_2)^{-1} \, X_2^t \, D$$

5.1 ANALYSE CANONIQUE

RECHERCHE DE VARIABLES CANONIQUES

1. On cherche des vecteurs normés $x_1 \in E_1$ et $y_1 \in E_2$ dont le produit scalaire est maximal en valeur absolue (minimise l'angle formé par les vecteurs) :

obtenu si la projection $y_1' = A_2\, x_1$, de x_1 sur E_2 , est colinéaire à y_1 , et si

$x_1' = A_1\, y_1$, de y_1 sur E_1 , est colinéaire à x_1 , donc

$$y_1' = A_2\, x_1 = \lambda_2\, y_1 \text{ , et } x_1' = A_1\, y_1 = \lambda_1\, x_1 \text{ , d'où}$$

$$A_2\, A_1\, y_1 = \lambda_1\, A_2\, x_1 = \lambda_1\, \lambda_2\, y_1$$

$$A_1\, A_2\, x_1 = \lambda_2\, A_1\, y_1 = \lambda_2\, \lambda_1\, x_1$$

On cherche donc les vecteurs propres des matrices $A_2\, A_1$ et $A_1\, A_2$

Remarque : $A_1\, A_2$ et $A_2\, A_1$ sont diagonalisables. Les valeurs propres sont réelles positives.

5.1 ANALYSE CANONIQUE

RECHERCHE DE VARIABLES CANONIQUES

2. On cherche des vecteurs normés $x_k \in E_1$ et $y_k \in E_2$ pour deux suites orthogonales, dont le produit scalaire est maximal en valeur absolue. On obtient des relations analogues :

$$y_k' = A_2\, x_k = \lambda_{k,1}\, y_k, \text{ et } x_k' = A_1\, y_k = \lambda_{k,2}\, x_k$$

$$A_2\, A_1\, y_k = \lambda_{k,1}\, A_2\, x_k = \lambda_{k,1}\, \lambda_{k,2}\, y_k = \mu_k\, y_k$$

$$A_1\, A_2\, x_k = \lambda_{k,2}\, A_1\, y_k = \lambda_{k,2}\, \lambda_{k,1}\, x_k = \mu_k\, x_k$$

$$x_k'\, D\, x_k = 1 \text{ et } y_k'\, D\, y_k = 1$$

$$x_i'\, D\, x_j = 0 \text{ et } y_i'\, D\, y_j = 0,\ 1 \le i \ne j \le m$$

$$y_i'\, D\, x_j = 0 \text{ et } x_i'\, D\, y_j = 0,\ 1 \le i \ne j \le m$$

$A_1\, A_2$ et $A_2\, A_1$ sont diagonalisables. Les vecteurs propres sont orthogonaux.

5.1 ANALYSE CANONIQUE

RECHERCHE DE VARIABLES CANONIQUES

Ecriture des vecteurs x_k et y_k :

$$x_k = X_1\, a_k \quad \text{et} \quad y_k = X_2\, b_k$$

Les vecteurs a_k et b_k sont les facteurs canoniques. On pourra les calculer directement en diagonalisant des matrices d'ordre p et q. Si n est grand devant ces valeurs, ce sera beaucoup plus simple que de diagonaliser les matrices d'ordre n, $A_2\, A_1$ et $A_1\, A_2$

On peut les évaluer, en notant pour $r, s \in \{1,2\}$, $V_{r,s} = X'_r\, D\, X_s$

Si $X'_1\, D\, 1 = 0$ et $X'_2\, D\, 1 = 0$, les matrices $V_{r,s}$ sont des matrices de covariance.

$$V_{11}^{-1}\, V_{12}\, V_{22}^{-1}\, V_{21}\, a_k = \lambda_k\, a_k$$

$$V_{22}^{-1}\, V_{21}\, V_{11}^{-1}\, V_{12}\, b_k = \lambda_k\, b_k$$

5.2 TABLEAUX DE CONTINGENCES

Tableau de contingence, pour deux variables qualitatives

Si deux variables peuvent prendre respectivement r et s valeurs, on écrit les fréquences n_{ij} de ces valeurs dans un tableau $\boldsymbol{N}$ à r lignes et s colonnes.

On note $n_{i\bullet}$ et $n_{\bullet j}$ les sommes des éléments par colonnes et par lignes. Les sommes de ces valeurs sont aussi égales toutes deux à l'effectif total, n.

Ces valeurs $n_{i\bullet}$ et $n_{\bullet j}$ sont appelées marges, et on peut normaliser les valeurs sur les lignes et sur les colonne du tableau. On obtient deux nouveaux tableaux : un tableau **profils lignes**, $\boldsymbol{N_L}$, et un tableau **profils colonnes** $\boldsymbol{N_C}$.

On peut caractériser l'indépendance des variables : $n_{ij} = \frac{1}{n} n_{i\bullet} n_{\bullet j}$

Les lignes du tableau profils colonnes. Et les colonnes du tableau profils lignes sont alors identiques

5.2 TABLEAUX DE CONTINGENCES

Tableau de contingence, pour deux variables qualitatives

Si on définit les matrices diagonales D_1 et D_2, d'ordre r et s, dont les éléments diagonaux sont les éléments $n_{i.}$ pour D_1 et $n_{.j}$ pour D_2, les tableaux de profils s'écrivent :

$N_L = D_1^{-1} N$ et $N_C = N D_2^{-1}$

Les lignes de la matrice N_L sont des éléments de $\mathbb{R}^s$, et elles forment un nuage de r points.

Son centre de gravité est : $g_L = \frac{1}{n} (D_1^{-1} N)^t D_1 1 = \begin{bmatrix} \frac{n_{.1}}{n} \\ \frac{n_{.1}}{n} \\ \vdots \\ . \end{bmatrix} = \begin{bmatrix} p_{.1} \\ p_{.2} \\ \vdots \\ . \\ p_{.s} \end{bmatrix}$

Et de manière analogue, les colonnes de N_C forment un nuage de s points, et on a $g_C = \begin{bmatrix} p_{1.} \\ p_{2.} \\ \vdots \\ . \\ p_{r.} \end{bmatrix}$

En cas d'indépendance, les nuages sont réduits à leurs centres de gravité.

5.3 MÉTRIQUE DU KHI DEUX ET ACP

Métrique du khi deux :

Distance entre 2 profils lignes : $D(i, i') = d^2(i, i') = \sum_{j=1}^{s} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2$

C'est une métrique diagonale qui peut s'écrire D_2^{-1}

On peut également définir cette métrique pour les colonnes

V. CLASSIFICATION NON SUPERVISÉE

1. CLASSIFICATION, MÉTRIQUES
2. CLASSIFICATION HIÉRARCHIQUE
3. APPROCHES DE CENTRES MOBILES
4. MODÈLES PARAMÉTRIQUES : MÉLANGES GAUSSIENS
5. EXEMPLES
6. UNE CLASSIFICATION SUPERVISÉE ?

V. CLASSIFICATION NON SUPERVISÉE

1. CLASSIFICATION, MÉTRIQUES

a) La classification non supervisée est fondée sur des critères de ‘distances’ ou de métriques entre les points et entre les classes,

b) La classification non supervisée est fondée sur des notions d’inertie intra classe et inter classes.

Les méthodes sont fondées sur des règles et des critères d’agrégation de points et de classes (méthodes constructives), ou sur une minimisation d’une fonction globale.

Deux méthodes principales :

Classification hiérarchique : chaque point définit une classe, et à chaque étape on regroupe les classes les plus ‘proches’. Si on itère, on n’obtient à la fin qu’une seule classe (question du nombre de classes ?)

V. CLASSIFICATION NON SUPERVISÉE

1. CLASSIFICATION, MÉTRIQUES

Deux méthodes principales :

2. Classification *k*-means (moyennes mobiles, nuées dynamiques) : On crée *k* classes, à partir de *k* points, et on affecte successivement chaque point à une classe.

Définition des métriques entre points : distances, similarités, dissimilarités

Définition des métriques entre classes : métriques entre centres de gravité, métriques max et min, pondération ou poids des classes

Inertie intra-classe et inter-classes : critères de qualité de la classification.

Utilisées pour déterminer le nombre de classes (hiérarchique) ou caractériser le choix de *k* classes.

V. CLASSIFICATION NON SUPERVISÉE

1. CLASSIFICATION, MÉTRIQUES

Similarité : sur un ensemble E

Application de $E \times E \rightarrow R+$ vérifiant

i) $s(x, y) = s(y, x)$, pour tout couple (x, y) de $E \times E$

ii) $s(x, x) = S > 0$, pour tout élément x de E

iii) $s(x, y) \leq S$, pour tout couple (x, y) de $E \times E$

Exemples : 1. constante S positive

2. $s(x, y) = \min(|x|, |y|) + C$, pour $E \subset R$, $C = 0 \notin E$ ou $C > 0$

3. $s(x, y) = 2 \frac{|x\,y|}{x^2 + y^2} + C$, pour $E \subset R$, $C = 0 \notin E$ ou $C > 0$

V. CLASSIFICATION NON SUPERVISÉE

1. CLASSIFICATION, MÉTRIQUES

Dissimilarité : sur un ensemble E

Application de $E \times E \rightarrow R+$ vérifiant

i) $s(x, y) = s(y, x)$, pour tout couple (x, y) de $E \times E$

ii) $s(x, y) = 0 \Leftrightarrow x = y$, pour tout couple (x, y) de $E \times E$

Exemples : 1. $s(x, y) = |x - y|$, pour $E \subset R$

2. $s(x, y) = \big|\,|x| - |y|\,\big|$, pour $E \subset R$

3. $s(x, y) = 1$ si $x \neq y$, et $s(x, x) = 0$, pour E quelconque

V. CLASSIFICATION NON SUPERVISÉE

1. CLASSIFICATION, MÉTRIQUES

Distance : sur un ensemble E

Application de $E \times E \rightarrow R+$ vérifiant

i) $d(x, y) = d(y, x)$, pour tout couple (x, y) de $E \times E$

ii) $d(x, y) = 0 \Leftrightarrow x = y$, pour tout couple (x, y) de $E \times E$

iii) $d(x, z) \leq d(x, y) + d(y, z)$

Exemples : 1. distance euclidienne

2. distance de Hamming

3. distance du Khi deux

V. CLASSIFICATION NON SUPERVISÉE

2. CLASSIFICATION HIÉRARCHIQUE ASCENDANTE

Pour un nuage de *n* points : On constitue *n* classes élémentaires (singletons)

On définit une métrique interclasses fondée sur :

Des caractéristiques quantitatives ou qualitatives des points,

L'effectif des classes,

Des modèles de classes.

A chaque étape:

On regroupe les classes les plus 'proches'

On crée une nouvelle classe, dont on détermine les caractéristiques

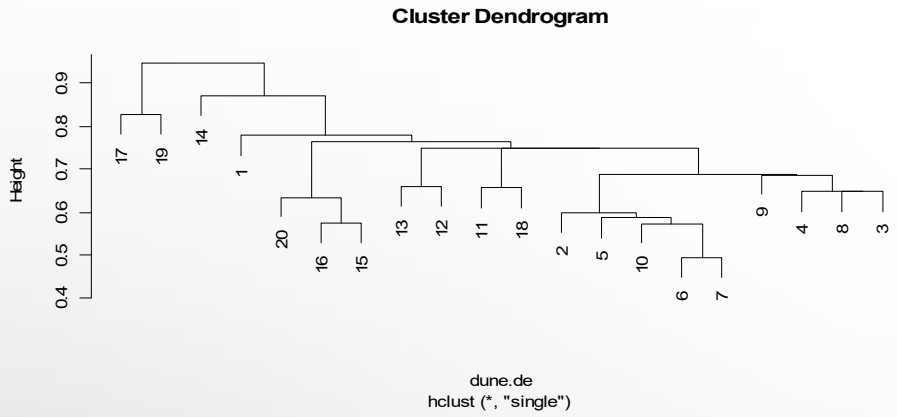
V. CLASSIFICATION NON SUPERVISÉE

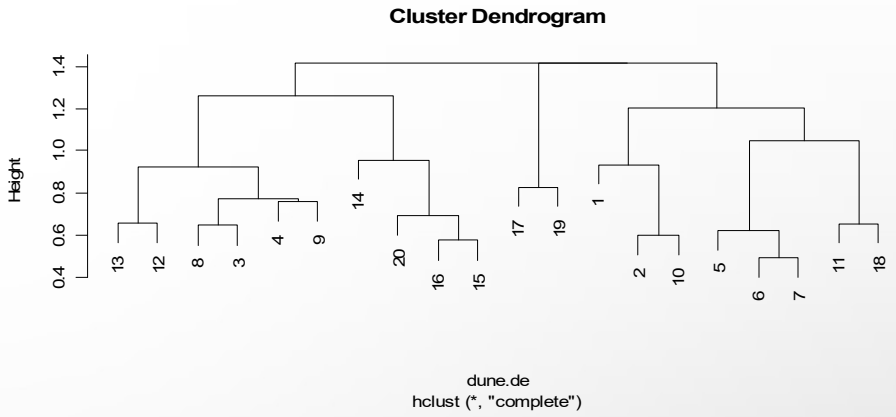
2. CLASSIFICATION HIÉRARCHIQUE ASCENDANTE

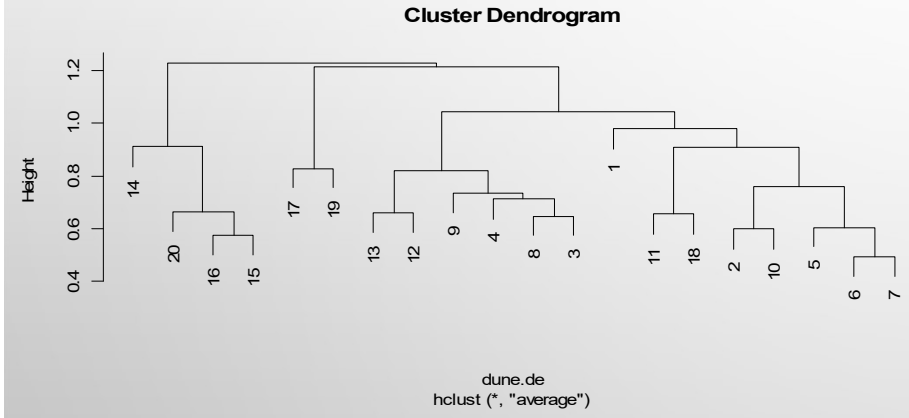
On calcule une inertie interclasses normalisée et pondérée.

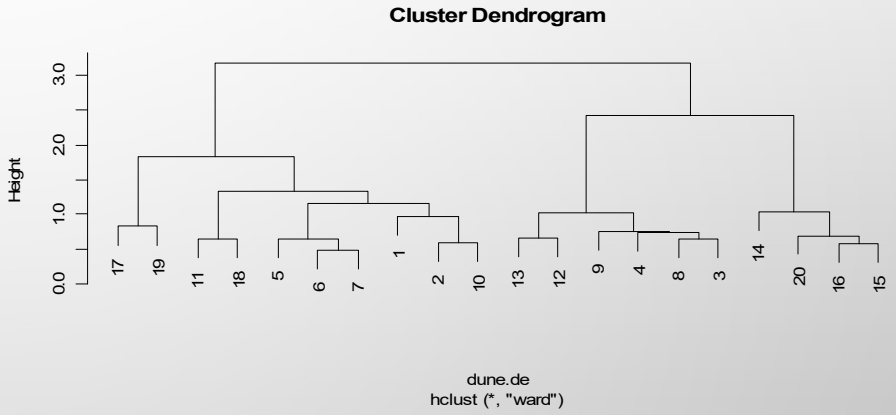
Si cette inertie dépasse un seuil donné, on arrête l'agrégation.

III. MÉTHODES DE CLASSIFICATION









Graphique 1=> mettre en évidence les gradients

Graphique 2=> bien séparer les groupes