Sommaire

19/09/2023

La prévalance d'une maladie est la probabilité qu'un individu pris au hasard dans une population, soit malade.

Deux types de statistiques descriptives :

- La statitistique univarié : on gère variable par variable
- La statistique bivarié constiste à prendre les variables deux à deux dans le but de déterminer le lien entre les deux.
- La statistique multivariée : on s'intéresse à plus de deux variables à la fois. Ici, on va voir quelques méthodes de réduction de variables dans le but de représenter nos informations dans le plan.

Dans la classification supervisée, on cherche à expliquer une variable qui est quantitative par un ensemble de variables mais dans l'autre(la classification , on ve ut créer des groupes …

On ne va pas aborder la collecte et le netoyage des données dans le cadre de ce cours.

20/09/23

Introduction et rappels

Introduction

L'analyse des données soustend l'idée de valoriser les données dont on dispose en y extrayant des informations utiles.

Les méthoes utiliseées peuvent $etre classé comme suit :

Les méthodes exploratoires

- **Les méthodes de description des données**
- **Les méthodes factorielles (reduction des dimensions de representation des données)**

Les méthodes inférentielles (modèles statisituqes)

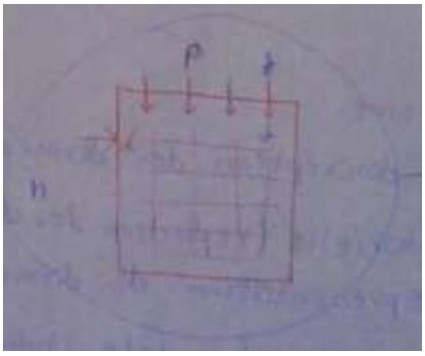
Les méthodes de prédiction

En fonction de la nature des données à analyser, ces méthoes d'analyse des données peuvent être regroupées en trois classes principales

1) **La classe des méthodes non supervisées**

elle vise à resumer les données sans privilégier une des variables qui se presente sous la forme **d'entrées (sans sortie)**

le jeu de données étant de la forme $X = \left(x_{ij}\right)_{1\leq i\leq n, 1\leq j\leq p} = (x_i)_{i=1}^{n}$ où $x_i = \left(x_{ij}\right)_{j=1}^{p}$



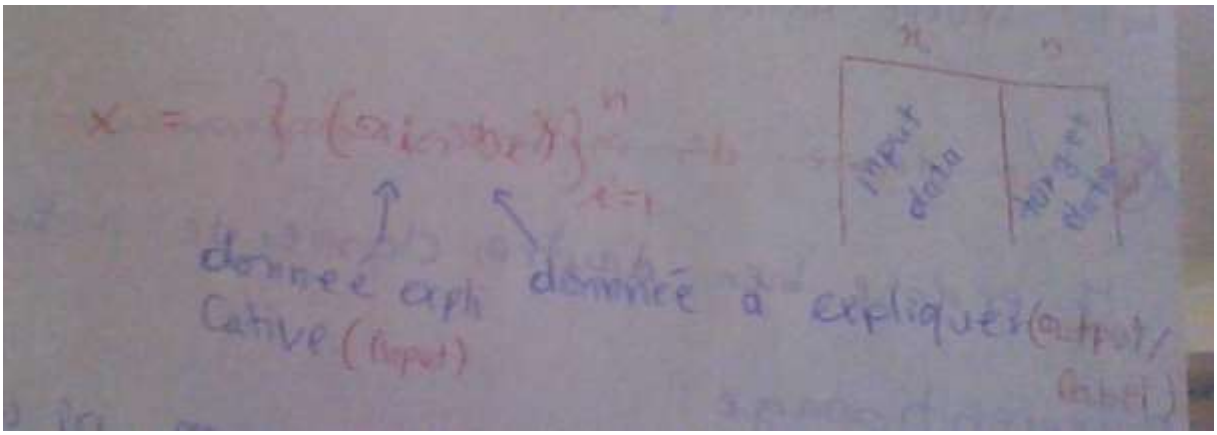
shape of the data set

2) **La classe des méthodes supervisées**

Elles sont applicables sur les données dans lesquelles on priviligie une ou plusieurs varialbes à expliquer, les autres étant des variables eplicatives.

Les données sont alors dsous la forme

$$X = \{(x_i, y_i)\}_{i=1}^{n}$$



Ici on peut chercher à comprendre le lien entre les deux tyypes de données pour en faire des prédictions pour les prochaines entrées, ou bien juste pour trouver la qualité du lien entre x et y.

Ces méthodes servent d'analyser les/le lien/s

3) La classe des méthodes d'apperntissage par renforcement

Ici généralemenet les données changent au cours du temps et cette classe de méthodes est plus utilisée en robotique.

Données : $(x_t, a_t), (x_{t+1}, a_{t+1})$ **t étant le temps**

(x_t, a_t) **est l'état du système à t**

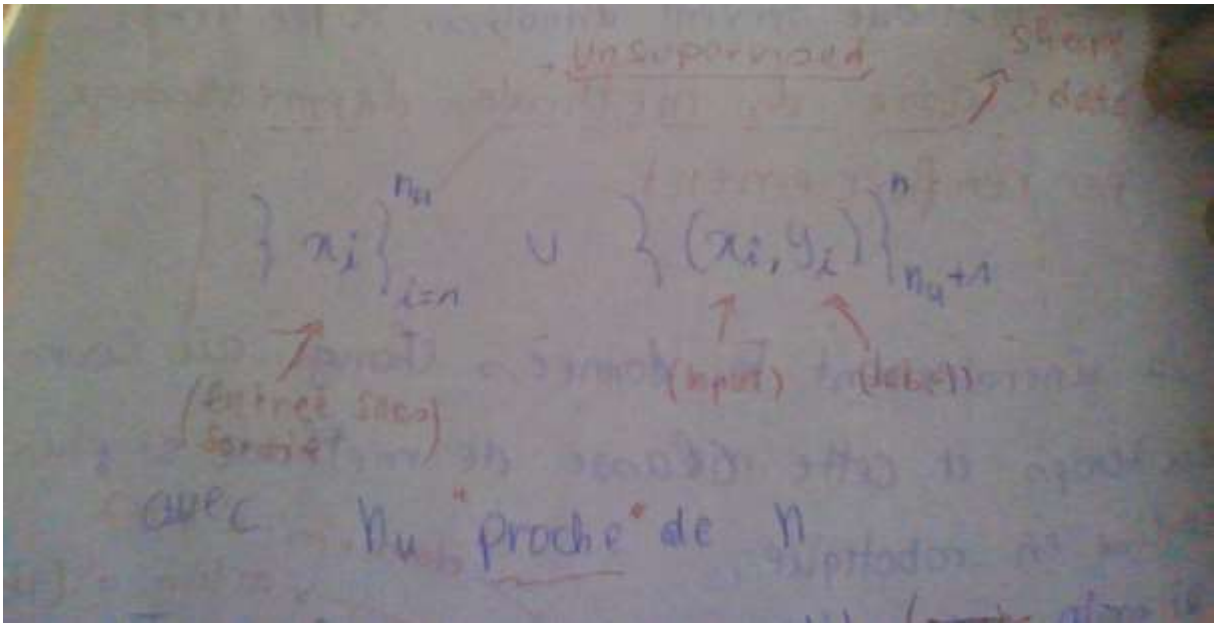
$$(x_{t+1}, a_{t+1}) = f(x_t, a_t)$$

But : quelle action pour un état optimal ?

Il existe bien d'autres classes de méthodes d'apprentissage :

- **Méthodes d'apprentissage semi-supervisé**

Ici, il y a majoritairement des données à entrée sans sortie et une petite partie pour des données avec sortie.



Toutes fois si n_u est petit, alors il serait mieux d'appliquer les données supervisées aux données non labelisés et utiliser un model supervisé.

Le but du semi-supervisé est de coller les labels aux données non labelisées.

$(x_i)_{i=1}^n$	- Segmentation des clients d'une téléphonie mobile - Segmentation des patrimoines génétiques dans une ville

$(x_i, y_i)_{i=1}^n$	- Reconnaissance de forme
$(\{x_i\} U \{x_i, y_i\})$	

Prochain cours :

- **Vecteur et valeur propre**
- **Projection orthogonale dans un espace vectoriel**
- **Covariance empirique, corrélation empirique**
- **Base orthogonale**
- **Décomposition SVD d'une matrice**

Notes :

Une variable aléatoire est donnée par

$X: (\Omega, A, P) \to (E, \epsilon), \forall B \in \epsilon, X^{-1}(B) \in A$

La loi de X est donnée par

$$P_X: \epsilon \to R, B \to P_x(B) = P(X^{-1}(B))$$

Fonction de répartition théorique

$$F_X(x) = P^{¨}(X \leq x)$$

Fonction de repartition empirique

Pour un ensemble de données $(x_1, .., x_n) = x_{\_}$

On a $\widetilde{F_x} = \frac{1}{n}\sum_{i=1}^n 1_{x_i \leq x}$

Déplacement d'un nuage de ponts dans le temps

C'est généralement le jeu de données qui nous guide vers le choix des méthodes à utiliser, dans certains jeux de données, on eput souvent être ammené à utiliser pluseireurs méthodes.

26/09/23

Chapitre 1 : ACP

Dans ce cours, on suppose que toutes nos variables sont quantitatives

données : $\boldsymbol{X = (x_{ij})_{i=1,\dots,n, j=1,\dots,p} \in R^{n \times p}}$ avec $\boldsymbol{p \geq 2}$, p potentiellement très grand.

si p=1, on peut placer les données sur une droite

si p=2, on peut placer ces données dans un plan muni d'un repère et se rendre compte de la dispersion des points et de la forme des nuages.

Lorsque p est plus grand que 4, on ne peut plus les representer à moins de reduire la dimension de représentation des données en perdant le moins d'informations possibles dans le nuage déformé.

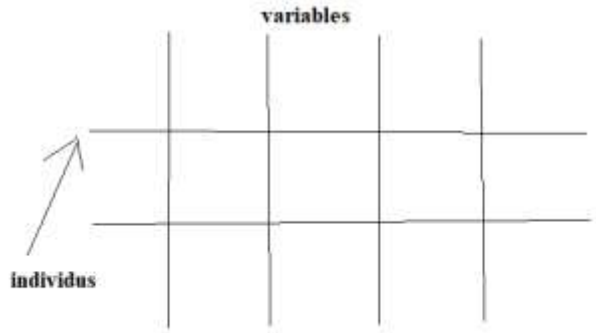
I. Statistiques descriptives

1) Espace des individus

Données d'un individu

$$x_{i,.} = \left(x_{i,j}\right)_{j=1}^{p} \in R^p$$

L'espace des individus est R^p



$\beta = \left(e_j\right)_{j=1}^{p}$ base canonique de R^p

On munit l'espace des données R^p d'une métrique M.

M est étant une matrice symétrique définie positive.

$$d^2\left(x_{i,.}, x_{i'_{,.}}\right) = {}^t\left(x_{i,.} - x_{i'_{,.}}\right)M\left(x_{i,.} - x_{i'_{,.}}\right)$$

Le rôle de la métrique est de ramener les variables au même ordre de grandeur.

Métrique usuelle

$$M = diag\left(\frac{1}{S_j^2}\right)_{j=1}^{p}$$

Où S_j est l'écart type de la variable

$$x_{.,j} = (x_{i,j})_{i=1}^{n}$$

$S_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x_{.,j}})^2 = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 - \overline{x_{.,j}}^2$ $(E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E(X)^2)$

Où $\overline{x_{.,j}} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$ $(**)$

Point moyen= centre de gravité

$$\overline{x} = (\overline{x_{i,j}})_{j=1}^{p} \in R^p$$

Notee : dans les formules (*) et (**), on a attriué le même poid $\frac{1}{n}$ à tous les individus.

Si par exemple, on veut déterminer la moyenne générale de polytech en analyse, on ne peut pas attribué le même poids aux moyennes de chaque classe. Car les effectifs diffèrent.

On définit les poids convenables $w_i > 0, \sum_i w_i = 1$

Matrice des poids : $D = diag(w_i)_{i=1}^{n}$

(**) et (*) deviennent alors $\overline{x_{i,j}} = \sum_{(i=1)}^{n} w_i x_{i,j}$

$$S_j^2 = \sum_{i=1}^{n} w_i(x_{i,j} - \overline{x_{.,j}})\text{^}2$$

- Matrice des données centrées

$$X_C = (x_{i,j} - \overline{x_{.,j}})_{i=1,\dots,n}$$

On enlève à chaque donnée, la moyenne de sa colonne.

- Matrice des données centrées réduites

$$X_{C\Omega} = \left(\frac{x_{ij} - \overline{x_{i,j}}}{S_j}\right)_{i=1,\dots,n, j=1,\dots,p}$$

Remarque :

$$\overline{x} = {}^t XD1_n \in R^p$$

$1_n \times L =$

$$\begin{pmatrix} l_1 & l_2 & \dots & l_p \\ \dots & \dots & \dots & \dots \\ l_1 & l_2 & \dots & l_p \end{pmatrix}$$

Avec $L = (l_1 \dots l_p)$

$D \times 1_n$

$$= \begin{pmatrix} \omega_1 & & \\ & \dots & \\ & & \omega_n \end{pmatrix}$$

Avec $D = \begin{pmatrix} \omega_1 \\ \dots \\ \omega_n \end{pmatrix}$

Où $1_n = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix}$

$$X_C = X - 1_n .^t \overline{x}$$

$\overline{x}$ était un vecteur colonne.

$$S_j^2 =.^t \left(x_{.,j} - \overline{x_{.,j}}.\, 1_n\right) D \left(x_{.,j} - \overline{x_{.,j}}.\, 1_n\right)$$

C'est la matrice des variances covariances.

$$S^2 =.^t X_C D X_C = \left[\sum_{i=1}^{n} w_i .^t (x_{.,j} - \overline{x_{.,j}})(x_{.,j'} - \overline{x_{.,j'}})\right]_{j,j'}$$

$$S_j^2 = [S^2]_{j,j'}$$

Note : espace des individus : (X,M,D)

2) Espace des variables

Variable $x_{.,j} = \left(x_{i,j}\right)_{i=1}^{n} \in R^n$

On va aussi définir une métrique et une matrice de poids comme on l'a fait dans l'espace des individus.

- Métrique : D

Remarque :

$$< x_{.,j} - \overline{x_{.,j}}, x_{.,j'} - \overline{x_{.,j'}} >_D = \sum_i w_i (x_{.,j} - \overline{x_{.,j'}})(x_{.,j'} - \overline{x_{.,j'}}) = cov(x_{.,j}, x_{.,j'})$$

$$\left|\left| x_{.,j} - \overline{x_{.,j'}} \right|\right|_D^2 = \sum_i w_i .^t (x_{.,j} - \overline{x_{.,j}})(x_{.,j'} - \overline{x_{.,j'}})$$

$$cor\left(x_{.j}, x_{.,j'}\right) = \frac{cov\left(x_{.,j}, x_{.,j'}\right)}{S_j S_{j'}}$$

- **Matrice des poids : M**

$\boldsymbol{n \times p}$	$\boldsymbol{p \times p}$	$\boldsymbol{n \times n}$
$\boldsymbol{p \times n}$	$\boldsymbol{n \times n}$	$\boldsymbol{p \times p}$

Note : espace des variables $(.^t X, D, M)$

$.^t X$ Contient nos données

D est notre métrique

M est la matrice des poids.

II. ACP

Rappel du problème : réduire la dimension de représentation des données.

1) Problème de l'ACP comme modèle

On suppose que $x_{i,.} = \left(x_{i,j}\right)_{j=1}^{p}$ est une réalisation du vecteur aléatoire $X_{i,.}$ A valeurs dans R^p.

Modèle :

$$\left\{\begin{array}{c} \boldsymbol{x_{i,.} = \mu_i + \epsilon_i, i = 1, \dots, n.} \ (\boldsymbol{\epsilon_i} \text{ é}\boldsymbol{tant\ al\acute{e}atoire, \mu_i d\acute{e}terministe\ donc\ non\ al\acute{e}atoire}) \\ \boldsymbol{\mu_i \in \overline{\mu} + E_q\ avec \dim E_q = q \leq p} \\ \boldsymbol{E(\epsilon_i) = 0_{R^p}\ esp\acute{e}rance\ (1)} \\ \boldsymbol{var((\epsilon_i)_{i=1}^{n}) = \Delta\ diagonale\ par\ blocs\ (2)} \\ \boldsymbol{cov(\epsilon_i, \epsilon_{i'}) = 0_{R^{p\times p}}\ si\ i \neq i'} \end{array}\right\}$$

Note : (1) et (2) font de $(\epsilon_i)_{i=1}^{n}$ un bruit blanc

Paramètres à estimer :

$$\theta = \left((\mu_i)_{i=1}^{n}, E_q, \Lambda\right)$$

quand on ne fait pas d'hypothèses de loi sur les données, on utilise la méthode des moindres carrées.

2) Estimation par les moindres carrées

Pb :

$$\boldsymbol{min_{E_q, (\mu_i)_{i=1}^{n}}} \left\{ \sum_i \boldsymbol{w_i} \left|\left| \boldsymbol{x_i - \mu_i} \right|\right|_{\boldsymbol{M}}^{\boldsymbol{2}}, \boldsymbol{dim\ E_q = q, \mu_i - \overline{\mu} \in E_q} \right\}$$

Notons $\overline{Y} = \begin{pmatrix} .^t(\mu_1 - \overline{\mu}) \\ .^t(\mu_2 - \overline{\mu}) \\ \dots \\ .^t(\mu_n - \overline{\mu}) \end{pmatrix}$

Exercice : (par Pythagore)

$$\sum_i w_i \left|\left| x_{i,.} - \mu_i \right|\right|_M^2 = \sum_i w_i \left|\left| x_{i,.} - \overline{x} + \overline{\mu} - \mu_i \right|\right|_M^2 + \left|\left| \overline{x} - \overline{\mu} \right|\right|_M^2$$

Il suffit de remplacer $||\quad||_M^2$ par un produit matriciel

Il faut aussi se rappeler que $x_{i,.}$ est un vecteur colonne.

Pour minimiser notre terme, il faut et il suffit de minimiser chaque terme de la somme.

$$\arg\min_{\overline{\mu}} ||\overline{x} - \overline{\mu}||_M^2 = \overline{x}$$

Il reste à résoudre

$$\min_{E_q,(\mu_i)_{i=1}^n} \sum_i w_i \left|\left|[X_C]_{i,.} - \overline{Y_{i,.}}\right|\right|_M^2$$

Devoir :

Chercher la décomposition SVD et l'approximation d'une matrice à un rang donné.

28\09\23

Proposition : l'estimation Y de rang q est obtenue par la décomposition en valeurs singulières de (X_c, M, D).

On l'obtient en faisant :

$$Y_q = \sum_{j=1}^{q} \lambda_j^{\frac{1}{2}} u^j .^t v_j = U_q \Lambda^{\frac{1}{2}} .^t V_q$$

Où U_q, Δ_q et V_q sont données dans le théorème suivant :

Théorème : toute matrice $X_c \in R^{n\times p}$ de $rg(X_c) = r$, de colonnes centrées (la somme de tous les élements de la colonne va donner 0 (vérifier)) se met sous la forme

$$X_c = U\Lambda^{\frac{1}{2}} .^t V = \sum_{j=1}^{r} \lambda^{\frac{1}{2}} . u^j .^t v^j$$

Le terme $u^j .^t v^j$ est une matrice.

Si on veut une approximation de rang q, on s'arrête à q dans la somme.

Où

i) $U \in R^{n\times p}$ des vecteurs propres de la matrice $X_c M .^t X_c D$

ii) $\Lambda = diag\left(\lambda_j\right)_{j=1}^{r}$ où les r premières valeurs $\lambda_1 > \cdots > \lambda_r > 0$ sont les valeurs propres de la matrice $.^t X_c D X_c M = S^2 M$

iii) $V \in R^{p\times r}$ matrices des vecteurs propres de $S^2 M$

Notes :

i) $\boldsymbol{U = [u1\ u2 \dots u^r], V = [v^1\ v^2 \dots v^r]}$

ii) $\boldsymbol{X_c M .^t X_c D\ u^j = \lambda_j u^j}$

$$\boldsymbol{S^2 M v^j = \lambda_j v^j}$$

iii) Liens entre U et V

$$\boldsymbol{U = X_c M V \Lambda^{-\frac{1}{2}} \in R^{n\times r}}$$

$$\boldsymbol{V = .^t X_c D U \Lambda^{-\frac{1}{2}} \in R^{p\times r}}$$

iv) les colonnes de U sont D orthonormées et donc $\boldsymbol{.^t UDU = I_r}$
les colonnes de V sont M orthonormées et donc $\boldsymbol{.^t VMV = I_r}$

Définition :

Axes principaux : les v^j

Facteurs principaux : les u^j

notes :

espcaes principaux :

1)

$$E_q = R\{v^1, v^2, \dots, v^q\}, q \leq r$$

2) $E_1 \subset E_2 \subset \cdots \subset \cdots$

Composantes principales

La matrice des composantes principales C est celle dans laquelle la ligne i est la projection M-orthogonale de la ligne i de la matrice X_c dans l'espace E_r.

Notons $C_i = \sum_{j=1}^{r} .^t (x_i - \overline{x}) M v^j \, . v^j = \sum_{j=1}^{r} c_i^j v^j$

$$c_i^j = .^t (x_i - \overline{x}) M v^j$$

Et $C = \left(c_i^j\right)_{i=1,\dots,n, j=1,\dots,r}$=matrice des composantes principales.

La colonne $C^j = \left(c_i^j\right)_{i=1}^n$ est la composante principale j

Remarque : $\boldsymbol{C = X_c MV}$

Note :

La matrice des composantes principales C représente le jeu de données transformée par l'ACP.

Propriétés :

1) $\sum_{i=1}^n \boldsymbol{c_i^j = 0}$
2) $\sum_{i=1}^n \boldsymbol{w_i \left(c_i^j\right)^2 = \lambda_j}$

   $\sum_{i=1}^n \boldsymbol{w_i \left(c_i^j\right)^2}$ **est la variance de la composante $\boldsymbol{C^j}$**
3) **Ces composantes sont décorrélés.**

$$\boldsymbol{cov\left(c^j, c^{j'}\right) = \begin{cases} \lambda_j \; si \; j = j' \\ 0 \end{cases}}$$

4) $\boldsymbol{tr(S^2 M) = \sum_j \lambda_j}$

   Pour la preuve de celle-ci, $\boldsymbol{X_C = U\Lambda^{\frac{1}{2}} .^t V}$ et $\boldsymbol{S^2 = .^t X_c D X_c}$

En effet :

Pour montrer 1, on applique la somme sur i sur $C_i = \sum_{j=1}^r .^t (x_i - \overline{x}) M v^j \, . v^j$

Pour montrer 2,

La matrice de covariances-variances du jeu de données C qui est centrée est $.^t CDC$ et il faut alors montrer que cette matrice est diagonale et qu'on a les λ_j sur la diagonal

RESUME DE L'ACP ($\boldsymbol{X_C, M, D}$)

Données : X, M, D

On construit M en essayant d'éliminer les unités dans nos données

Si on peut justifier le même poids à donner aux individus, on prend D la matrice $diag\left(\frac{1}{n}\right)_{i=1}^{n}$.

Sinon, on se sert du contexte de l'exercice pour trouver la matrice des poids. Par exemple dans l'exemple des moyennes d'analyse de chaque classe de po dont j'ai parlé plus haut.

Calculer $X_c = X - 1_n.^t \overline{x}$

Calculer $S^2 =.^t X_c D X_c$

Calculer $S^2 M$

Décomposition spectrale de $S^2 M$

Valeurs propres $\lambda_1 > \cdots > \lambda_r > 0$

Vecteurs propres $\widetilde{v^1}, \widetilde{v^2}, \ldots, \widetilde{v^r}$

$$v^j = \frac{1}{\sqrt{.^t \widetilde{v^j} M \widetilde{v^j}}} \widetilde{v^j}$$

$$V = [v^1, \ldots, v^r]$$

$$C = X_c M V$$

Les λ_j representent les variablités des composantes principales. Les λ_j doivent être différentes si non l'aCp ne nous a pas apporté car les distances entre les individus dans notre projection sont quasiment les même.

L'ACP est intéressant dans le cas où l'on a une décroissance forte entre les λ_i

Pour savoir la part de l'information qu'on conserve dans l'espace qu'on garde est donnée par $\frac{\lambda_1 + \cdots + \lambda_q}{\sum_j \lambda_j}$

Car $\sum_j \lambda_j$ représente toute l'information.

Inertie totale

$$I_t = \sum_i w_i \left|\left|x_{i,\cdot} - \overline{x}\right|\right|_M^2$$

Remarque :

Pour M=I,

$$I_t = \sum_i w_i \sum_j \left(x_i^j - \overline{x^j}\right)^2 = \sum_j S_j^2$$

Inertie expliquée par un sous espace vectoriel E de R^p

$$\boldsymbol{I_E = \sum_i w_i \left|\left|P_E^{+}(x_i - \overline{x})\right|\right|_M^2}$$

P_E^{+} est la projection M-orthogonale sur E.

Soit $v \in R^p$ tels que $\left|\left|v\right|\right|_M^2 = 1$

Proposition :

$$\boldsymbol{I_{\Delta_v} =.^t vMS^2Mv}$$

Preuve :

$$\boldsymbol{P_{\Delta_v}^{+}(x_i - \overline{x}) =.^t (x_i - \overline{x})Mv.v}$$

$$I_{\Delta_v} = \sum_i w_i [[.^t (x_i - \overline{x})Mv.v||_M^2$$

$$\boldsymbol{= \sum_i w_i |.^t (x_i - \overline{x})Mv|^2}$$

$$\boldsymbol{= \sum_i w_i .^t vM(x_i - \overline{x}).^t (x_i - \overline{x})Mv}$$

$$\boldsymbol{=.^t vM\left[\sum_i w_i (x_i - \overline{x}).^t (x_i - \overline{x})\right]Mv}$$

$$\boldsymbol{=.^t vMS^2Mv}$$

Lemme

$v_1, v_2 \in R^p$, $v_1 \perp_M v_2$, alors $I_{R\{v_1,v_2\}} = I_{v_1} + I_{v_2}$

Notes :

Posons $r = rg(X)$

1) $I_t = \sum_{j=1}^{n} I_{\Delta_{v^j}}$ où $v = [v^1 \dots v^r]$

$$I_{\Delta_{vj}} = \lambda_j$$

2) Rappeler que les espaces principaux sont données par :

$$\boldsymbol{E_1 = Rv_1}$$

$$\boldsymbol{E_2 = R\{v_1, v_2\}}$$

...

$$\boldsymbol{E_k = R\{v_1, \dots, v_k\}}$$

I) $\boldsymbol{v^1 = arg\ max}_{\substack{v\in R^p \\ ||v||_M^2=1}} \boldsymbol{I_{\Delta_v}}$

Pb : $\boldsymbol{max}_{\substack{v\in R^p \\ ||v||_M^2=1}} \{.^t \boldsymbol{vMS^2Mv}\}$

Rappel : Lagrangien $L(v,\lambda) =.^t vMS^2Mv + \lambda(\left|\left|v\right|\right|_M^2 - 1)$

λ est le multiplicateur de Lagrange.

La matrice MS^2M est symétrique.

Donc $\frac{\partial(.^t v\,(MS^2M)v)}{\partial v} = 2\,MS^2Mv$

En fait $\frac{\partial(.^t XAX)}{\partial X} = (A +.^t A)X$

La point où L(v,λ) est maximal est un point critique

$$\begin{Bmatrix} \nabla_\lambda L(v,\lambda) = \left|\left|v\right|\right|_M^2 - 1 = 0 \\ \nabla_v L(v,\lambda) = 2MS^2Mv - 2\lambda Mv = 0 \end{Bmatrix}$$

M est sdp donc inversible

$MS^2Mv = \lambda Mv \leftrightarrow S^2Mv = \lambda v$ **(on a multiplié par l'inverse de M)**

En multipliant ceci $MS^2Mv = \lambda Mv$ par la transposée de v,

On retrouve alors $\boldsymbol{\lambda =.^t vMS^2Mv}$

Remarque :

V est un vecteur propre de S^2M associée à la valeur propre $\lambda = .^t vMS^2Mv = I_{\Delta_v}$ à maximiser → **v est un vecteur propre M-normé associé à la plus grande valeur propre $\boldsymbol{\lambda}$**

Poser : $v^1 = v \; et \; \lambda_1 = \lambda$

ii) exo : Déterminer $\boldsymbol{\lambda_2 = arg\,max}_{\substack{v\in R^p \\ ||v||_M^2=1 \\ .^t vMv^1=0}} \; \boldsymbol{I_{\Delta v}}$

Qualité de représentation des individus

1) **Qualité globale**

$$\frac{\boldsymbol{I_{E_k}}}{\boldsymbol{I_t}} = \frac{\sum_{i=1}^{k} \boldsymbol{\lambda_i}}{\sum_{i=1}^{n} \boldsymbol{\lambda_i}} \in]\mathbf{0}, \mathbf{1}]$$

Une mesure de la qualité globale de représentation des individus dans $\boldsymbol{E_k}$

2) **Qualité de représentation d'un individu**

Elle est donnée $\cos^2((x-\overline{x}), P_{E_k^{\perp}}(x-\overline{x})) = \frac{\sum_{i=1}^{k}\left(c_i^j\right)^2}{\sum_{i=1}^{r}\left(c_i^j\right)^2} \in]0,1]$

En fait $\cos^2\left(\left(x_{i,.}-\overline{x}\right), P_{E_k^{\perp}}\left(x_{i,.}-\overline{x}\right)\right) = \frac{\left|\left|P_{E_k^{\perp}}(x_{i,.}-\overline{x})\right|\right|_M^2}{\left|\left|x_{i,.}-\overline{x}\right|\right|_M^2}$

$$= \frac{\sum_{j=1}^{k} < x_{i,.}-\overline{x}, V_j >_M^2}{\sum_{j=1}^{r} < x_{i,.}-\overline{x}, V_j >_M^2}$$

$$= \frac{\sum_{j=1}^{k}\left(.^t\left(x_{i,.}-\overline{x}\right)MV_j\right)^2}{\sum_{j=1}^{r}\left(.^t\left(x_{i,.}-\overline{x}\right)MV_j\right)^2}$$

Se garder d'interpréter les individus mal représentés $\cos^2(\ldots)$ proche de 0

Lire le slide 20 du cours.

Pour le prochain tp,

A partir d'un jeu de données, utiliser la bibliothèque scibit-learn.

Aller dans la documentation de scibit learn, regarder datasets et choisir le jeu de données iris. (on ne va pas tenir compte de l'espèce des fleurs)

Juste les 4 dimensions quantitatives qu'il y a dedans et essayer de les condenser sur deux dimensions.

```
Entrée [17]: import sklearn
             from sklearn import datasets
             from sklearn.preprocessing import StandardScaler
             from sklearn.decomposition import PCA
```

```
Entrée [18]: iris0=datasets.load_iris()

Entrée [19]: dir(iris0)

Out[19]: ['DESCR',
          'data',
          'data_module',
          'feature_names',
          'filename',
          'frame',
          'target',
          'target_names']
```

```
Entrée [21]: print(iris0.DESCR)
```

```
.. _iris_dataset:

Iris plants dataset
--------------------

**Data Set Characteristics:**

    :Number of Instances: 150 (50 in each of three classes)
    :Number of Attributes: 4 numeric, predictive attributes and the class
    :Attribute Information:
        - sepal length in cm
        - sepal width in cm
        - petal length in cm
        - petal width in cm
        - class:
                - Iris-Setosa
                - Iris-Versicolour
                - Iris-Virginica

    :Summary Statistics:

    ============== ==== ==== ======= ===== ====================
                    Min  Max   Mean    SD   Class Correlation
    ============== ==== ==== ======= ===== ====================
    sepal length:   4.3  7.9   5.84   0.83    0.7826
    sepal width:    2.0  4.4   3.05   0.43   -0.4194
    petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
    petal width:    0.1  2.5   1.20   0.76    0.9565  (high!)
    ============== ==== ==== ======= ===== ====================
```

```
Entrée [23]: x=iris0.data
             x.shape

Out[23]: (150, 4)
```


```
Entrée [23]: x=iris0.data
             x.shape

Out[23]: (150, 4)
```


```
Entrée [24]: scaler= StandardScaler(with_mean=True, with_std=True)
             Xcr=scaler.fit_transform(x)
```

StandardScaler permet de centrer et réduire.

Vérifions que les données sont centrées :

```
Entrée [25]: Xcr.mean(axis=0)
Out[25]: array([-1.69031455e-15, -1.84297022e-15, -1.69864123e-15, -1.40924309e-15])
```

Vérifions que les données sont réduites

```
Entrée [26]: Xcr.var(axis=0)
Out[26]: array([1., 1., 1., 1.])
```

Faisons maintenant l'ACP

Ici M c'est l'identité

```
Entrée [28]: acp=PCA()
             acp.fit(Xcr)
Out[28]: PCA()
```

```
Entrée [29]: dir(acp)
Out[29]: ['__abstractmethods__',
          '__annotations__',
          '__class__',
          '__delattr__',
          '__dict__',
          '__dir__',
          '__doc__',
          '__eq__',
          '__format__',
          '__ge__',
          '__getattribute__',
          '__getstate__',
          '__gt__',
          '__hash__',
          '__init__',
          '__init_subclass__',
          '__le__',
          '__lt__',
          '__module__',
          '__ne__',
          '__new__',
          '__reduce__',
          '__reduce_ex__',
```

Faire des recherches sur les méthodes suivantes :

```
'components_',
'copy',
'explained_variance_',
'explained_variance_ratio_',
'fit',
'fit_transform',
'get_covariance',
'get_feature_names_out',
'get_metadata_routing',
'get_params',
'get_precision',
'inverse_transform',
'iterated_power',
'mean_',
'n_components',
'n_components_',
'n_features_',
'n_features_in_',
'n_oversamples',
'n_samples_',
'noise_variance_',
'power_iteration_normalizer',
'random_state',
'score',
'score_samples',
'set_output',
'set_params',
'singular_values_',
'svd_solver',
'tol',
'transform',
```

```
Entrée [30]: C=acp.fit_transform(Xcr)

Entrée [31]: acp.explained_variance_ratio_
  Out[31]: array([0.72962445, 0.22850762, 0.03668922, 0.00517871])
```

Ici, on a la première variance sur la somme des variances, et ainsi de suiste

Et justement la somme donne 1

```
Entrée [36]: acp.explained_variance_ratio_.sum()
  Out[36]: 1.0
```

Ici, les deux premières axes contiennent plus de 90% de la variabilité totale.

$$ACP(X_c, M, D) \rightarrow C = X_c MV \leftarrow X_c = U\Lambda.^t V$$

A partir ce cet ACP des individus, on peut déduire l'ACP des variables :

$ACP(.^t X_c, D, M) \rightarrow F = .^t X_c DU \leftarrow .^t X_c = V\Lambda.^t U$ **et on connait la relation entre U et V**

En fait $F = .^t X_c D X_c MV\Lambda^{-\frac{1}{2}} = S^2 MV\Lambda^{-\frac{1}{2}} = V\Lambda\Lambda^{-\frac{1}{2}} = V\Lambda^{\frac{1}{2}}$

Choix de la dimension

On rappelle que l'ACP des individus crée de nouvelles variables qui sont des combinaisons linéaires des premières, telles que l'inertie expliquée est maximale de la première à la dernière.

Question : combien de nouvelles variables conserver ?

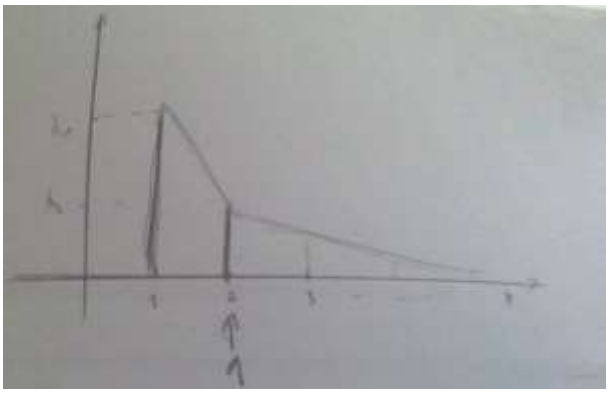
Note :

Si le but de cette réduction de dimension, est de représenter graphiquement le nuage des individus, alors, on conservera $q \leq 2$**(en général q=2)**

il existe dans la littérature, des suggestions de critères de sélection. Parmi, il y en a qui sont des heuristiques (c'est une procédure qu'on utilise sans avoir une preuve que ça marche)

i) **Eboulis des valeurs propres**

Cela consiste à construire le graphique des valeurs propres.



Lorsque l'ACP est pertinent pour un jeu de données, le diagramme à barres des valeurs propres présente un coude qui sépare une partie à forte décroissance d'une autre à décroissance modérée :

On sélectionne alors, la valeur q qui correspond à la position du coude. (ce résultat est intuitif).

En fait cela se justifie par le fait que quand on prend les premières dimensions, le reste n'apporte pas grande chose.

ii) Part de l'inertie expliquée

On se donne un seuil s de l'inertie expliquée par les composantes principales à sélectionner.

On sélectionne alors la plus petite valeurs q telle que la parte de l'inertie expliquée $\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{r} \lambda_j} \geq \boldsymbol{s}$

iii) Règle de Kaiser

Elle repose sur l'idée qu'un axe est intéressant si sa variabilité est supérieure à la moyenne.**(c'est aussi heuristique).**

On sélectionne la plus grande valeur de q pour la quelle $\lambda_q \geq \lambda_s = \frac{1}{r}\sum_{j=1}^{r} \boldsymbol{\lambda_j}$

REPRESENTATIONS GRAPHIQUES

- **Nuage des individus pour q=2**

Un individu supplémentaire est un individu qu'on a pas utilisé dans le calcul de l'ACP mais qu'on veut représenter pour faire des interprétations.

Supposons qu'on appelle ses données : $x^s = \left(x_j^s\right)_{j=1}^{p} \in R^p$

Pour représenter un tel individu, on fait sa projection dans notre plan (c_i^1, c_i^2)

$$c^{s,1} =.^t (x^s - \overline{x})Mv^1$$

$$c^{s,2} =.^t (x^s - \overline{x})Mv^2$$

on n'oublie pas de centre les données de x^s.

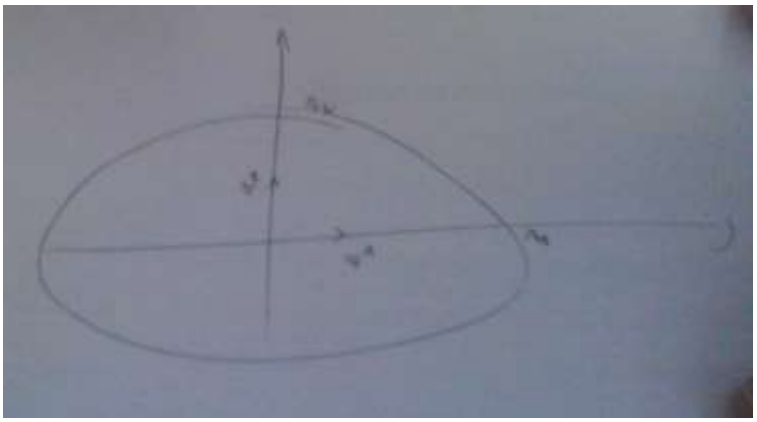
Ellipse de Hotelling :

$$\boldsymbol{q = 2}$$

$$\boldsymbol{\delta = \frac{q(n^2-1)}{n(n-q)} Q_{0,95}\left(F(q, n-q)\right)}$$

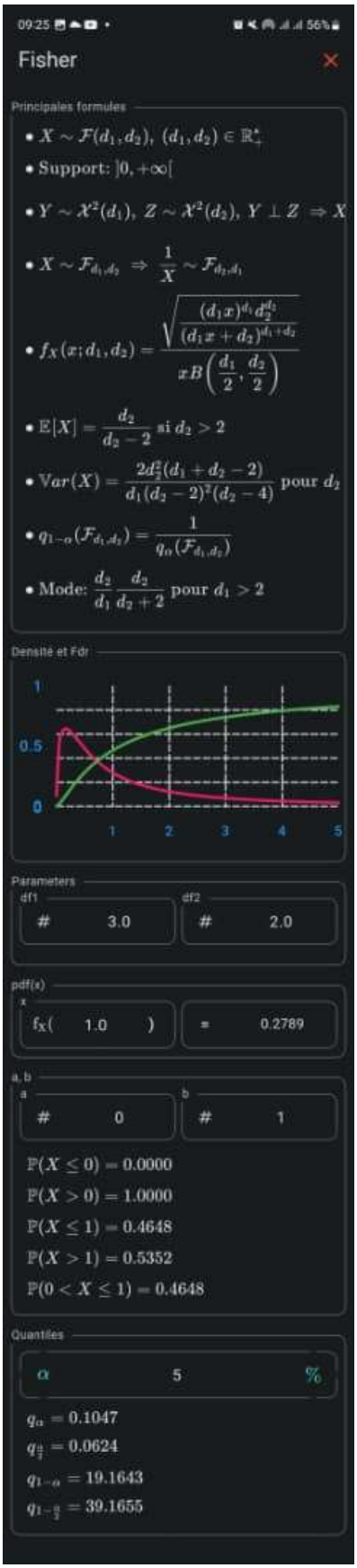
F est la loi de Fisher.

$$\boldsymbol{r_1 = \sqrt{\delta\lambda_1}, r_2 = \sqrt{\delta\lambda_2}}$$

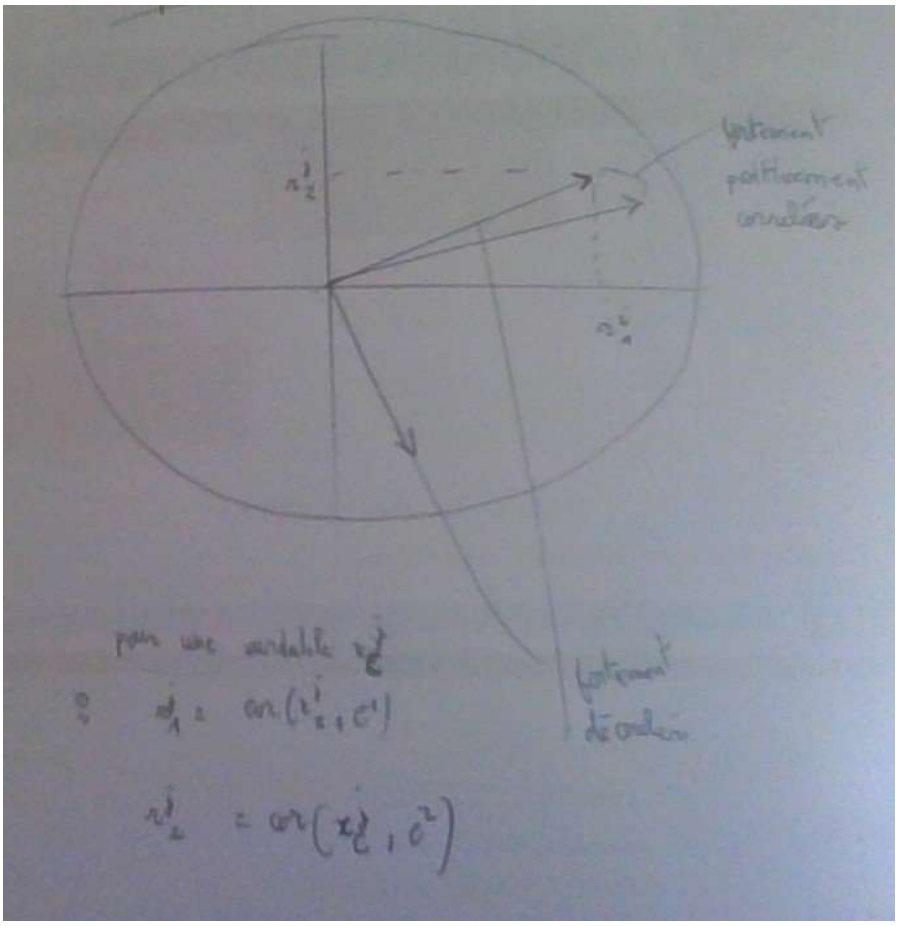


Pour avoir plus d'informations sur la loi de Fisher, on peut aller sur l'application statchat ou bien statistics tutor.

Voici par exemple le contenu sur statchat :



- **Représentation des variables : cercle des corrélations**



Dans ce qui suit, c^1 est la première colonne dans C, c^2 pareil

- Chaque variable est représentée par un vecteur de norme ≤ 1 ; la variable est d'autant bien représentée que la norme associée est proche de 1.
- Deux variables sont fortement positivement corrélées lorsque l'angle entre leurs vecteurs est proche de 0.
- Lorsque l'angle est proche de 180 degrés, les deux variables sont fortement négativement corrélées.
- Lorsque l'angle est proche de 90 degrés, les deux variables sont décorrélées.

on évite de faire des commentaires pour des variables qui sont de norme très petite car on voir en fait une telle variable sous le mauvais angle.

On peut avoir deux variables de norme très faible qui semblent décorrélés dans notre cercle mais qui sont fortement corrélées dans la vraie vie.

A savoir : quand on dit de faire l'ACP normée, cela signifie qu'il faut faire l'ACP sur le jeu de données centrées-réduites (sur $X_{c,r}$ avec $M = I$))

Le nombre de valeurs propres devra être p (**si ce n'est pas le cas, il y a une erreur**).

Chapitre 2 : analyse factorielle discriminante (AFD)

Données

X

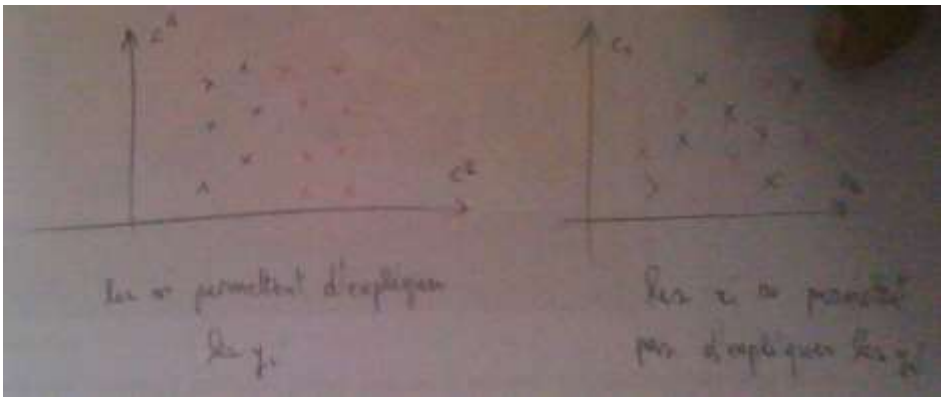
Ligne i : (x_i, y_i)

$x_i \in R^p$

y_i un label(indique la classe ou groupe de l'individue)

Problème : explorer les données pour savoir si les entrées $x_i$permettent d'expliquer les sorties y_i(réduction de la dimension des x_i pour permettre une visualisation optimale du lien entre les x_i et les y_i)

Si on réduit les informations des données dans les x_i dans deux variables par exemples et que notre nuage de points nous donne(dans le nuage, on ajoute les y_i)



I- Stat – descriptives

- $X = \left(x_i^j\right)_{\substack{i=1,\dots,n \\ j=1,\dots,p}} \in R^{n\times p}$
- $x_i = \left(x_i^j\right)_{j=1}^{p} \in R^p$
- Supposons que le nombres de classes (ou de groupes) est $K > 1$
- Coder les classes par

  $Y = \left(y_{i,k}\right)_{\substack{i=1,\dots,n \\ k=1,\dots,K}} \in \{0,1\}^K$ (**un individu ne peut appartenir qu'à une seule classe**

  Avec $\sum_{k=1}^{K} y_{i,k} = 1$

Exemples

Pour K=3,

$(1, 0, 0)$ correspond à la classe 1

(0,1,0) correspond à la classe 2

(0,0,1) correspond à la classe 3.

- **Effectif de chaque classe k :**

$$m_k = \sum_{i=1}^{n} Y_{i,k}$$

- **Centre de gravité (point moyen)**

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \in R^p = \frac{1}{n}.{}^t X 1_n \in R^p$ **avec** $1_n = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$

On transpose parce qu'on veut parcourir les individus et non les variables.

- **Centre de gravité de la classe k**

$$\overline{x_k} = \frac{1}{n_k}\sum_{i=1}^{n} y_{i,k}x_i \in R^p$$

- **Variance totale**

$\boldsymbol{S_T^2 = \frac{1}{n}.{}^t X_c X_c}$ **Où** $\boldsymbol{X_c = X - 1_n.{}^t\overline{x}}$ **données centrées**

Remarque : $\boldsymbol{S_T^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}).{}^t(x_i - \overline{x}) \in R^{p\times p}}$

- **Variance infra-classe :** mesure de la variabilité entre les groupes.

$$S_W^2 = \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n} y_{i,k}(x_i - \overline{x_k}).{}^t(x_i - \overline{x_k}) \in R^{p\times p}$$

- **Variance interclasse**

Mesure de la variabilité entre les groupes

B pour between

$$S_B^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\overline{x_k} - \overline{x}).{}^t(\overline{x_k} - \overline{x}) \in R^{p\times p}$$

Proposition :

$$\boldsymbol{S_T^2 = S_W^2 + S_B^2}$$

II- PRINCIPE DE L'AFD

$\mathrm{AFD}(X \times Y) = ACP(G, I, D)$

$$G = \left(\overline{x}_k^j\right)_{\substack{k=1,\dots K \\ j=1,\dots,p}},$$

$$\overline{x}_k^j = \frac{1}{nk}\sum_{i=1}^{n} y_{i,k}x_i^j$$

$$D = diag\left(\frac{n_k}{n}\right)_{k=1}^{K}$$

Remarque : Si X est centrée,

$\overline{\boldsymbol{x}} = \mathbf{0}_{R^p}$ **; alors** $\overline{\boldsymbol{g}} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{k}} \boldsymbol{n_k}\overline{\boldsymbol{x_k}} = \overline{\boldsymbol{x}} = \mathbf{0}_{R^p}$

Note : si X n'est pas centrée, il faut le faire dans les calculs de l'ACP en posant

$$G_c = G = 1_n.^t \overline{x}$$

- Il est conseillé de réduire les colonnes de X avant d'effectuer une AFD.
- $\boldsymbol{S_B^2} =.^t \boldsymbol{G_c D G_c}$
- **L'AFD** repose sur la décomposition spectrale de S_B^2 c'est-à-dire $\boldsymbol{S_B^2 M}$ **avec M=I**
- Décomposition spectrale $(S_B^2) \rightarrow$

- **Valeurs propres** $\boldsymbol{\lambda_1 > \cdots > \lambda_p > 0}$
- **Valeurs propres normés :** $\boldsymbol{v^1, \ldots, v^p}$

On pose $\boldsymbol{V = \{v^1, \ldots, v^p\}}$

- **Données transformées par l'AFD** $\boldsymbol{X_c V}$

Xc V

Remarque : dans certains documents la métrique $M = (S_T^2)^{-1}$ (ça permet de déformer le nuage)

$$\widetilde{\boldsymbol{G}} = \boldsymbol{GMV}$$

$$\widetilde{\boldsymbol{X}} = \boldsymbol{X_c MV}$$

Est la formule qui sépare au mieux les classes.

Exercice : reprendre l'analyse du jeu de données d'iris en faisant une AFD avant l'ACP et comparer à ce qu'on a obtenu quand on a uniquement fait l'ACP.

Chapitre 3 : Analyse Factorielle des correspondances(AFC)

Données : $\{(x_i, y_i)\}_{i=1}^{n}$ **où** $(x_i, y_i) \in X(\Omega) \times Y(\Omega)$ avec $X(\Omega) = \{a_1, .., a_k\}$ et $Y(\Omega) = \{b_1, \dots, b_L\}$

Les a_i et les b_i sont les modalités.

Des ensembles finis non ordonnées.

Remarque : les données sont les observations de deux variables qualitatives X et Y.

Problème : représenter les liens (les correspondances) entre les modalités de X et de Y en perdant le moins d'informations possibles.

Note : la question plus globale est celle de l'étude du lien ou de façon équivalente de l'étude de l'indépendance entre les deux variables en question. Au niveau global, elle peut être abordée par le test d'indépendance du khi-deux (**qui est un test global mais pour avoir la nature du lien entre les modalités, il faut faire l'analyse factorielle**).

1) **CONCEPTS**

   i) Tableau de contingence

A savoir : quand on analyse du texte, on dit qu'on fait du text mining.

Le tableau de contingence $M = (m_{k,l})_{\substack{k=1,\dots,k \\ l=1,\dots,L}} \in N^{k \times l}$

Où $m_{k,l}$ **est l'effectif du couple** $(a_k, b_k) \in X(\Omega) \times Y(\Omega)$

X\Y	… b_l ….	total
…		
a_k	… $m_{k,l}$ …	$n_{k,t}$
…	…	
total	$m_{t,l}$	n

$$n_{k,t} = \sum_{l=1}^{L} m_{k,l}$$

$$m_{t,l} = \sum_{k=1}^{K} n_{k,l}$$

$$n = \sum_{l=1}^{L}\sum_{k=1}^{K} m_{k,l} = \sum_{l} m_{t,l} = \sum_{k} n_{k,t}$$

Fréquences :

$F = (f_{k,l})_{\substack{k=1,\dots,K \\ l=1,\dots,L}} \in [0,1]^{K\times L}$ **avec** $f_{k,l} = \frac{n_{k,l}}{n}$

X\Y	… b_l ….	total
…		
ω_k …	… $f_{k,l} = \frac{n_{k,l}}{n}$ …	$f_{k,t}$
…		
total	$f_{t,l}$	1

$$f_{k,t} = \sum_{l=1}^{L} f_{k,l}$$

$$f_{t,l} = \sum_{k=1}^{K} f_{k,l}$$

$$1 = \sum_{l=1}^{L}\sum_{k=1}^{K} f_{k,l} = \sum_{l} f_{t,l} = \sum_{k} f_{k,t}$$

ii) Profils

Profils lignes

$$L_k = \left(\frac{f_{k,l}}{f_{k,t}}\right)_{l=1}^{L} \in [0,1]^{l}$$

$$= \left(\frac{n_{k,l}}{n_{k,t}}\right)_{l=1}^{L}$$

Si on regarde bien, ce sont des probabilités conditionnelles(avoir la modalité (k,l) sachant que X a la modalité k.

Posons $L = \begin{bmatrix} .^t L_1 \\ \dots \\ .^t L_k \\ \dots \\ .^t L_K \end{bmatrix} \in [0,1]^{K,L}$ **matrice des profils lignes.**

Profils colonnes

On va noter le profil de la colonne l par C_l

$$C_l = \left(\frac{f_{k,l}}{f_{t,l}}\right)_{k=1}^{K} \in [0,1]^K$$

$$= \left(\frac{n_{k,l}}{n_{t,l}}\right)_{k=1}^{K}$$

Si on regarde bien, ce sont des probabilités conditionnelles (avoir la modalité (a_k, a_l) sachant que Y a la modalité b_l. (en fait ce n'est pas vraiment une probabilité vu qu'on travaille sur un échantillon et non sur la population mais c'est une approximation de cette probabilité).

En fait $P\big((X = a_k), (Y = b_l)\big) = \frac{n_{kl}}{n}$

$P(Y = b_l) = \frac{n_{t,l}}{n}$ **donc** $P\big((X = a_k)\big|(Y = b_l)\big) = \frac{n_{k,l}}{n} \times \frac{n}{n_{t,l}} = \frac{n_{k,l}}{n_{t,l}}$

Posons $L = [C_1 \quad \dots \quad C_l \dots C_L] \in [0,1]^{K,L}$ **matrice des profils colonnes.**

Posons :

$$D_{K,.} = diag\left((f_{k,t})_{k=1}^{K}\right) \in R^{K \times K}$$

$$D_{.,L} = diag\left((f_{k,t})_{l=1}^{L}\right) \in R^{L \times L}$$

Propriétés : (exo)

Montrer que

$$L = D_{K,.}^{-1}.F$$

$$C = D_{.,L}^{-1}.^{t}F$$

iii) Métrique du Khi2 des profils lignes

$$d^2(L_k, L_{k'}) = \sum_{l=1}^{L} \frac{1}{f_{t,l}} \left(\frac{f_{k,l}}{f_{k,t}} - \frac{f_{k',l}}{f_{k',t}} \right)^2 = \sum_{l=1}^{L} \frac{1}{f_{t,l}} \left(\frac{n_{k,l}}{n_{k,t}} - \frac{n_{k',l}}{n_{k',t}} \right)^2$$

Remarque(exercice) :

La métrique du khi2 sur les profils lignes est définie par la matrice $D_{.,L}^{-1}$ c'est-à-dire :

$$d^2(x, y) =.^{t}(x - y) D_{.,L}^{-1} (x - y)$$

iv) Métrique du Khi2 sur les profils colonnes

Elle est définie par $D_{K,.}^{-1}$:

$$d^2(C_l, C_{l'}) = \sum_{k=1}^{K} \frac{1}{f_{k,t}} \left(\frac{f_{k,l}}{f_{t,l}} - \frac{f_{k,l'}}{f_{t,l'}} \right)^2 = \sum_{k=1}^{K} \frac{1}{f_{k,t}} \left(\frac{n_{k,l}}{n_{t,l}} - \frac{n_{k,l'}}{n_{t,l'}} \right)^2$$

2) Test d'indépendance du khi 2

Quand on fait un test en statistique, Il faut préciser les hypothèses qu'on veut tester :

- **Hypothèses :**
- (H_0) **: X et Y sont indépendantes (H0 est appelé hypothèse nulle dans le jargon)**
- (H_1) **: X et Y sont liées**

De façon générale, il faut s'assurer que H0 et H1 ne peuvent pas être vraies à la fois. Les hypothèses ne sont pas obligées d'être négations l'une de l'autre comme ici.

H0 doit être celle à laquelle on a plus confiance mais surtout l'hypothèse pour laquelle on est capable de construire un critère qui nous permet de savoir l'éloignement à la réalité sous cette hypothèse.

Il faut chercher un critère qui nous permette de savoir si on s'éloigne de H0 au profit de H1(c'est ce qu'on fait par la suite)

- **Statistique de test**

Posons $\widehat{f_{k,l}} = f_{k,t} \times f_{t,l} = \frac{n_{k,t} \times n_{t,l}}{n^2}$

$\widehat{f_{k,l}}$ est la fréquence du couple (a_k, b_l) sous (H_0)

Dans la formule qui suit, $n \times \widehat{f_{k,l}}$ est l'effectif attendu sous H_0

$$X^2 = n \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(f_{k,l} - \widehat{f_{k,l}}\right)^2}{\widehat{f_{k,l}}} = \sum_{k} \sum_{l} \frac{\left(n_{k,l} - n \times \widehat{f_{k,l}}\right)^2}{n \times \hat{f}_{k,l}}$$

X^2 est d'autant plus petite que l'hypothèse H_0 est proche de la réalité)

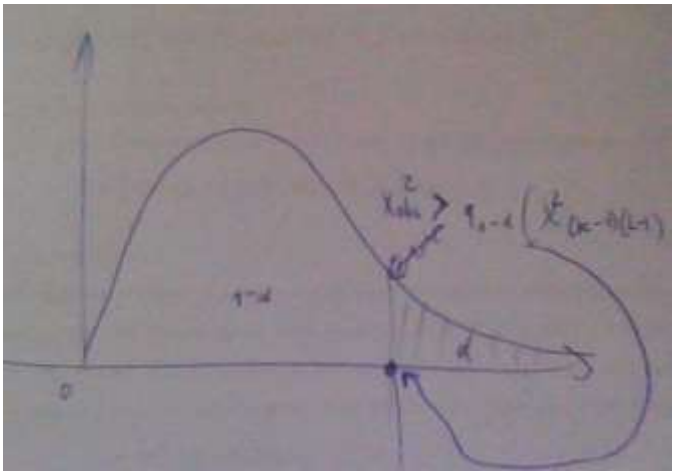
$X^2 \rightarrow^{(H_0)}_{n \to +\infty} X^2_{(K-1)(L-1)}$ (convergence en loi)

$X^2_{(K-1)(L-1)}$ est la loi du khi deux

Décision au seuil α(en général $\alpha = 5\%$)

Rejeter $(H_0) \leftrightarrow X^2_{obs} > q_{1-\alpha}\left(X^2_{(K-1)(L-1)}\right)$

$q_{1-\alpha}$ est la valeur qui permet d'avoir la fonction de repartition égale à $1 - \alpha$.



X_{obs} est juste la valeur de X calculée avec les valeurs observées.

On rejète en fait H0 lorsque les valeurs de X^2 sont grandes.

p-valeur $= P(X^2_{(K-1)(L-1)}) > X^2_{obs}$

09\11\23

Matrice centrée : $L_c = D_{K,.}^{-1}F - 1_K.{}^t\overline{L}$

$$\overline{L} = (f_{t,l})_l \in R^L$$

$$C = (f_{k,t})_k \in R^K$$

Métrique : $\boldsymbol{D_{.,L}^{-1}}$

Poids : $\boldsymbol{D_{K,.}^{-1}}$

- Calculer $\boldsymbol{S^2 =.{}^t L_c D_{K,.}^{-1} L_c}$
- Calculer $\boldsymbol{S^2 D_{.,L}^{-1}}$
- Déterminer U qui est la matrice des vecteurs propres de $S^2 D_{.,L}^{-1}$ qui doivent être $D_{.,L}^{-1}$ orthonormés **et rangées dans l'ordre décroissant des valeurs propres.**
- Calculer CP : $L_c D_{.,L}^{-1} U$: **les deux premières colonnes donnent les coordonnées de représentation des modalités de X et Y.**

ACP des profils colonnes

Matrice centrée : $C_c = D_{.,L}^{-1}F - 1_L.{}^t\overline{C}$

Métrique : $D_{K,.}^{-1}$

Poids : $D_{.,L}^{-1}$

$$\overline{C} = (f_{K,t})_K \in R^K$$

Calculer $S^2_{.,L} =.{}^t C_c D_{.,L}^{-1} C_c$

Calculer $S^2_{.,L} D_{K,.}^{-1}$

On représente les modalités dans le même plan et deux modalités qui sont éloignés traduisent que ces deux modalités sont très peu corrélés.

21\11\23

Chapitre 4 : Analyse de variance a un facteur

I. Exemple et position du problème

Un forestier s'intéresse à la hauteur moyenne des arbres dans 3 forêts. Pour les estimer et les comparer, il échantillonne un certain nombre de leurs hauteurs.

Forêt	1	2	3
	23.4	22.5	18.9
	24.4	22.9	21.1
	24.6	23.0	21.1
	24.9	24.0	22.1
	25.0	23.5	22.1
	26.2	22.8	23.5
	25.5		24.5
	27.0		18.5
	25.4		20.3
	26.5		19.7

Une des questions qui intéressent le forestier :

Les arbres ont-ils la même hauteur moyenne dans les 3 forêts ?

Formalisation du problème

Notations :

- $Y_{j,i}$ **: la variable aléatoire qui donne la hauteur de l'arbre i dans la forêt j.**
- μ_j **: la hauteur moyenne de l'arbre dans la forêt j. c'est-à-dire l'espérance de** $Y_{j,i}$

$$\mu_j = E[Y_{j,i}]$$

Modèle :

$$\boldsymbol{Y_{j,i} = \mu_j + \epsilon_{j,i}} \textbf{ avec } \left\{ \begin{array}{c} \boldsymbol{E[\epsilon_{j,i}] = 0} \\ \boldsymbol{Var(\epsilon_{j,i}) = \sigma^2} \\ \boldsymbol{\epsilon_{j,i} \rightarrow^{iid} N(0, \sigma^2)} \\ \boldsymbol{cov(\epsilon_{j,i}, \epsilon_{j',i'}) = 0 \; si \; (i,j) \neq (i',j')} \end{array} \right\}$$

La dernière condition traduit le fait que les arbres sont indépendants.

Remarque :

i) $Y_{j,i} \rightarrow^{iid} N(\mu_j, \sigma^2)$

ii) La question d'intérêt peut être formulée en termes d'hypothèses statistiques :
(H0) : $\mu_1 = \mu_2 = \mu_3$
(H1) : $\exists (j, j') \in \{1,2,3\}, \mu_j \neq \mu_{j'}$

II. Rappels du principe d'un test statistique paramétrique

On part d'un n-échantillon $X_1, \dots, X_n \rightarrow^{iid} \{P_\theta\}_{\theta \in \Theta}$ avec $\Theta \subset R^k$

i) **Bien formuler les hypothèses nulles(H0) et alternatives** (H1)

ii) **Construire une statistique de test :**

- **Une statistique sur un échantillon** est une fonction mesurable qui ne dépend d'aucun paramètre et qui ne dépend que de l'échantillon.
  Une statistique de test permet de mesurer l'écart à (H0)
- **On doit construire S(statistique de test) de sorte à connaître sa loi exacte ou asymptotique sous (H0)**
  Supposons $S \rightarrow^{H0}_{n \rightarrow +\infty} L$

iii) Déterminer les valeurs critiques : elles sont les plus souvent les quantiles associés à L.

EXEMPLES(1)

$X_1, \dots, X_n \rightarrow^{iid} \{N(\mu, \sigma^2)\}_{\mu \in R, \sigma^2 connue}$

Test de (H0) : $\boldsymbol{\mu = \mu_0}$ contre

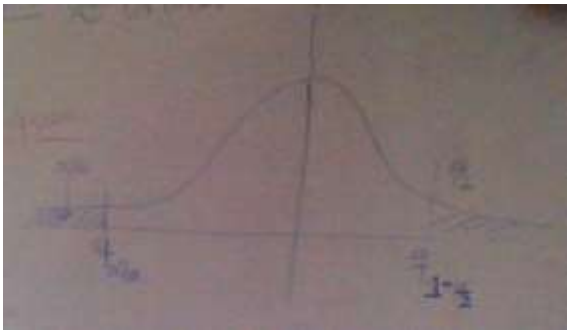
(H1) : $\boldsymbol{\mu \neq \mu_0}$

Statistique de test :

$$\overline{X_n} - \mu_0 \rightarrow^{H0} N\left(0, \frac{\sigma^2}{n}\right)$$

$$S = \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} \rightarrow^{H0} N(0,1)$$

Valeurs critiques : (elles permettent de déterminer la région critique)



On rejette (H0) lorsque la valeur S est dans l'une des 2 zones hachurés du schéma précédent.

Région critique :

$$\boldsymbol{W} =]-\infty, \boldsymbol{q}_{\frac{\alpha}{2}}[\cup]\boldsymbol{q}_{1-\frac{\alpha}{2}}, +\infty[$$

Décision :

Rejeter $(H_0) \leftrightarrow S_{obs} \in W$

(2) $X_1, \dots, X_n \rightarrow^{iid} \{N(\mu, \sigma^2)\}_{\theta=\binom{\mu}{\sigma^2} \in R \times R_+^*}$

Test de (H0) :

$\mu = \mu_0$ contre $(H1)$: $\mu < \mu_0$

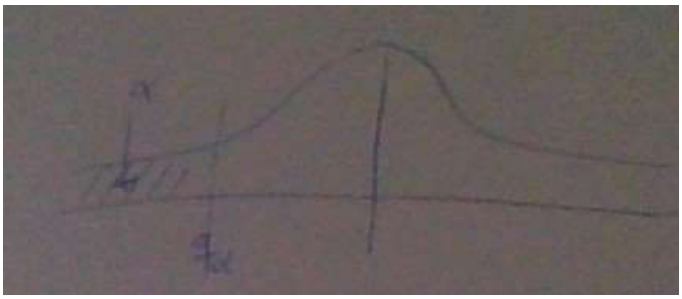
Stat de Test :

$\widehat{\sigma_n^2} = \frac{1}{n-1} \Sigma_{i=1}^{n} \left(X_i - \overline{X_n}\right)^2$ (Estimateur non biaisé de la variance)

$$T = \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sqrt{\widehat{\sigma_n}^2}} \rightarrow T_{n-1}$$

Région critique $(\boldsymbol{\alpha < 0,5})$: si non on donne le même poids à nos deux hypothèses

$$W =]-\infty, q_\alpha(T_{n-1})[$$



α est le risque maximale qu'on s'autorise à prendre lorsque H0 est vraie.

Décision :

Rejeter $(H_0) \leftrightarrow T_{obs} \in W$

III. Anova 1 (analyse de la variance à un facteur)

- **Hypothèses :**

- $(H0): \mu_1 = \mu_2 = \cdots = \mu_j$
- $(H_1): \exists j, j' \in \{1, \dots, J\}, \mu_j \neq \mu_{j'}$

  $J =$ Nombre de groupes
- n_j est l'effectif du groupe j
- $n = \sum_{j=1}^{J} n_j$

- **Statistiques usuelles :**

$$\overline{Y_{j,\cdot}} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{j,i}$$

$$\overline{Y} = \frac{1}{n} \sum_{j=1}^{J} n_j \overline{Y_j} = \frac{1}{n} \sum_{i=1}^{J} \sum_{i=1}^{n_j} Y_{j,i}$$

- **Somme des carrées**

$WSS = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{j,i} - \overline{Y_j}\right)^2$ = **somme des carrées intragroupes**

$BSS = \sum_{j=1}^{J} n_j \left(\overline{Y_j} - \overline{y}\right)^2$ = **somme des carrées inter-groupes**

$TSS = \sum_j \sum_i \left(Y_{j,i} - \overline{y}\right)^2$ = **WSS+BSS**

- **Dégrés de liberté**

Intra-groupe: **n-J**

Intergroupe : **J-1**

Total: n-1

Statistique de test

$\boldsymbol{F = \frac{\frac{BSS}{J-1}}{\frac{WSS}{n-J}} \rightarrow^{(H_0)} F_{J-1,n-J}}$**, loi de Fisher à (J-1,n-J) dégrés de liberté.**

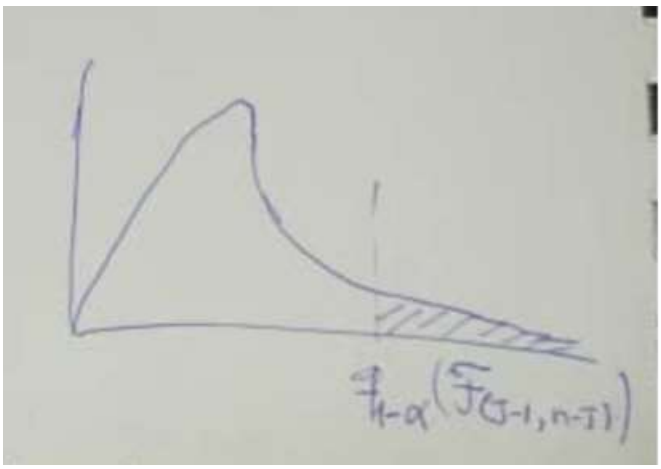
Décision au niveau $\boldsymbol{\alpha}$

(Rejeter (H0) ↔ $\boldsymbol{F_{obs} > q_{1-\alpha}(F_{J-1,n-J})}$

p-valeur :

$$\boldsymbol{p - valeur = P(F > F_{obs}|H_0) = 1 - F_{dr_{F_{J-1,n-J}}}(F_{obs})}$$

(regjeter (H0))↔ $\boldsymbol{p - valeur < \alpha}$



Données : $\boldsymbol{X = (x_{ij})_{\substack{i=1,\dots,n \\ j=1,\dots,p}}}$

x_{ij} est une donnée de la variable j sur l'individu i

Problème : regrouper les n individus des classes (clusters) de sorte que 2 individus d'une même classe soient moins dissemblables que 2 individus de 2 classes différentes.

1) Classification ascendante hiérarchique (CAH)

La CAH est une méthode de classification non supervisée (clustering ou segmentation) qui procède par agglomération récursive des individus de l'échantillon.

Il en résulte une suite de partitions emboités, de la plus fine (elle est composé de plusieurs éléments (éléments qui contiennent chacun un singleton) à la plus grossière (c'est la partition à un élément (élément qui contient tous les éléments)).
Cette méthode de classification nécessite de se donner une mesure de la dissemblance entre les groupes d'individus.

2) Dissimilarité – Dissemblance

Notation : $x_i = (x_{ij})_{j=1}^{p}$

$$\boldsymbol{X = \{x_i\}_{i=1}^{n}}$$

$X(\Omega)$ =l'ensemble des valeurs possibles des x_i

Définition : Dissimilarité

$d: X(\Omega) \times X(\Omega) \to R$ vérifiant

i) $\boldsymbol{d(x, x') \geq 0}$

ii) $\boldsymbol{d(x, x') = 0 \leftarrow x = x'}$

iii) $\boldsymbol{d(x, x') = d(x', x)}$

Remarque :

1) toute distance est une dissimilarité.
2) Tout ce qui manque à la dissimilarité est l'inégalité triangulaire $d(x, x'') \leq d(x, x') + d(x', x'')$

Exemples :

1) $X(\Omega) = R^p$

$$d(x, x') = \sqrt{\sum_{j=1}^{p} (x_j - x'_j)^2}$$

$$d(x, x') = \sum_{j=1}^{p} (x_j - x'_j)^2$$

$$d(x, x') = \left(\sum_{j=1}^{p} (x_j - x'_j)^p \right)^{\frac{1}{p}} \; avec \; p > 1$$

2) $X(\Omega) = \{0,1\}$

$$d(x, x') = \sum_{j=1}^{p} 1_{\{x_j \neq x'_j\}}$$

3) Données qualitatives

Appliquer par exemple le codage disjonctif complet, puis considérer d précédent.

Exemple de codage disjonctif complet

X^1	X^2	X^3
a	**bb**	**ddd**
a	**cc**	**bbb**
b	**cc**	**ccc**
a	**bb**	**ddd**
b	**aa**	**aaa**

$X^{1,1}$	$X^{1,2}$	$X^{2,1}$	$X^{2,2}$	$X^{2,3}$				
1	**0**	**0**	**1**	**0**				
1	**0**	**0**	**0**	**1**				
0	**1**	**0**	**0**	**1**				
1	**0**	**0**	**1**	**0**				
0	**1**	**1**	**0**	**0**				

Définition (Dissemblance)

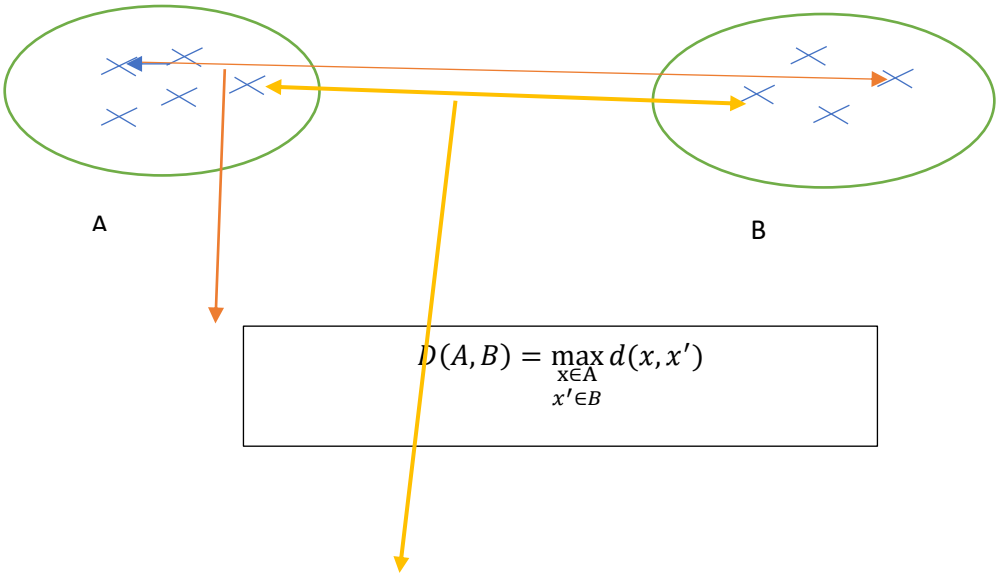
La notion de dissemblance permet de juger de la dissimilarité entre deux groupes ou deux classes d'individus. Pour la définir, on a besoin d'une dissimilarité.

Note : dissemblance ≡ ultra-métrique

Exemples :

Soit d une dissimilarité.

1) **Dissemblance lien complet(lien maximum)**



2) **Dissemblance lien simple(lien minimum)**

$$D(A,B) = \min_{\substack{x\in A \\ x'\in B}} d(x,x')$$

3) **Dissemblance de Ward**

Supposons que $x_i \in R^p, \forall i$

On considère $d(x,x') = \sqrt{\sum_{i=1}^{n}(x_j - x'_j)^2}$

Posons $\mu_A = \frac{1}{card\ A}\sum_{x\in A} x\ , \mu_B = \frac{1}{card\ B}\sum_{x\in B} x$

$$\boldsymbol{D(A,B) = \frac{1}{card\ A + card\ B} d(card\ A \times \mu_A, card\ B \times \mu_B)}$$

4) **Ward 2**

$$\boldsymbol{D(A,B) = \frac{1}{card\ A + card\ B} d^2(card\ A \times \mu_A, card\ B \times \mu_B)}$$

Note : $A = \{x\}$ et $B = \{x'\}$, on pose $D(A,B) = d(x,x')$

Comme propriétés des dissemblances, on peut noter que $D(A,A) = 0$ si A est un singleton.

Les dissemblances sont aussi à valeurs positives.

CAH(il faut arranger les numérotations des titres)

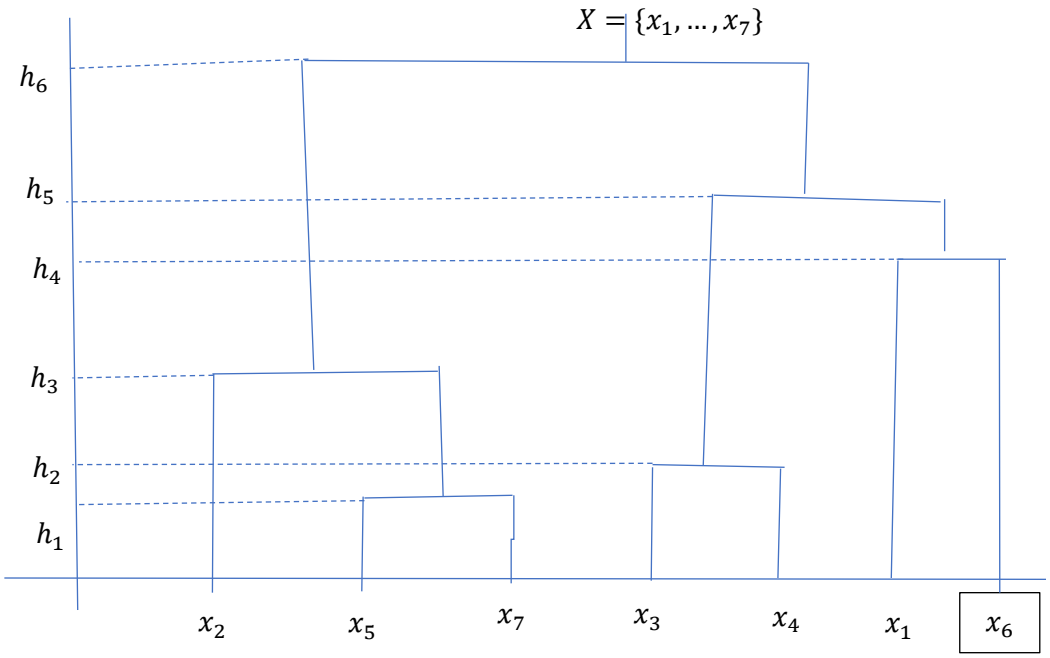
1) Hiérachie valuée(Dendogramme)

Les résultats d'une cah peuvent être représentés dans une hiérachie valuée en appelée **dendogramme.**

Note : Hiérachie ≡ arbre binaire.

Exemple :

$$\boldsymbol{X = \{x_1, x_2, \dots, x_7\}}$$



Quand la dissemblance est petite alors la similarité est grande.

1) Les nœuds sont sous-ensembles de X
2) Singleton ≡ feuille
3) Nœud intermédiare ≡ Nœud qui n'est pas une feuille
4) On a n-1 nœuds intermédiares où $n = card\ X$ =nbre d'individus

5) Fonction de coupure

La hiérachie (*) peut être représentée par

la **fonction de coupure :**

$\boldsymbol{c}$: $\boldsymbol{R^+}$ →**Ensemble des partitions de X**

$$\boldsymbol{c(h)} = \begin{Bmatrix} \{\boldsymbol{x_1}\}, \dots, \{\boldsymbol{x_7}\}\ \boldsymbol{si} \leq \boldsymbol{h} < \boldsymbol{h_1}\} \\ \{\{\boldsymbol{x_2}\}, \{\boldsymbol{x_5}, \boldsymbol{x_7}\}, \{\boldsymbol{x_3}\}, \{\boldsymbol{x_4}\}, \{\boldsymbol{x_1}\}, \{\boldsymbol{x_6}\}\ \boldsymbol{si}\ \boldsymbol{h_1} \leq \boldsymbol{h} < \boldsymbol{h_2} \\ \dots \\ \{\boldsymbol{X}\}\boldsymbol{si}\ \boldsymbol{h} \geq \boldsymbol{h_6} \end{Bmatrix}$$

6) $\boldsymbol{h_1} = \boldsymbol{D}(\{\boldsymbol{x_5}\}, \{\boldsymbol{x_7}\})$

7) Algorithme :

Données : $X = \{x_1, \dots, x_n\}$

Dissemblance D

Résultat : suite de partitions

Début :

$P^{(0)} \leftarrow \{\{x_1\}, \dots, \{x_n\}$

$k \leftarrow 0$

Répéter

$(A^*, B^*) = argmin_{A,B \in P^{(k)}} \{D(A, B)\}$

$P^{(k+1)} \leftarrow P^{(k)}$

Dans $P^{(k+1)}$, remplacer A^* et B^* par $A^* \cup B^*$

$k \leftarrow k + 1$

Jusqu'à $card\ P^{(k)} = 1$

Renvoyer $\left(P^{(k)}\right)_{k=0,\dots}$

Fin.

Quand on fait l'algo, il faut stocker les dissimilarités entre A* et B* car ce sont ces dissimilarités qui nous donnent les valeurs des h_i.

Il est préférable de calculer les dissemblances dans un tel tableau :

	$\{\boldsymbol{x_1}\}$	$\{\boldsymbol{x_2}\}$	$\{\boldsymbol{x_3}\}$	$\{\boldsymbol{x_4}\}$	$\{\boldsymbol{x_5}\}$	$\{\boldsymbol{x_6}\}$	$\{\boldsymbol{x_7}\}$
$\{\boldsymbol{x_1}\}$							
$\{\boldsymbol{x_2}\}$							
$\{\boldsymbol{x_3}\}$							
$\{\boldsymbol{x_4}\}$							

$\{x_5\}$							
$\{x_6\}$							
$\{x_7\}$							