# ANALYSE DE DONNEES

Johann PAVIS (selon source :)

# Table des matières

# Chapitre 1

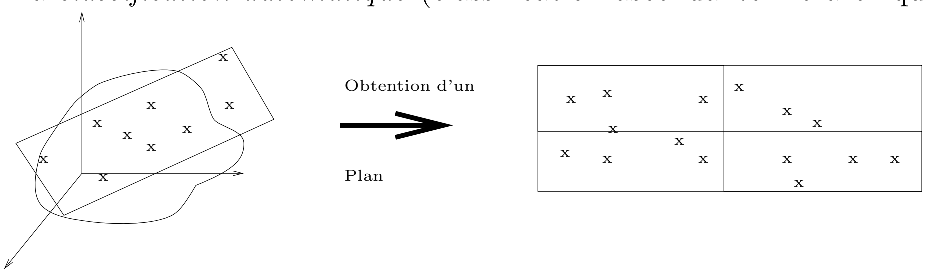
# INTRODUCTION A L'ANALYSE DE DONNÉES

L'analyse de données traite des données statistiques, souvent triées sous forme de tableaux, de dimension $n \times p$.

Il existe deux grandes méthodes pour réduire la taille des données à analyser :

- l'analyse *factorielle* (analyse en composantes principales, analyse factorielle des correspondances)

- la *classification automatique* (classification ascendante hiérarchique)




ANALYSE FACTORIELLE

Pour la classification, on mesure toutes les distances, puis on trace un arbre. Plus les individus sont éloignés, plus les palliés sont importants.

## 1.1 Retapes préliminaires

### 1.1.1 Construction du tableau de données

Il se fabrique en général à partir d'un questionnaire. Pour pouvoir le traiter en analyse de données, il faut que ce tableau soit **homogène**, c'est à dire qu'il ne doit contenir qu'un seul type de données.

### 1.1.2 Notion d'individu et de variable

**Notion d'individu** Un individu est un élément de la population $\Omega$ étudiée (cette population est supposée finie). Il correspond à une **ligne** du tableau.

**Notion de variable** A chaque paramètre choisi par l'utilisateur pour décrire les individus, on associe une ou plusieurs variables. Une variable est définie par les données suivantes :
- un ensemble O, appelé espace d'observation ;
- une structure S sur O ;
- une application v de $\Omega$ sur (O,S)

#### 1.1.2.1 Typologie de Diday

| | O continu | O fini/dénombrable | |
|---|---|---|---|
| O est sous structure $= \neq$ | | CSP, yeux, département | nominales |
| O sous structure d'ordre $\leq$ | Age, température | rang | ordonnables |
| O corps ordonnable | Revenu (mensuel, ...) | | mesurables |
| | quantitatives | qualitatives | |

#### 1.1.2.2 Typologie de Lerman

**Variables d'incidence**

**Attributs descriptifs**

$$x \in \Omega : v(x) = \begin{cases} 1, & \text{si x vérifie v} \\ 0, & \text{sinon} \end{cases}$$

Ceci permet parfois d'homogénéiser un tableau, en "éclatant"les individus.

**Variables numériques**

$$x \to v(x) \in \mathbf{R}$$

Elles peuvent servir pour la description des propriétés agricoles (surface agricole utile, ...), par exemple.

**Variables relationnelles**

**Ensemble de modalités totalement ordonné**

$$x \rightarrow v(x) = C_k$$

$$C_k \in \{C_0, C_1, C_2, ..., C_p\} \text{ où } \forall (i,j) \in [1,p]^2,\ i \leq j \Rightarrow C_i \leq C_j$$

**Ensemble de modalités sans relation d'ordre**

$$x \rightarrow C_k = v(x) \in \{C_0, C_1, ..., C_p\}$$

**Variable "rang"**
Note, sur une échelle suffisamment fine pour que 2 individus n'aient pas la même note.

**Variable mesure sur $\Omega \times \Omega$**
Elle peut être utile pour les exportations vers d'autres pays, par exemple.

### 1.1.3 Tableaux de données

Ils doivent être **homogènes**.

#### 1.1.3.1 Tableau de données quantitatives

| | Profondeur | | |
|---|---|---|---|
| | 5m | 10 m | 15 m |
| Sondage $S_1$ | | | |
| Sondage $S_2$ | | | |
| Sondage $S_3$ | | $x_2^3$ | |

#### 1.1.3.2 Tableau de contingence

Il s'agit d'un tableau d'effectifs (sans unité donc...).
En voici un exemple :

| | Alezan | Corsaire | Directoire | Orly |
|---|---|---|---|---|
| distingué | 14 | 12 | | |
| mièvre | | | | |
| ridicule | | | | |
| pour 1 femme | | | | |

Pour ce genre de tableaux, on utilise la méthode d'analyse factorielle des correspondances.

#### 1.1.3.3 Tableau d'incidence

Ce sont des tableaux de variables attributs descriptifs (département, de rérsidence, entre autres). Les tableaux étant en général très volumineux, on tente de les réduire (classe d'âge, région, etc...)

#### 1.1.3.4 Tableau de préférences

Voici un exemple, un échantillon de vins :

| | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ | $V_6$ | $V_7$ | $V_8$ | $V_9$ |
|---|---|---|---|---|---|---|---|---|---|
| n ° 1 | 1 | 3 | 2 | 7 | 9 | 4 | 6 | 5 | 8 |

#### 1.1.3.5 Tableau de modalités

Ils peuvent servir pour les enquêtes sur la fréquence de lecture de revues ou de journaux par exemple. Il peut, dans ces tableaux, y avoir *plusieurs fois* la même réponse sur une ligne.

#### 1.1.3.6 Tableau de proximité

Ils servent à traiter les mesures sur $\Omega \times \Omega$. Par exemple, pour les échnges internationnaux :

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $M_1$ | 20 | 3 | 5 | 10 |
| $M_2$ | | 20 | | |
| $M_3$ | | | 20 | |

## 1.2 Eléments descriptifs d'un tableau de données

### 1.2.1 Les notions

$n$ est le nombre d'individus, soit celui de lignes. $p$ est le nombre de variables, soit le nombre de colonnes.

Le tableau est souvent noté $X$, les valeurs pour l'individu $i$ pour la variable $j$ est noté $x_i^j$.

On note $\underline{x_i^j}$ l'individu correspondant à la ligne $i$. C'est le vecteur : $\begin{pmatrix} x_i^1 \\ \vdots \\ x_i^j \\ \vdots \\ x_i^p \end{pmatrix} \in$ $\mathbf{R}^p$

De même, on note $\underline{x^j}$ la variable correspondant à la ligne $j$. C'est le vecteur : $\begin{pmatrix} x_1^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{pmatrix} \in \mathbf{R}^n$

$\boxed{\mathbf{R}^p \text{ est l'espace des individus, } \mathbf{R}^n \text{ celui des variables.}}$

### 1.2.2 Eléments descriptifs

#### 1.2.2.1 Au niveau des variables

On peut considérer chaque variable comme une série statistique de n observations. On peut donc calculer sa moyenne, notée $\overline{x^j} = \sum_{i=1}^n p_i.x_i^j$ où $p_i$ est le poids de l'individu $i$ (en général, on a : $p_i = \frac{1}{n}$).

On peut également calculer :
- <u>la variance</u> :

$$Var(\underline{x}^j) = \sum_{i=1}^n p_i.(\underline{x}_i^j - \overline{x}^j)^2$$

- <u>l'écart-type</u> :

$$\sigma_{\underline{x}^j} = \sqrt{Var(\underline{x}^j)}$$

Pour deux variables, on peut calculer la covariance :

$$Cov(\underline{x}^j, \underline{x}^{j'}) = \sum_{i=1}^n p_i.(x_i^j - \overline{x}^j).(x_i^{j'} - \overline{x}^{j'})$$

De même, on calcule la corrélation :

$$Coor(\underline{x}^j, \underline{x}^{j'}) = \frac{Cov(\underline{x}^j, \underline{x}^{j'})}{\sigma_{\underline{x}^j}\sigma_{\underline{x}^{j'}}}$$

#### 1.2.2.2 Au niveau des individus

On peut considérer chaque individu comme une série statistiques de p observations. On peut donc en calculer la moyenne, l'écart-type, etc. Il y a cependant ici un problème de signification.

En revanche, on représente souvent le nuage des individus ($N_I$) comme un solide auquel il est assimilé. On peut donc en calculer le centre de gravité :

$$\overline{\underline{x}} = \sum_{i=1}^{n} p_i.x_i$$

Les coordonnées de ce centre de gravité (en développant) sont les moyennes des variables. On calcule également l'inertie de ce solide :

$$I = \sum_{i=1} np_i.d^2(\underline{x}_i, \overline{\underline{x}}$$

où d est une distance. En analyse de données, la distance dépend du nombre de variables.

## 1.3 Choix d'une mesure de ressemblance

#### 1.3.0.3 Quelques définitions

**Indice de similarité** Il s'agit d'une application s de $\Omega^2$ dans $\mathbf{R}^+$ telle que :

$$\begin{cases} \text{s est symétrique} \\ et \\ \forall(\omega,\omega^{'}) \in \Omega^2, \omega \neq \omega^{'} \Rightarrow s(\omega,\omega) = s(\omega^{'},\omega^{'}) \geq s(\omega,\omega^{'}) \end{cases}$$

**Indice de dissimilarité** Il s'agit d'une application de $\Omega^2$ dans $\mathbf{R}^+$ telle que :

$$\forall \omega \in \Omega, s(\omega,\omega) = 0$$

**Distance** Il s'agit d'un indice de dissimilarité vérifiant de plus :

$$\forall(\omega,\omega^{'},\omega'') \in \Omega^3, \begin{cases} s(\omega,\omega^{'}) = 0 \Leftrightarrow \omega = \omega^{'} & (a) \\ s(\omega,\omega^{''}) \leq s(\omega,\omega^{'}) + s(\omega^{'},\omega^{''}) & (b) \end{cases}$$

**Ultramétrique** Il s'agit d'une distance vérifiant (a) et :

$$\forall(\omega,\omega^{'},\omega^{''}) \in \Omega^3, s(\omega,\omega^{''} \leq max(s(\omega,\omega^{'}), s(\omega^{'},\omega^{''})$$

#### 1.3.0.4 Calcul de distance

Le choix de la distance dépend du type de variables étudiées.

**Variables quantitatives** - distance euclidienne
- distance du $\chi^2$ (tableau de contingence) :

$$d^2(\underline{x}_i, \underline{x}_{i'}) = \sum_{j=1}^{p} \frac{1}{x^j} (\frac{x_i^j}{x_i} - \frac{x_{i'}^j}{x_{i'}})^2$$

$$\text{avec : } x^j = \sum_{i=1}^{n} x_i^j \text{ et } x_i = \sum_{j=1}^{n} x_i^j$$

**Tableaux de contingence**

| | 1 | 0 |
|---|---|---|
| 1 | a | b |
| 0 | c | d |

a est le nombre de fois où $x_i^j = x_{i''}^j = 1$ ($\underline{x}_i$ est la ligne du haut, $\underline{x}_{i'}$ la colonne de gauche).

**Distance de Sokal et Michener**

$$d(\underline{x}_i, \underline{x}_{i'}) = \frac{a+b}{a+b+c+d}$$

**Distance de Russel et Rao**

$$d(\underline{x}_i, \underline{x}_{i'}) = \frac{a}{a+b+c+d}$$

**Tableaux de variables qualitatives ordinales**
Coefficient de corrélation sur rang de Sperman.

### 1.3.1 Mesures de ressemblance entre groupes d'objets

**Distance du lien maximum**

$$d(A, B) = max\{d(a, b)/a \in A, b \in B\}$$

**Distance du lien maximum**

**Distance des centres de gravité**

**Distance faisant intervenir l'inertie**

- $d^2(A,B) = I(A \cup B) - I(A) - I(B)$

$$(I(A) = \sum_{\omega \in A} p(\omega) d^2(\omega, g_A))$$

- $d^2(A,B) = \frac{I(A \cup B)}{p(A \cup B)} - \frac{I(A)}{p(A)} - \frac{I(B)}{p(B)}$

# Chapitre 2

# ANALYSE EN COMPOSANTES PRINCIPALES

## 2.1 Problème à traiter

Il s'agit de réduire un tableau de n lignes et p colonnes à un tableau de n lignes et q colonnes, avec $q < p$. On souhaite donc réduire le nombre de variables décrivant les n individus en perdant un minimum d'information. La technique la plus utilisée est la projection du nuage des individus sur un espace de dimension q, en minimisant la déformation.

## 2.2 Formulation mathématique

On note $W$ l'espace de projection.

Dans le cas où p et q sont quelconques, déterminer un sous-espace de $W$ revient à déterminer une origine $a \in \mathbf{R}^p$ et un système générateur libre $(\underline{u}_1, ..., \underline{u}_q)$.

On choisit $W$ tel que la déformation lors de la projection soit minimale. Les q nouvelles variables sont les composantes des projections orthogonales des points du nuage dans $W$.

On note : $\underline{\widehat{x}_i}$ le vecteur projeté sur $W$. Les nouvelles variables sont notées $y_i^j$. On a donc la relation suivante :

$$\underline{\widehat{x}}_i = \sum_{k=1}^{q} y_i^k . \underline{u}_k + \underline{a}$$

**Déformation**

La déformation est l'inertie autour de l'axe z, soit :

$$I_W = \sum_{i=1}^{n} p_i d^2(\underline{x}_i, \widehat{\underline{x}}_i)$$

## 2.3 Résolution du problème

### 2.3.1 Choix de $\underline{\mathbf{a}}$

**Proposition 1** *Toute solution $W$ du problème passe par $\underline{g}$, centre de gravité du nuage des individus.*

On prend donc $\underline{\mathrm{g}}$ comme origine pour $W$. On centre alors les données. On note alors les nouvelles variables :$\widetilde{\underline{x}}^i$.

### 2.3.2 Choix des vecteurs générateurs

**Proposition 2** *Soit $(\underline{u}_1, ..., \underline{u}_q)$ un système orthonormé relativement à une métrique $M$. Alors : $M \in \mathcal{D}_p(\mathbf{R})$.*

*- Si les variables sont dans une même identité, alors $M = I_p$*

*- Si les variables sont dans des unités différentes, alors $M = Var(\underline{x}^j)I_p$*

*Soit $W = Vect(\underline{u}_1, ..., \underline{u}_q)$ tel que ${}^t\underline{u}_k M \underline{u}_{k'} = \delta_{kk'}$. Alors on a :*

$$I_W = \sum_{i=1}^{n} p_i ||\underline{x}_i||^2 - \sum_{k=1}^{q} {}^t\underline{u}_k MVM\underline{u}_k$$

*$M$ étant la métrique et $V$ la matrice de variance / covariance.*

Si on veut minimiser $I_W$, il faut maximiser la seconde somme. On effectue un changement de variable : $\begin{cases} \underline{v}_k = M^{1/2}\underline{u}_k \\ S = M^{1/2}VM^{1/2} \end{cases}$

Minimiser la déformation globale revient donc ainsi à maximiser :

$$\sum_{k=1}^{q} {}^t\underline{v}_k S \underline{v}_k$$

S est diagonalisable dans $\mathbf{R}$. On note $\lambda_1, \ldots, \lambda_q$ ses valeurs propres. Soit $(\underline{w}_1, ..., \underline{w}_p)$ une base de vecteurs propres (de S).

On montre alors que le problème admet pour solution :

$$\begin{aligned} &\underline{v}_1 = \pm\underline{w}_1 \\ &\vdots \\ &\underline{v}_q = \pm\underline{w}_q \end{aligned}$$

## 2.4 Résultats

### 2.4.1 Axes principaux d'inertie

On a : $\underline{u}_k = n^{-1/2}\underline{v}_k$

Les vecteurs $\underline{u}_i$ sont donc les vecteurs propres de VM, associés aux valeurs propres $\lambda_i$ qui sont également valeur propre de $S = M^{1/2}.V.M^{1/2}$.

**Définition 3** *On appelle k-ème axe principal d'inertie l'axe engendré par* $\underline{u}_k$.

L'inertie expliquée est maximale pour le plan $[\underline{u}_1\ ,\ \underline{u}_2]$ (ensuite, elle diminue).

### 2.4.2 Les composantes principales

**Proposition 4** *On appelle k-ème composante principale le vecteur $\underline{y}^k$ dont les composantes sont les projections du nuage des n individus sur l'axe $[\underline{u}_k]$.*

*Les $\underline{y}^k$ ($k \in [1, q]$) sont les nouvelles variables à interpréter à l'aide des p premières variables.*

### 2.4.3 La valeur du critère $I_W$

On a :

$$I_W = \sum_{i=1}^{n} p_i ||\underline{x}_i||^2 - \sum_{k=1}^{q} \lambda_k$$

On suppose que l'on recherche un espace de projection $W_0$ de dimension p. On a alors : $I_W = 0$. Donc :

$$\sum_{i=1}^{n} p_i ||\underline{x}^i||^2 = \sum_{k=1}^{q} \lambda_k$$

Plus l'inertie expliquée est proche de 1, meilleure sera la représentation.

### 2.4.4 Interprétation des résultats : les individus

#### 2.4.4.1 Coordonnées

On représente en général les individus dans les plans d'inertie.

Les coordonnées de l'individu $\underline{x}_i$, projeté en $\widehat{\underline{x}_i}$ sont : $\underline{y}_i^k =^t \underline{x}_i M \underline{u}_k$ (sur $[\underline{u}_k]$).

#### 2.4.4.2 Qualité globale de la représentation

$$IE_{[\underline{u}_1,\underline{u}_2]} = \frac{\lambda_1 + \lambda_2}{tr(VM)}$$

#### 2.4.4.3 Qualité ponctuelle de représentation

$\rightarrow$ Indique si un point es bien représenté sur le graphique. Si cette qualité est voisine de 0, il ne faut pas interpréter ces points.

> On peut mesurer la qualité ponctuelle de la représentation de l'individu $\underline{x}_i$, sur le plan $[\underline{u}_k, \underline{u}_{k'}]$,
> en considérant le cosinus carré de l'angle $\alpha$ qu'il fait avec ce plan.
> Cette valeur est également appelée contribution relative du plan à l'individu $\underline{x}_i$ et est notée :
> $\rho_i = \cos^2\alpha = \frac{||\widehat{\underline{x}}_i||^2}{||\underline{x}_i||^2} = \frac{(y_i^k)^2+(y_i^{k'})^2}{||\underline{x}_i||^2}$.

Plus l'individu est proche du plan, plus l'angle sera faible et donc le cosinus carré proche de 1.

La contribution relative de l'axe $[\underline{u}_k]$ à l'individu $\underline{x}_i$ est : $\rho_i^k = \frac{(y_i^k)^2}{||\underline{x}_i||^2}$

#### 2.4.4.4 Contribution absolue d'un individu à l'axe

Elle sert à interpréter les nouvelles variables $(\underline{y}^k)$

On note $\alpha_i^k$ la contribution absolue de l'individu $\underline{x}_i$ à la détermination de l'axe $[\underline{u}_k]$ et on a :

$$\alpha_i^k = \frac{p_i(y_i^k)^2}{\sum_{i=1}^n p_i(y_i^k)^2}$$

C'est la part d'inertie d'un individu à l'inertie totale du nuage. C'est aussi l'influence prise par un individu pour l'orientation de l'axe. On ne retient que les individus ayant une contribution importante par rapport à l'axe $[\underline{u}_k]$.

Plus $\alpha_i^k$ est proche de 1, plus l'individu contribue à la détermination de l'axe.

### 2.4.5 Interprétation des résultats : les variables

**Proposition 5** *Les composantes principales sont des vecteurs orthogonaux de l'espace* $\mathbf{R}^n$

Les variables initiales sont projetées dans l'espace des composantes principales. Si on retient comme plan de projection des individus le plan $[\underline{u}_k\ ,\ \underline{u}_{k'}]$, on projettera les p variables initiales dans l'espace $[\underline{y}^k\ ,\ \underline{y}^{k'}]$.

La coordonnée de $\underline{x}^j$ sur l'axe $[\underline{y}^k]$ est le coefficient de corrélation entre $\underline{x}^j$ et $\underline{y}^k$. On interprétera la nouvelle variable $\underline{y}^k$ à l'aide des variables initiales qui lui sont le plus corrélées.

L'ensemble des p variables initiales se projettent dans une cercle de rayon 1 appelé cercle des corrélations. On a la relation suivante :

$$r_{\underline{x}^j,\underline{y}^k} = \frac{\sqrt{\lambda_k}.u_k^j}{\sigma_j}$$

## 2.5 Méthodologie

1. Calcul de la matrice S.
2. Calcul des valeurs propres de S.
3. On projette.
4. On regarde le graphique de projection des individus.
5. On regarde le $1^{er}$ axe de projection.
6. Pour chaque individu, on regarde le $\cos^2$.

# Chapitre 3

# ANALYSE FACTORIELLE DES CORRESPONDANCES

Méthode d'analyse factorielle, traitant de tableaux de contingence. Le but est d'empêcher le biais occasionné par les colonnes / lignes de poids trop fort.

## 3.1 Le but de la méthode

Comme en ACP (Analyse en Composantes Principales), on cherche à décrire les proximités entre points ligne d'une part, et point colonne d'autre part. La technique utilisée est similaire à celle employée en ACP. On fera cependant simultanément :

- une ACP des profils ligne
- une ACP des profils colonne

On aura donc deux matrices de données au départ.

## 3.2 La méthode

### 3.2.1 Transformation des données initiales : passage aux matrices de profil

On part de la matrice $Z = (n_{i,j} \in \mathcal{M}_{n,p}(\mathbf{R})$.

On a pose :

$$n_{i.} = \sum_{j=1}^{p} n_{ij} \text{ et } n_{.j} = \sum_{i=1}^{n} n_{ij}$$

$$n_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} n_{ij}$$

Le tableau des profils ligne est noté : $X_I = \left(\frac{n_{ij}}{n_{i.}}\right)_{(i,j)\in[1,n]\text{x}[1,p]} = \underline{x}_{Ii}$

Le tableau des profils colonne est noté : $X_J = \left(\frac{n_{ij}}{n_{.j}}\right)_{(i,j)\in[1,n]\text{x}[1,p]} = \underline{x}_{Jj}$

### 3.2.2 Les nuages $N_I$ et $N_J$

On dispose de deux tableaux à analyser : $X_I$ et $X_J$. Aux n lignes du tableau $X_I$ on associe le nuage : $N_I = \{\underline{x}_{Ii}, \frac{n_{i.}}{n_{..}}/i \in [1,n]\} \in \mathbf{R}^p$. Aux p lignes du tableau $X_J$, on associe le nuage : $N_I = \{\underline{x}_{Jj}, \frac{n_{.j}}{n_{..}}/j \in [1,p]\} \in \mathbf{R}^n$.

On va calculer $\widetilde{X}_I$ et $\widetilde{X}_J$

### 3.2.3 Choix des métriques : matrices $M_I$ et $M_J$

Comme en ACP, on fait intervenir les matrices $M_I$ et $M_J$. Pour les individus $N_I$, on a : $M_I = \left(\frac{n_{..}}{n_{.j}}\right) \in \mathcal{D}_p(\mathbf{R})$

La distance entre 2 points $\underline{x}_{Ii}$ et $\underline{x}_{Ii'}$ est donnée par :

$$d^2 = \sum_{j=1}^{n} \frac{n_{..}}{n_{.j}} . \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right) \text{(distance dîte du } \chi^2)$$

De la même façon on calcule $M_J = \left(\frac{n_{..}}{n_{i.}}\right) \in \mathcal{D}_n(\mathbf{R})$

### 3.2.4 Calcul de $\widehat{X}_I$ et de $\widehat{X}_J$

On calcule les deux centres de gravité $\underline{g}_I$ et $\underline{g}_J$, pour $N_I$ puis pour $N_J$ :

$$\underline{g}_I = \begin{pmatrix} \frac{n_{.1}}{n_{..}} \\ \vdots \\ \frac{n_{.p}}{n_{..}} \end{pmatrix} \qquad \underline{g}_J = \begin{pmatrix} \frac{n_{1.}}{n_{..}} \\ \vdots \\ \frac{n_{n.}}{n_{..}} \end{pmatrix}$$

### 3.2.5 Calcul de $V_I$ et de $V_J$

En ACP, on avait : $V = \frac{1}{n} {}^t\widetilde{X}\widetilde{X}$

Pour $N_I$, on pose : $D_I = \left(\frac{n_{i.}}{n_{..}}\right) \in \mathcal{D}_n(\mathbf{R})$

On a donc :

$$V_I = {}^t\,\widetilde{X}_I D_I \widetilde{X}_I \in \mathcal{M}_p(\mathbf{R}) = (v_{ij})$$

$$\text{où : } v_{jj'} = \sum_{i=1}^{n} \frac{n_{i.}}{n_{..}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n_{..}}\right)\left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{.j'}}{n_{..}}\right)$$

### 3.2.6 Recherche des axes principaux d'inertie

De la même façon qu'en ACP, on cherche les valeurs propres et les vecteurs propres de $C_I = M_I^{\frac{1}{2}}.V_I.M_I^{\frac{1}{2}}$ pour le nuage $N(I)$ (idem pour le nuage $N(J)$). On a les relations suivantes :

$$\begin{cases} c_{j,j'} = & \frac{n_{..}}{\sqrt{n_{.j}.n_{.j'}}}.v_{j,j'} \\ \quad\;\; = & \sum_{i=1}^{n} \frac{n_{ij}.n_{ij'}}{n_{i.}.\sqrt{n_{.j}.n_{.j'}}} - \frac{\sqrt{n_{.j}.n_{.j'}}}{n_{..}} \end{cases}$$

(analogue pour $c_{i,i'}$)

**Proposition 6** *On note $H = (h_{i,j})$ la matrice définie par :*

$$h_{i,j} = \frac{n_{ij}}{\sqrt{n_{i.}.n_{.j}}} - \frac{\sqrt{n_{i.}.n_{.j}}}{n_{..}}$$

*Alors, on a : $C_I = {}^tH.H$ et $C_J = H.{}^tH$.*

**Proposition 7** *$C_I$ et $C_J$ ont les mêmes valeurs propres.*

*De plus, si $\lambda$ est valeur propre simple positive et vérifiant*

$$\begin{cases} C_I.\underline{v_I} = \lambda.\underline{v_I} \text{ avec } {}^t\underline{v_I}.\underline{v_I} = 1 \\ C_J.\underline{v_J} = \lambda.\underline{v_J} \text{ avec } {}^t\underline{v_J}.\underline{v_J} = 1 \end{cases}$$

*Alors on a :* $\begin{cases} \underline{v_I} = \pm\frac{1}{\sqrt{\lambda}}{}^tH.\underline{v_I} \\ \underline{v_J} = \pm\frac{1}{\sqrt{\lambda}}{}^tH.\underline{v_J} \end{cases}$

On peut donc écrire les vecteurs propres de $C_J$ à l'aide de ceux de $C_I$.

**Conséquences pratiques** On ne fera les calculs que sur la matrice $C_I$. Par des formules de transpositions, on obtiendra les résultats de l'ACP sur $N(J)$ (les vecteurs propres de $C_J$ étant d'duits de ceux de $C_I$).

**Proposition 8** *Le vecteur de composantes $\left(\sqrt{\frac{n_{.1}}{n_{.p}}}\right)$ est vecteur propre de $C_I$ relativement à la valeur propre 0. Le vecteur de composantes $\left(\sqrt{\frac{n_{.1}}{n_{n.}}}\right)$ l'est de même pour $C_J$.*

**Conséquence** On ne retient pas la valeur propre 0, car l'inertie expliquée (IE) par l'axe correspond est nulle.

**Corollaire 9** *Soient les matrices $A_I$ et $T_I$ définies par : $C_I = A_I - T_I$, c'est à dire :*

$$a_{j,j'} = \sum_{i=1}^{n} \frac{n_{ij}.n_{ij'}}{n_{i.}.\sqrt{n_{.j}.n_{n.j'}}} \text{ et } t_{j,j'} = \frac{\sqrt{n_{.j}.n_{.j'}}}{n_{..}}$$

*Alors :*

1. *Tout vecteur propre de $C_I$ associé à la valeur propre $\lambda > 0$ est vecteur propre de $A_I$ pour la valeur propre $\lambda$.*
2. *Le vecteur propre $(\sqrt{n_{i.}})_{i\in[1,p]}$ de $C_I$ (relativement à $\lambda = 0$) est vecteur propre de $A_I$ relativement à 1.*

**Proposition 10** *Formules de passage entre $\underline{v_I}$ et $\underline{v_J}$.*
*Si $C_I.\underline{v_I} = \lambda.\underline{v_I}$, avec $\lambda \neq 0$, alors :*

$$\forall i \ , \ \underline{v_J}^i = \pm\frac{1}{\sqrt{\lambda.n_{i.}}}.\sum_{j=1}^{p}\frac{n_{ij}}{\sqrt{n_{.j}}}.\underline{v_I}^j$$

## 3.3 Les résultats

### 3.3.1 Les composantes principales

La *k-ième* composante principale de $N(I)$ sur le *k-ième* axe d'inertie est : $y_{Ii}^k = {}^t\underline{x_{Ii}}.M_I.\underline{u_{Ik}}$ (projection du *i-ème* individu de $N(I)$ sur $[\underline{u}_{Ik}]$).
De même, pour $N(J)$, on a : $y_{Jj}^k = {}^t\underline{x_{Jj}}.M_I.\underline{u_{Jk}}$.

**Remarque :** Contrairement aux résultats de l'ACP, en analyse factorielle des correspondances (AFC) on calcule les projections des points variable et des point individus (sur $W_I = \text{Vect}[\underline{u_{Ik}}, \underline{u_{Ik'}}]$).

**Remarque :** Entre $y_I^k$ et $y_J^k$, il existe une formule de passage. On peut montrer que les $y_I^k$ sont barycentres des $y_j^k$ ($k \in [1,p]0$. On peut alors faire coïncider les plans de projections.

### 3.3.2 Contributions absolue et relative

- $IE_{[\underline{u}_{Ik},\underline{u}_{Ik'}]} = \frac{\lambda_k+\lambda_{k'}}{\sum\lambda_i}$
- Contribution absolue de l'élément $\underline{x}_{Ii}$ à l'axe $[\underline{u}_{Ik}]$

$$\alpha_{Ik} = \frac{n_{i.}}{n_{..}}.\left(y_{Ik}^k\right)^2.\frac{1}{\lambda_k}$$

On peut donc déterminer l'axe soit en cherchant les points individus ayant une forte contribution absolue, soit en cherchant les points variables (de forte contribution absolue).

– Contribution relative de l'axe au point : $\cos^2$

$$\rho_{ik} = \frac{\left(y_{Ii}^k\right)^2}{||\underline{x}_I||_{\chi^2}^2} \text{ (pour } \underline{x}_{Ii}\text{, à l'axe } [\underline{u}_{Ik}])$$

De même pour les variables.

### 3.3.3 Représentation graphique

On a l'habitude de représenter les individus et les variables. On peut cependant ne représenter que les individus ou que les variables. On essaiera d'interpréter les relations entre ces variables.

Si on détermine les axes à l'aide des individus, on doit accrocher les variables (ou réciproquement).

# Chapitre 4

# CLASSIFICATION AUTOMATIQUE

## 4.1 But

Il s'agit ici de répartir une population en catégories :

– soit on obtient les catégories attendues

– soit les catégories attendues n'apparaissent pas

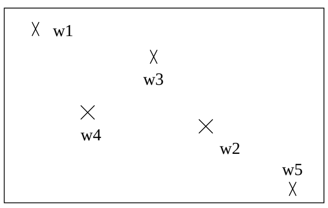
Dans le second cas, les variables choisies sot dites à faible pouvoir séparateur.

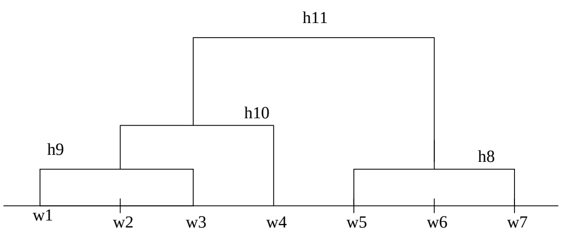
Si les catégories attendues apparaissent, il faut vérifier les hypothèses par d'autre(s) méthode(s).

## 4.2 Classification par hiérarchie et par arbre

### 4.2.1 Principe de la méthode

On cherche à représenter des points $w_1, \ldots, w_N \in \Omega^N$ par un ensemble de parties <u>*hiérarchiquement emboîtées*</u>.




L'interprétation de la hiérarchie est la suivante :

$\rightarrow$ chaque palier sous-tend un groupe de points

$\rightarrow$ la hauteur des paliers est une mesure du degré d'agrégation de la classe sous-tendue

### 4.2.2 Définition d'une hiérarchie

**Définition 11** *Soit $\Omega$ un ensemble fini. Soit $H$ un ensemble de parties de $\Omega$ appelées* ***paliers****.*

*Alors, $H$ est une hiérarchie ssi*

1. *$\Omega \in H$ : le palier le plus haut sous-tend l'ensemble de la population*
2. *$\forall \omega \in \Omega$ , $\{\omega\} \in H$ : les points terminaux de la hiérarchie contiennent des paliers*
3. *$\forall (h, h') \in H^2$ , $h \bigcap h' \neq \emptyset \Rightarrow (h \subseteq h') \vee (h' \subseteq h)$*

Dans la représentation précédente, on a les paliers suivants :
- $h_i = \{\omega_i\}$ , $i \in [1, 7]$
- $h_8 = \{\omega_5, \omega_6, \omega_7\}$
- $h_9 = \{\omega_1, \omega_2, \omega_3\}$
- $h_{10} = \{\omega_1, \omega_2, \omega_3, \omega_4\} = h_9 \bigcup h_4$
- $h_{11} = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7\} = \Omega$

Dans ce cas, on a : $H = \{h_i / i \in [1, 11]\}$.

Une hiérarchie est dîte **bianire** si chaque palier est la réunion des deux autres paliers de niveau(x) inférieur(s).

Une hiérarchie **indicée** est notée $(H, f)$ où $f$ est une application telle que :

1. $f : H \rightarrow \mathbb{R}^+$
2. $f(H) = 0 \Leftrightarrow Card(H) = 1$
3. $\forall (h, h') \in H^2$ , $h \subset h' \Rightarrow f(h) \subset f(h')$

**Définition 12** ***Indice de dissimilarité à partir d'une hiérarchie indicée***

*Un indice de dissimilarité est une application $s$ vérifiant :*

$$\begin{cases} s : \Omega^2 \rightarrow \mathbb{R}^+ \text{ et } s \text{ symétrique} \\ \forall \omega \in \Omega \ , \ s(\omega, \omega) = 0 \end{cases}$$

*On peut associer à une hiérarchie indicée $(H, f)$ un indice de dissimilarité :*

$$\begin{cases} \sigma : \Omega^2 \rightarrow \mathbb{R}^+ \\ \sigma(k, l) = min\{f(h)/(k, l) \in h^2\} \end{cases}$$

Pour construire une hiérarchie, on calcule la dissimilarité entre les points. On calcule ensuite les distances entre groupes de points, par l'intermédiare de *l'indice d'agrégation*.

**Définition 13** *On appelle indice d'agrégation une application $\delta$ qui vérifie :*

1. $\delta : \mathcal{P}^2 \to \mathbb{R}^+$
2. $\forall (h,h') \in \mathcal{P}^2$ , $\begin{cases} \delta(h,h') = \delta(h',h) \\ \delta(h,h') \geq 0 \end{cases}$

**Principaux indices d'agrégation utilisés**

- indice d'agrégation du lien maximum

$$\delta(a,b) = max\{(d(a,b))/(a,b) \in A \times B\}$$

- indice d'agrégation du lien minimum

$$\delta(a,b) = min\{(d(a,b))/(a,b) \in A \times B\}$$

- indice d'agrégation des centres de gravités

$$\delta(a,b) = d(G(A),G(B)) \text{ , avec } G(X) \text{ centre de gravité de } X \subset \Omega$$

- indice d'agrégation pour le couplage ACH et AFC (ou ACP)
  → indice d'agrégation de l'inertie de la réunion de 2 classes (Jambu, 1978)

$$\delta(A,B) = I\,(A \cup B) = I(A) + I(B) + \frac{p(A).p(B)}{p(A)+p(B)}.d^2(G(A),G(B))$$

  (cette agrégation utilise la distance du $\chi$-deux)
  → indice d'agrégation de la variance de la réunion de 2 classes

$$\delta(a,b) = Var\,(A \cup B) = \frac{I\,(A \cup B)}{p\,(A \cup B)}$$

### 4.2.3 Construction de hiérarchies indicées

Il existe deux grands types de classification hiérarchique :
→ classification hiérarchique descendante (non utilisée) On part de la population entière, que l'on coupe de manière à obtenir des classes de plus en plus petites, jusqu'à obtenir des classes réduites à un point.
→ classification hiérarchique ascendante
Algorithme

1. On part de la partition $P$ dont les classes sont réduites à un élément
2. On construit une nouvelle partition en regroupant les classes de partition précédentes, de manière à minimiser $\delta$

3. On réitère le procédé jusqu'à obtention d'une unique classe, contenant toute la population

On emploie principalement deux classifications hiérarchiques ascendantes :
→ l'agrégation de saut minimal (utilisation de l'indice d'agrégation du line minimum)
→ l'agrégation selon la variance (lors du couplage avec une AFC)

### 4.2.4 Classification hiérarchique ascendante de saut minimal et arbre des poids minimaux

Rappel : Un arbre est un *graphe connexe sans arc.*

Lorsque l'on construit une hiérarchie de saut minimum à partir de la table des distances, on ne connaît plus ces distances à la fin. On ne possède d'ailleurs plus d'information à ce sujet, *i.e.* on ne peut pas retrouver les distances inter points à partir de l'arbre.

On utilise alors une nouvelle distance : la distance ultramétrique.

**Définition 14** *$d$ est une ultramétrique si :*
- *$d$ vérifie les 2 premières propriétés des distances*
- $\forall (x, y, z) \in \Omega^3$ , $d(x, z) \leq max(d(y, z), d(x, y))$

L'ultramétrique ici obtenue est la plus grande ultramétrique inférieure à la distance. Elle est appele **sous-dominante**.

**Propriété 15** *Soient $d_1$ et $d_2$ deux ultramétriques.*
*Alors :* $d_1 \leq d_2 \Leftrightarrow \forall (x, y) \in \Omega^2$ , $d_1(x, y) \leq d_2(x, y)$

**Construction de l'arbre des poids minimaux** par l'algorithme de KRUCKAL

Soient $\Omega$ une population et $A$ un ensemble (vide au départ)

1. Ranger par croissant les arrêtes selon leur valeur
2. Mettre dans $A$ la première arrête de la liste
3. Mettre dans $A$ l'arrête suivante, sauf si un cycle est fermé
4. Si $(\Omega, A)$ est connexe, on s'arrête. Sinon, on revient en 3