RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

UNIVERSITÉ MOULOUD MAMMERI, TIZI-OUZOU
Faculté des Sciences
Département de Mathématiques

Mémoire de Master en Mathématiques Appliquées
Option : *Processus Aléatoires et Statistique de la Décision*

THÈME :

: ÉTUDE COMPARATIVE EN CLASSIFICATION NON-SUPERVISÉE

Présenté par :

TOUCHERIFTE Samira

Devant le jury d'examen composé de :

M^r M-A. BOUDIBA	Maître de Conférence A	Président
M^r M. MEHIRI	Maître Assistant A	Rapporteur
M^r H. FELLAG	Professeur	Examinateur
M^r Y. BERKOUN	Maître de Conférence A	Examinateur

Soutenu le 12 / 10 / 2011

Table des matières

Table des figures

Remerciements

Je tiens à exprimer ma gratitude à Monsieur MEHIRI Mohamed, pour avoir accepté de diriger ce mémoire tout en ayant l'obligeance de consacrer de son temps pour donner des conseils très utiles qui m'ont permis d'enrichir et de mener à bien ce travail.

Je tiens à remercier les jurés et je suis très honorée de leur présence et de leur bienveillance à lire et juger ce travail. J'exprime donc ma gratitude à :

M^r BOUDIBA Mohand-Arezki, pour l'honneur qu'il me fait pour sa participation au jury en qualité de président.

M^r FELLAG Hocine et M^r BERKOUN Youcef, pour l'honneur qu'ils me font en acceptants d'être membres de mon jury de mémoire en qualité des examinateurs.

Je remercie chaleureusement tous les membres de l'équipe Master Probabilités et Statistique de l'Université Mouloud MAMMERI de TIZI-OUZOU.

Je remercie également : mes parents, mes grands-parents, mes frères (Hacen, Rabah et Sofiane), mes sœurs (Faroudja, Sadia, Naima et surtout ma petite sœur Amel) et mes cousins (Ahmed et Brahim) qui n'ont pas cessé de nourrir ma motivation pour ce travail et qui m'ont assisté de leurs encouragements.

Finalement, que toutes celles et tous ceux qui -hélas nombreux pour les citer tous- m'ont aidée d'une manière ou d'une autre à la réalisation de ce travail soient ici remerciés.

Introduction Générale

Le seul moyen de faire une méthode instructive et naturelle, est de mettre ensemble les choses qui se ressemblent et de séparer celles qui diffèrent les unes des autres.

Georges BUFFON, Histoire naturelle, 1749.

Cette phrase du célèbre naturaliste et écrivain *Georges BUFFON* peut servir de définition générale à un modèle de classifcation.

Les modèles les plus classiquement utilisés en classiffication sont les partitions et les hiérarchies de parties. Dans les deux cas, les objets qui se ressemblent sont regroupés en classes (ou clusters). Pour les partitions, les classes sont deux à deux disjointes ; pour les hiérarchies, elles peuvent être emboîtées. Dans les deux cas, elles ne sont pas empiétantes, dans le sens où l'intersection de deux d'entres elles n'en produira jamais de troisième.

La notion de classification est essentielle en science car elle permet aux scientifiques de mettre de l'ordre dans l'information et les connaissances qu'ils ont sur le monde. Aussi, depuis longtemps des scientifiques et des chercheurs de divers bords ont essayé de classer des espèces (animales ou autres), et -plus généralement- des données de diverses natures.

De nombreuses classifications ont été créées. Face à celles-ci, les scientifiques sont souvent incapables de désigner la meilleure d'entre elles, c'est à dire celle qui a une prévalence sur toutes les autres pour tout ensemble de données ; car chacune présente un intérêt, au moins légèrement supérieur, par rapport à d'autres et en fonction de la t$\hat{a}$che considérée.

Le terme "*classification*" est associé à la notion d'abstraction. En effet, une classification permet de synthétiser des informations dans des groupes ou ensembles de données très généraux ; c'est une forme d'abstraction dans le sens où l'on va mettre de côté les descriptions exactes des objets et ne faire ressortir que les traits particuliers que certains d'entre eux ont en commun.

L'importance de la classification dans les sciences se reflète dans la grande variété des domaines où tant leur nature que leur construction ont fait l'objet de recherches ; on citera à ce propos SOKAL [1963], BONNER[1964], FORGY [1965], Mac QUEEN[1967], LANCE et WILLIAMS [1967], DIDAY [1971], BENZECRI [1973], GORDON [1987], CELEUX et al. [1989], ...

Dans le cadre d'un problème de classification, on dispose d'un ensemble de données qui représente une collection d'individus (objets) Les classes sont encore inexistantes. L'objectif est alors d'obtenir des classes d'objets homogènes, en favorisant l'hétérogénéité entre ces différentes classes.

Pour cela, la définition de la "*classification*" amène à se poser les questions suivantes :

- Comment les objets à classer sont-ils définis ?
- Comment définir la notion de ressemblance (dissemblance) entre objets ?
- Comment sont structurées les classes (clusters) ?
- Comment préférer une classification par rapport à une autre ?

Dès le départ il est nécessaire de différencier la classification non supervisée et la classification supervisée ou analyse discriminante. La classification supervisée consiste à construire des règles de décision en se basant sur un ensemble de données pour lesquelles les étiquettes des classes sont connues a priori. Le but de la classification non supervisée est de trouver une organisation des données cohérente et valide, qui puisse mettre en évidence les vraies structures dans un ensemble de données sans aucune connaissance a priori sur les données traitées.
Parmi les différentes méthodes, on peut considérer deux grands types d'approches :

1. **Non-paramétriques :** Les approches dites non-paramétriques (classification hiérarchique, méthode des centres mobiles) ne considèrent qu'une seule hypothèse : plus deux individus sont proches, plus ils ont

de chance de faire partie de la même classe.

2. **Probabilistes :** Les approches dites probabilistes utilisent une hypothèse sur la distribution des individus à classifier. Par exemple, on peut considérer que les individus de chacune des classes suivent une loi normale. Le problème qui se pose alors est de déterminer quels sont les paramètres de la loi et à quelles classes les individus ont le plus de chances d'appartenir.

Dans notre travail, nous nous intéressons *exclusivement* aux méthodes non-paramétriques qui sont des méthodes de classification automatique qu'on appelle aussi classification non supervisée.

Ce mémoire est organisé en quatre (4) chapitres de la manière suivante :

Dans le premier chapitre, nous définissons et donnons quelques types de déscrétisations ; on en donnera certaines caractéristiques, et l'on rappellera quelques méthodes de calcul du nombre de classes (méthodes de la racine carrée, de SCOTT, ... etc).

Dans le deuxième chapitre, on présentera les méthodes classiques de classification automatique utilisées en analyse de données dans le cadre non-paramétrique ; nous reviendrons -au préalable- sur quelques rappels concernant des notions telles que : tableaux de données ou de contingence, tableau disjonctif, matrices de distance, de variance-covariance et de corrélation, inerties et mesures de ressemblance.

Le troisième chapitre étudie en détail quelques indices de validation des classes (clusters), ainsi que d'autres (indices) qui serviront à comparer des partition sur un même ensemble d'individus.

Dans le chapitre quatre, après une brève présentation du logiciel R, on fait une (re)présentation des données ; la représentation graphique en dendogramme nous aidera à sélectionner le critère d'agrégation utilisé. Après une première analyse (une ACP), on appliquera la méthode ascendante hiérarchique qui nous donne un nombre optimal de clusters auxquels on applique la méthode des k-means ; après calcul de quelques indices, on détermine après comparaison la partition optimale.

Enfin, on va terminer avec une conclusion générale et essayer d'indiquer quelques perspectives et voies de recherche dans le vaste domaine que reste la classification en analyse statistique des données.

Chapitre 1

La Discrétisation

Dans ce chapitre, on s'intéresse -dans le cadre unidimensionnel- à classer un ensemble de données de manière que le résultat soit aussi clair et aussi fiable que possible.

Une grande utilisation de cet outil qu'est la discrétisation (classfication uni-dimensionnelle) est du ressort de la cartographie, où on essaie de donner une représentation graphique (choix de l'échelle, des figures, de l'implantation des variables visuelles, ... etc). Un compromis à ces représentations consiste, justement, à traiter-analyser ces données par diverses méthodes de classification. Ceci permet d'obtenir une forme simplifiée des données initiales ; cette simplification est -quelquefois- appelée *discrétisation.*

1.1 Définition [14]

La discrétisation est l'opération qui permet de découper en classes une série de données.

Cette opération simplifie l'information en regroupant les objets présentant les mêmes caractéristiques en classes distinctes.

Ainsi, la discrétisation doit conserver au mieux l'information tout en la simplifiant. Et, pour satisfaire ces deux (2) conditions, il faut respecter certaines règles :

1. Indiquer les descripteurs clés de la distribution d'une série statistique,

tels que le mode, la moyenne, le minimum/maximum et l'écart type.

2. Respecter la forme de la distribution (la plupart des séries de données ressemblent à un type de distribution connue : uniforme, symétrique -type gaussien, asymétrique). La méthode de discrétisation tiendra compte du modèle supposé de la distribution.
3. Appliquer le principe de ressemblance/dissemblance. Pour ce faire, une méthode de discrétisation s'efforce au mieux de minimiser la variance intra-classe tout en maximisant la variance inter-classe.

Définition 1. ***Coefficient d'asymétrie (Skewness)***
Le coefficient d'asymétrie est un moment standarisé qui mesure l'asymétrie de la densité de probabilité d'une variable aléatoire réelle.
L'asymétrie est le 3^{eme} *moment standarisé, il est défini et se note par :*

$$\gamma = \frac{\mu_3}{\sigma^3} = \frac{E[(X-\mu)^3]}{\sigma^3}$$

où $\mu_3 = E[(X-\mu)^3]$ *est le* 3^{eme} *moment centré et* σ *est l'ecart type.*

En pratique, nous utilisons un coefficient d'asymétrie sans biais :

$$\gamma_c = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (\frac{x_i - \overline{x}}{\sigma})^3.$$

Cette fonction γ *ainsi définie caractérise le degré d'asymétrie d'une distribution par rapport à sa moyenne : une asymétrie positive indique une distribution unilatérale décalée vers les valeurs les plus à droite, tandis qu'une asymétrie négative indique une distribution unilatérale décalée vers les valeurs les plus à gauche.*

Définition 2. ***Coefficient d'aplatissement (Kurtosis)***
Ce coefficient est une mesure d'aplatissement de la distribution d'une variable réelle.
On le définit pour une variable aléatoire X *d'espérance* μ *et d'ecart type* σ *comme suit :*

$$\beta = E[(\frac{X-\mu}{\sigma})^4]$$

lorsque μ *existe. Ce coefficient est très utile quand nous avons affaire à des échantillons ou répartitions symétriques. Dans le cas où* $\beta = 0$*, nous disons*

que la répartition des observations est de type gaussien.

En pratique, on utilise un coefficient d'aplatissement sans biais :

$$\beta_c = \{(\frac{n(n+1)}{(n-1)(n-2)(n-3)}) \sum_{i=1}^{n} (\frac{x_i - \overline{x}}{\sigma})^4\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Le coefficient d'aplatissement caractérise la forme du pic ou l'aplatissement relatifs d'une distribution comparativement à une distribution normale standard : un coefficient positif indique une distribution relativement pointue, tandis que s'il est négatif, nous avons une distribution relativement aplatie.

1.2 Les méthodes de discrétisation [15]

Avant de citer les différentes méthodes de discrétisation, on doit d'abord donner quelques règles à respecter pour réaliser une discrétisation correcte :

1. Aucune classe ne doit être vide. En général, cette règle est facile à respecter, sauf dans le cas de séries fortement dissymétriques (à coefficient d'asymétrie élevé !).

2. Les limites des classes extrêmes doivent couvrir l'ensemble du domaine de variation de la variable (la série). Il y a nécessité d'avoir une ressemblance des classes. Les limites de classes ne doivent pas laisser de valeurs en dehors du champ couvert par les bornes des classes.

3. Les limites de classes ne doivent pas se chevaucher : une valeur ne doit appartenir qu'à une seule classe.

A présent, on cite quelques méthodes de discrétisation :

1.2.1 Discrétisation par équidistance

Appelée aussi discrétisation en classes d'amplitude égale, cette méthode est très facile à réaliser et à implémenter, mais risque de donner lieu à des classes vides, notamment quand la distribution est asymétrique.

1.2.2 Discrétisation par progression arithmétique

Contrairement à la discrétisation par équidistance, dans cette méthode, l'amplitude des classes augmente selon une progression arithmétique de rai-

son r telle que :

$$r = \frac{e}{k}$$

où $e = max - min$ et k est le nombre de classes.
Cette méthode est conçue pour les distributions *asymétriques*.

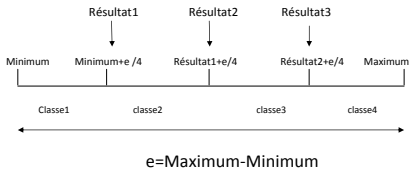

FIG. 1.1 – Discrétisation par progression arithmétique

1.2.3 Discrétisation selon les moyennes emboîtées

La moyenne arithmétique dans cette méthode divise la série de données en deux classes, et leur moyennes respectives divisent à leur tour en deux classes, et ainsi de suite, ce qui donnera un nombre de classes qui est toujours une puissance de 2. Cette méthode peut être utilisée avec n'importe quelle forme de distribution, mais présente l'inconvénient d'une restriction quant au nombre de classes.

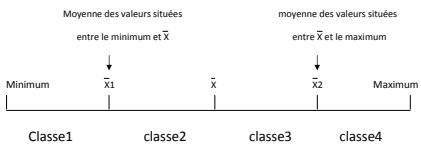

FIG. 1.2 – Discrétisation selon les moyennes emboîtées

1.2.4 Discrétisation standardisée

Dans cette méthode, on calcule la moyenne arithmétique et l'écart type en tant qu'indicateurs de l'amplitude des classes. Si le nombre de classes est pair, la moyenne sépare les deux classes centrales. Si le nombre de classes est impair, la moyenne est au centre de la classe centrale. L'amplitude des classes est généralement (proche) d'un écart type.

Cette méthode est d'autant plus adaptée que la distribution des valeurs de la série est proche de la distribution normale.
Si la distribution des valeurs n'est pas (proche de la loi) normale -mais pourrait la devenir en tenant compte d'un facteur de progression (lognormal ou autre), il convient de l'appliquer. Dans le cas contraire, il est possible de la normaliser.

Cette méthode est très utilisée en géographie. Elle a l'intérêt supplémentaire de produire des classes d'égale amplitude.

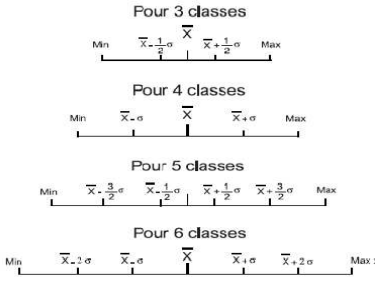

FIG. 1.3 – Discritésation standardisée

1.2.5 Discrétisation selon la méthode de JENKS

Elle se base sur le principe de ressemblance/dissemblance en calculant les distances entres toutes les paires de valeurs de la série. La méthode minimise la variance intra-classe et maximise la variance inter-classe. Il s'agit d'un algorithme itératif formant, dans un premier temps, autant de couples qu'il y a de combinaisons de valeurs (triées par ordre croissant) pour un nombre de classes donné. On calcule alors les variances intra-classe et inter-classe.
Cette méthode est certainement la plus adaptée des méthodes de discrétisation, même si sa mise en œuvre nécessite l'emploi d'un logiciel statistique.

Remarque

Il existe d'autres méthodes comme la discrétisation par équi-fréquence et manuelle.

Avant de penser à appliquer l'une des méthodes qu'on a déjà citées, la question de savoir combien de classes créer se pose d'elle-même.

1.3 Le nombre de classes

Le nombre de classes doit évidemment être en rapport avec le nombre d'individus de la série de données.
Si on veut mathématiquement définir le nombre de classes en fonction de la taille de la série, il existe plusieurs méthodes dont les plus utilisées sont :

Méthode	**Formule de k**
La racine carrée	$\sqrt{n}$
Brooks-Carruthers	$5log_{10}(n)$
Yule	$2,5\sqrt[4]{n}$
Huntesberger	$1+3,3log_{10}(n)$
Sturges	$log_2(n+1)$
Scott	$(b-a)/(3,5.S.n^{\frac{-1}{3}})$
Freedman-Diaconis	$(b-a)/(2.eiq.n^{\frac{-1}{3}})$

Table 1.1 Nombre de classes

où k est le nombre de classes (entier le plus proche),
n est le nombre d'individus,
a et b sont le minimum et le maximum de la série de données,
s est l'ecart type et eiq désigne l'ecart inter-quartile.

Définition 3. *L'intervalle inter-quartile est l'intervalle* $[Q_1; Q_3]$
L'écart inter-quartile est le nombre $Q_3 - Q_1$. *C'est la longueur de l'intervalle inter-quartile.*

On peut aussi utiliser l'inter-décile.

Définition 4. *L'intervalle inter-décile (ou l'intervalle de* KELLEY *est l'écart entre le dernier et le premier déciles* $(D_9 - D_1)$ *(cet intervalle contient 80% des observations et exclue les 20% de valeurs extrêmes ; il exprime la dispersion autour de la médiane. Il est très utilisé en granulométrie où on évite les deux extrêmes qu'on estime être dûs à des phénomènes exceptionnels ou aléatoires.*
On utilise cet intervalle dans les séries à effectifs élevés où le décile a un sens significatif.

Chapitre 2

Les Méthodes De Classification

2.1 Introduction

Ce chapitre est basé sur les méthodes de classification comme les autres méthodes de l'Analyse des Données (*Cluster Analysis*) dont elle fait partie.

La classification a pour but d'obtenir une représentation schématique simple d'un tableau de données dont les colonnes -suivant l'usage- sont des descripteurs (variables) de l'ensemble des observations placées, elles, en lignes.

L'objectif le plus simple d'une classification est de répartir l'échantillon en groupes (clusters) d'individus homogènes, chaque groupe étant bien différencié (séparé) des autres.

Dès les premières tentatives de classification, il s'est posé le problème du nombre de classes, de leur validation. Les questions sont aussi simples que les réponses sont complexes.

Il y a deux grands types de méthodes de classification :

1. Classification non hiérarchique (partitionnement). Il s'agit de décomposer l'espace des individus en classes disjointes.
2. Classification hiérarchique. A chaque étape du processus itératif, on a une décomposition de l'espace des individus en classes disjointes.

Avant de présenter les différentes méthodes de classification, on rappellera certaines notions utiles en classification (inertie, variance, et mesures de

ressemblance, ... etc).

Définition 5. *La classification (*clustering*) est l'opération statistique qui consiste à regrouper des objets (individus ou variables) en un nombre limité (réduit) de groupes (classes) qui ont les deux propriétés suivantes :*

1. *ils ne sont pas prédéfinis par l'expert, mais découverts au cours de l'opération ;*
2. *ils regroupent des objets ayant des caractéristiques similaires et séparent ceux ayant des caractéristiques différentes (homogénéité interne et hétérogénéité externe), ce qui peut être mesuré par des critères tels que les inerties inter-classe et intra-classe, ou les variances correspondantes.*

2.2 Généralités [5]

2.2.1 Tableau de données

On suppose que l'on dispose des mesures de p variables quantitatives sur n individus. Les valeurs (numériques) sont ainsi rangées dans un tableau à n lignes et p colonnes qu'on appelle aussi "tableau *individus*$\times$*variables*", et on note X la matrice (rectangulaire) associée à ce tableau :

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & x_i^j & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$$

où x_i^j est la valeur prise par la variable j sur l'individu $i, (i = 1, \dots, n \, et \, j = 1, \dots, p)$.

La variable j sera alors identifiée par le vecteur $\mathbf{x}^j = \begin{pmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{pmatrix}$

2.2.2 L'espace des individus

Chaque individu étant un point défini par p coordonnées comme un éléments d'un espace vectoriel E appelé l'espace des individus. L'ensemble des

n individus est alors un nuage de points dans E et g en est le Centre de gravité.
L'space E est muni d'une structure euclidienne afin de pouvoir définir des distances entre individus. Un individu i sera identifie par le vecteur $\mathbf{e}_i = (x_i^1, x_i^2, \ldots, x_i^p)$

2.2.3 Tableaux de contingence et disjonctif

Tableau de contingence

Soient $\mathfrak{X}$ et $\mathfrak{Y}$ deux variables qualitatives à r et s catégories respectivement décrivant un ensemble de n individus. On présente usuellement les données sous la forme d'un tableau croisé appelé *tableau de contingence* à r lignes et s colonnes renfermant les effectifs n_{ij} d'individus tels que $\mathfrak{X} = x_i$ et $\mathfrak{Y} = y_j$

$\mathfrak{X}\backslash\mathfrak{Y}$	y_1	y_2	$\cdots\cdots\cdots$	y_s	Total
x_1	n_{11}	n_{12}		n_{1s}	$\mathbf{n_{1.}}$
x_2	n_{21}	n_{22}		n_{2s}	$\mathbf{n_{2.}}$
$\vdots$					
x_i	n_{i1}	n_{i2}			$\mathbf{n_{i.}}$
$\vdots$					
x_r	n_{r1}	n_{r2}		n_{rs}	$\mathbf{n_{r.}}$
Total	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$		$\mathbf{n_{.s}}$	$\mathbf{n_{..} = n}$

Table 2.1 Tableau de contingence

Dans ce tableau, $\mathbf{n_{i.}} = \sum_j n_{ij}$ et $\mathbf{n_{.j}} = \sum_i n_{ij}$ désignent les totaux marginaux lignes et colonnes respectivement, appelés aussi marges lignes et marges colonnes.
On appelle tableau des **profils-lignes** le tableau des fréquences conditinnelles $n_{ij}/\mathbf{n_{i.}}$ et tableau des **profils-colonnes** le tableau des fréquences conditionnelles $n_{ij}/\mathbf{n_{.j}}$.

Tableau disjonctif

Soit X un tableau à n lignes et p colonnes. Les éléments de ce tableau sont des codes arbitraires sur lesquels aucune opération arithmétique n'est licite.

La forme mathématique utile pour les calculs est alors le tableau disjonctif des indicatrices des p variables obtenu en juxtaposant les p tableaux d'indicatrices de chaque variable $\mathfrak{X}_i$.

Exemple

Soit le tableau brut suivant :

$$\left|\begin{array}{ccc} \mathfrak{X}_1 & \mathfrak{X}_2 & \mathfrak{X}_3 \\ 1 & 2 & 3 \\ 2 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 2 & 1 \\ 3 & 1 & 2 \end{array}\right|$$

correspondant à 5 observations de trois variables $\mathfrak{X}_1, \mathfrak{X}_2 et \mathfrak{X}_3$ à 3, 2 et 3 catégories respectivement. On construit le **tableau disjonctif** X (à 5 lignes et 8 colonnes) suivant :

$$X = (\mathfrak{X}_1|\mathfrak{X}_2|\mathfrak{X}_3) = \left|\begin{array}{ccc|cc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{array}\right|$$

La somme des éléments de chaque ligne de X est égale au nombre p de variables ;
la somme des éléments d'une colonne de X donne l'effectif marginal de la catégorie correspondante.

Propsition 1. *Le rang de X est donné par :* $\sum_{i=1}^{p} m_i - p + 1$

2.2.4 La matrice de poids affectés aux individus

Afin de calculer la distance entre deux variables, il est parfois nécessaire d'attribuer des poids p_i aux individus selon l'importance que l'on souhaite leur donner.

On notera la matrice des poids :

$$\boldsymbol{D} = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & p_n \end{pmatrix}$$

où les p_i vérifient $\sum_{i=1}^{n} p_i = 1$
Souvent, on prend $\boldsymbol{D} = \frac{1}{n} I_n$, où I_n est la matrice identité (ce qui donne la même importance à chaque individu).

2.2.5 Les matrices de variance-covariance et de corrélation

- La matrice variance-covariance est définie comme suit :

$$V = X^{'} \boldsymbol{D} X - g g^{'}$$

où $\boldsymbol{D}$ est la matrice donnée précedement.

$$V = \begin{pmatrix} S_{11}^2 & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & . \\ S_{p1} & \dots & . & S_{pp}^2 \end{pmatrix},$$

où

$$S_{jj'} = cov(\mathbf{x}^j, \mathbf{x}^{j'}) = \frac{1}{n} \sum_{i=1}^{n} x_i^j x_i^{j'} - \overline{\mathbf{x}^j} \overline{\mathbf{x}^{j'}}$$

- La matrice regroupant tous les coefficients de coorélation linéaire entre les p variables prises deux à deux est notée R :

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & . \\ r_{p1} & \dots & . & 1 \end{pmatrix},$$

où $r_{jj'} = \frac{S_{jj'}}{S_j S_j'}$

2.2.6 La matrice des données centrées réduites

La matrice des données centrées et réduites est donnée par $Z = (z_i^j)_{1<i<n,1<j<p}$ avec :

$$z_i^j = \frac{x_i^j - \overline{\mathbf{x}^j}}{S_j}$$

2.2.7 La matrice des distances

Les distances entre individus, calculées par paires, sont répertoriées dans une matrice notée $D = (d_{ij})_{1<i,j<n}$, de taille $n \times n$, telle que :

$$d_{ij} = d(i, j)$$

Remarquons que seuls $\frac{n(n-1)}{2}$ termes sont significatifs, étant donné que la matrice est symétrique $(d_{ij} = d_{ji})$, et que les termes sur la diagonale sont nuls (la distance entre un individu et lui-même est nulle !).

2.2.8 Le Centre de gravité

On appellera Centre de gravité associé à la matrice des poids $\boldsymbol{D}$ le vecteur $\mathbf{g}$ défini par :

$$\mathbf{g} = \begin{pmatrix} \overline{\mathbf{x}^1} \\ \overline{\mathbf{x}^2} \\ \vdots \\ \overline{\mathbf{x}^p} \end{pmatrix}$$

où

$$\overline{\mathbf{x}^j} = \sum_{i=1}^{n} p_i x_i^j$$

2.2.9 La variance et l'écart type

Ce sont les deux mesures les plus fréquement utilisées.
La variance $S^2(\mathbf{x}^j)$ est définie par :

$$S^2(\mathbf{x}^j) = \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{\mathbf{x}^j})^2 \; ou \; \sum_{i=1}^{n} p_i(x_i^j - \overline{\mathbf{x}^j})^2,\; j = 1, \ldots, p.$$

L'écart type $S(\mathbf{x}^j)$ s'exprime dans la même unité que la variable étudiée. On a la formule suivante :

$$S_{(\mathbf{x}^j)} = \sqrt{S^2(\mathbf{x}^j)}$$

Théorème 1. ***Théorème de König-Huyghens***
La variance est égale à :

$$S^2(\boldsymbol{x}^j) = \frac{1}{n}\sum_{i=1}^{n}(x_i^j)^2 - (\overline{\boldsymbol{x}^j})^2$$

Remarque

Cette formule est utilisée pour le calcul de la variance : c'est la différence entre la moyenne des carrés et le carré de la moyenne.

2.2.10 Variances inter-classe et intra-classe

Soit V la matrice de variance-covariance d'un tableau de données X. Elle peut être décomposée en une somme de deux matrices :

$$V = B + W,$$

avec $B = \frac{1}{n}\sum_{k=1}^{q} n_k(g_k - g)'(g_k - g)$ et $W = \frac{1}{n}\sum_{k=1}^{q} n_k V_k$ les matrices de variance inter-classe et intra-classe respectivement, où g est le centre de gravité global, g_k est le centre de gravité de la classe C_k et V_k est la matrices variance-covariance des classes C_k.

En effet, la variance empirique totale s'écrit :

$$\begin{aligned}
S^2 &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^{'}(x_i - x_j) \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - g + g - x_j)^{'}(x_i - g + g - x_j) \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}((x_i - g) + (g - x_j))^{'}((x_i - g) + (g - x_j)) \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[((x_i - g)^{'} - (x_j - g)^{'})((x_i - g) - (x_j - g))] \\
&= \frac{1}{2n^2}[2n\sum_{i=1}^{n}(x_i - g)^{'}(x_i - g)] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - g)^{'}(x_i - g) \\
&= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C_k}(x_i - g)^{'}(x_i - g).....(*)
\end{aligned}$$

Donc

$$\begin{aligned}
\sum_{i\in C_k}(x_i - g)^{'}(x_i - g) &= \sum_{i\in C_k}(x_i - g_k + g_k - g)^{'}(x_i - g_k + g_k - g) \\
&= \sum_{i\in C_k}[((x_i - g_k)^{'} + (g_k - g)^{'})((x_i - g_k) + (g_k - g))] \\
&= \sum_{i\in C_k}[(x_i - g_k)^{'}(x_i - g_k) + (g_k - g)^{'}(g_k - g)] \\
&= \sum_{i\in C_k}(x_i - g_k)^{'}(x_i - g_k) + n_k(g_k - g)^{'}(g_k - g)
\end{aligned}$$

où $n_k = card(C_k)$
Et donc la variance totale s'écrit

$$\begin{aligned}
(*) &= \frac{1}{n}[\sum_{k=1}^{q}\sum_{i\in C_k}(x_i - g_k)^{'}(x_i - g_k) + \sum_{k=1}^{q} n_k(g_k - g)^{'}(g_k - g)] \\
&= \frac{1}{n}[\sum_{k=1}^{q} n_k \frac{1}{n_k}\sum_{i\in C_k}(x_i - g_k)^{'}(x_i - g_k) + \sum_{k=1}^{q} n_k(g_k - g)^{'}(g_k - g)] \\
&= \frac{1}{n}\sum_{k=1}^{q} n_k[\frac{1}{n_k}\sum_{i\in C_k}(x_i - g_k)^{'}(x_i - g_k)] + \frac{1}{n}\sum_{k=1}^{q} n_k(g_k - g)^{'}(g_k - g) \\
&= \frac{1}{n}\sum_{k=1}^{q} n_k V_k + \frac{1}{n}\sum_{k=1}^{q} n_k(g_k - g)^{'}(g_k - g) \\
&= S_W^2 + S_B^2
\end{aligned}$$

2.2.11 L'inertie [18]

On appelle inertie totale du nuage de points la moyenne pondérée des carrés des distances des observations à leur centre de gravité :

$$\begin{aligned}
I_{\mathbf{g}} &= \sum_{i=1}^{n} p_i(\mathbf{e}_i - g)^{'} M(\mathbf{e}_i - g) \\
&= \sum_{i=1}^{n} p_i \parallel \mathbf{e}_i - g \parallel^2
\end{aligned}$$

où $M \in \mathbb{M}_p(\mathbb{R})$ est une matrice symétrique positive de taille p.
En *théorie*, le choix de la matrice M dépend de l'utilisateur qui, seul peut préciser la métrique adéquate. En *pratique*, les métriques les plus utilisées, notamment en Analyse en Composantes Principales (ACP), sont en nombre réduit : à part la métrique $M = I$ qui revient à utiliser le produit scalaire usuel, la plus utilisée (qui est souvent l'option par défaut des logiciels) est la métrique diagonale des inverses des variances :

$$M = \boldsymbol{D}_{\frac{1}{S^2}} = \begin{pmatrix} \frac{1}{S_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{S_2^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{S_p^2} \end{pmatrix}$$

Elle a pour effet de s'affranchir des unités de mesures, ce qui est très utile quand, comme c'est souvent le cas en analyse des données, les variables sont exprimées dans différentes unités de mesures.

Définition 6. *L'inertie en un point a quelconque est définie par :*

$$I_a = \sum_{i=1}^{n} p_i (\boldsymbol{e}_i - a)' M (\boldsymbol{e}_i - a) \tag{2.1}$$

On a la relation de HUYGENS *:*

$$\begin{aligned} I_a &= I_g + (g-a)' M (g-a) \\ &= I_g + \| g - a \|^2 \end{aligned}$$

où I_g est l'inertie par rapport au Centre de gravité $\boldsymbol{g}$.

2.2.12 Inerties inter-classe et intra-classe [18]

Etant donné une partition en k groupes d'un nuage de n points, on définira les quantités suivantes : $g_1, \ldots, g_k$ centres de gravité des k groupes et $I_1, \ldots, I_k$ leur inerties respectives.
L'inertie totale I des n points autour du centre de gravité global est égale à la somme de deux termes (*Théorème de HUYGENS*) :

$$I = I_B + I_W \tag{2.2}$$

où $I_W = \sum_{i=1}^{k} p_i I_i$ est l'inertie intra-classe et $I_B = \sum_{i=1}^{k} p_i d^2(g_i, g)$ est l'inertie inter-classe (ou inertie du nuage des k centres de gravité).

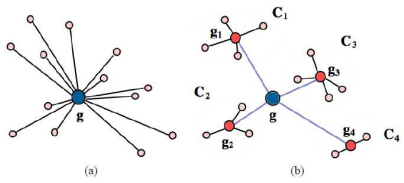

FIG. 2.1 – Inertie totale, (b) Inerties inter-classe (bleu) et intra-classe (rouge).

2.3 Les mesures de ressemblance [9]

La notion de ressemblance a fait l'objet d'importantes recherches dans des domaines extrêmement divers ; elle suscite encore l'intérêt des chercheurs.
On va donc définir ces différentes notions (similarité, dissimilarité et distance) et les liens qui existent entre elles ainsi que leurs différences. Il s'agira ensuite de faire une présentation des mesures de ressemblance définies pour les données suivant leur nature.

2.3.1 Définitions

On appelle similarité ou dissimilarité toute application de l'espace des individus à valeurs numériques qui permet de mesurer le lien entre deux individus d'un même ensemble.
Pour une *similarité*, le lien entre deux individus sera d'autant plus fort que sa valeur est grande.
Pour une *dissimilarité* le lien sera d'autant plus fort que sa valeur est petite.

Notations

On note que : $d_{ij} = d(i,j) = d(\mathbf{e}_i, \mathbf{e}_j)$ et $s_{ij} = s(i,j) = s(\mathbf{e}_i, \mathbf{e}_j)$

2.3.2 L'indice de dissimilarité

Un opérateur de dissemblance $d : E^2 \longrightarrow \mathbb{R}^+$ défini sur l'ensemble des individus $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ est dit *indice de dissimilarité* ou de dissemblance, s'il vérifie les propriétés suivantes :

1. $\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = d(\mathbf{e}_j, \mathbf{e}_i)$ (symétrie)

2. $\forall \mathbf{e}_i \in E; d(\mathbf{e}_i, \mathbf{e}_j) \geq d(\mathbf{e}_i, \mathbf{e}_i) = 0$ (séparabilité)

2.3.3 La distance

Un opérateur de dissemblance $d : E^2 \longrightarrow \mathbb{R}^+$ défini sur l'ensemble d'individus $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ est appelé *distance*, s'il vérifie en plus des deux propriétés 1 et 2 les propriétés d'identité et d'inégalité triangulaire suivantes :

3. $\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = 0 \Rightarrow \mathbf{e}_i = \mathbf{e}_j$ (identité)

4. $\forall \mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k \in E; d(\mathbf{e}_i, \mathbf{e}_k) \leq d(\mathbf{e}_i, \mathbf{e}_j) + d(\mathbf{e}_j, \mathbf{e}_k)$ (inégalité triangulaire)

2.3.4 L'indice de similarité

Un opérateur de ressemblance $s : E^2 \longrightarrow [0, 1]$ défini sur l'ensemble d'individus $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ est dit *indice de similarité*, s'il vérifie en plus de la propriété de symétrie (1), les deux propriétés suivantes :

$2'$**.** $\forall \mathbf{e}_i, \mathbf{e}_j \in E; s(\mathbf{e}_i, \mathbf{e}_j) > 0$ (positivité)

$3'$**.** $\forall \mathbf{e}_i, \mathbf{e}_j \in E \quad \mathbf{e}_i \neq \mathbf{e}_j; s(\mathbf{e}_i, \mathbf{e}_i) = s(\mathbf{e}_j, \mathbf{e}_j) > s(\mathbf{e}_i, \mathbf{e}_j)$ (maximisation)

Il convient de noter ici que le passage de l'indice de similarité s à la notion duale d'indice de dissimilarité qu'on notera d, est trivial.

Etant donné $S = s_{max}$, la similarité d'un individu avec lui-même ($S = 1$ dans le cas d'une similarité normalisée), il suffit de poser :

$$\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = S - s(\mathbf{e}_i, \mathbf{e}_j)$$

2.3.5 Une mesure de ressemblance entre individus

[9]Le processus de classification vise à structurer les données contenues dans $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ en fonction de leurs ressemblances, sous la forme

d'un ensemble de classes à la fois homogènes et contrastées.

L'ensemble d'individus E est décrit généralement par un ensemble de p variables $Y = \{Y_1, Y_2, \ldots, Y_p\}$ définies chacune par :

$$Y_h : E \rightarrow \Delta_h$$

$$\mathbf{e}_i \in E \rightarrow Y_h(\mathbf{e}_i)$$

où Δ_h est le domaine d'arrivée de la variable Y_h.
En conséquence, les données de classification sont décrites dans un tableau où chaque case du tableau contient la description d'un individu sur une des p variables. Ce tableau est en général un tableau homogène qui peut être de type quantitatif ou qualitatif.

Tableau de données numériques (continues ou discrètes)

La distance la plus utilisée pour les données de type quantitatif est la distance de *Minkowski d'ordre r* définie dans $\mathbb{R}^p$ par :

$$\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = (\sum_{h=1}^{p} |Y_h(\mathbf{e}_i) - Y_h(\mathbf{e}_j)|^r)^{\frac{1}{r}}$$

où $r \geq 1$, tel que si :

1. $r = 1$, d est la distance de *Cityblock* ou *Manhattan.*

$$d(\mathbf{e}_i, \mathbf{e}_j) = \sum_{h=1}^{p} |Y_h(\mathbf{e}_i) - Y_h(\mathbf{e}_j)|$$

2. $r = 2$, d est la distance *euclidienne* classique.

$$d(\mathbf{e}_i, \mathbf{e}_j) = \sqrt{\sum_{h=1}^{p} |Y_h(\mathbf{e}_i) - Y_h(\mathbf{e}_j)|^2}$$

3. $r \rightarrow +\infty$, d est la distance de *Tchebychev* définie comme suit :

$$d(\mathbf{e}_i, \mathbf{e}_j) = max_{1 \leq h \leq p} |Y_h(\mathbf{e}_i) - Y_h(\mathbf{e}_j)|$$

Il y a aussi d'autres types de distance, telle que la distance de *Mahalanobis*, qui est très utilisée (en ACP notamment).

Définition 7. *La distance de Mahalanobis prend en considération la corrélation entre les données ; de plus elle n'est pas dépendante de l'échelle de mesure des données. Elle est définie par :*

$$\forall \boldsymbol{e}_i,\ \ \boldsymbol{e}_j \in E; d(\boldsymbol{e}_i, \boldsymbol{e}_j) = (\boldsymbol{e}_i - \boldsymbol{e}_j)\Sigma^{-1}(\boldsymbol{e}_i - \boldsymbol{e}_j)^t$$

où Σ est la matrice de variance-covariance de X.

Remarques

- Souvent, on utilise la distance *euclidienne* mais la distance de *Manhattan* est aussi parfois utilisée, notamment pour atténuer l'effet de larges différences dues aux points atypiques (aberrants ou outliers) puisque leurs coordonnées ne sont pas élevées au carré.
- Il est à noter que dans la plupart des cas, la distance de *Manhattan* donne des résultats semblables à ceux de la distance *euclidienne*.

Tableau de données binaires

Les n individus à classer sont décrits par p variables binaires codées 0 ou 1. La ressemblance entre deux individus $\mathbf{e}_i$ et $\mathbf{e}_j$ se calcule à partir des informations du tableau de contingence 2×2 ci-dessous. Un tel tableau permet de compter le nombre de concordances $(a + d)$ et le nombre de discordances $(b + c)$ entre les individus $\mathbf{e}_i$ et $\mathbf{e}_j$.

		$\mathbf{e}_j$	
		1	0
$\mathbf{e}_i$	1	a	b
	0	c	d

Table 2.2 Tableau de contingence 2×2

où
a = nombre de propriétés que $\mathbf{e}_i$ et $\mathbf{e}_j$ possèdent simultanément ;
b =nombre de propriétés que $\mathbf{e}_j$ ne possède pas mais que $\mathbf{e}_i$ possède ;

c =nombre de propriétés que $\mathbf{e}_i$ ne possède pas mais que $\mathbf{e}_j$ possède ;
d =nombre de propriétés que $\mathbf{e}_i$ et $\mathbf{e}_j$ ne possèdent pas.

Il convient de noter que le rôle des modalités d'une variable binaire est très important dans le calcul d'une mesure de ressemblance entre les individus. En effet, une variable binaire peut être symétrique ou asymétrique.
Parmi les nombreux indices de similarité qu'on peut appeler aussi coefficients d'association entre deux individus $\mathbf{e}_i$ et $\mathbf{e}_j$, les plus connus sont :

. **L'indice de *Jaccard* (1908)** :

L'indice de *Jaccard* est un coefficient d'association connu pour étudier la similarité entre objets pour des données binaires de présence-absence. Cet indice s'écrit de la façon suivante :

$$J = \frac{a}{a+c+d}$$

La distance de *Jaccard* mesure la dissimilarité entre deux individus ou deux partitions. Elle consiste simplement à soustraire l'indice de *Jaccard* de 1.

$$d_J = 1 - \frac{a}{a+c+d}$$

Remarque

Cet indice varie entre 0 à 1 et ne tient compte que des associations positives (présences simultanées).

. **L'indice Simple Match de *Sokal* et *Michener* (1958) :**

Ce coefficient, aussi connu sous le nom *Coefficient de simple concordance*, consiste à calculer la proportion de valeurs communes entre les individus $\mathbf{e}_i$ et $\mathbf{e}_j$. Sa formule est donnée comme suit :

$$\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = \frac{a+d}{a+b+c+d}$$

. **L'indice de *Sokal* et *Sneath* (1963) :**

Il ignore également l'absence conjointe mais, contrairement à l'indice de *Jaccard*, il compte doublement les discordances :

$$\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = \frac{a}{a+2(b+c)}$$

. L'indice de *Russel* et *Rao* :

L'indice de *Russel* et *Rao*, décrit ci-dessous, est la proportion de *traits positifs communs*

$$\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = \frac{a}{a+b+c+d}$$

Tableau de données nominales

La distance la plus utilisée pour les données de type qualitatif est la distance de *Hamming.* Etant donné deux individus $\mathbf{e}_i, \mathbf{e}_j \in E$ décrits chacun par p variables nominales, la distance de *Hamming* entre $\mathbf{e}_i$ et $\mathbf{e}_j$ est donnée par le nombre de caractéristiques de $\mathbf{e}_i$ qui diffèrent de celle de $\mathbf{e}_j$, elle est définie comme suit :

$$\forall \mathbf{e}_i, \mathbf{e}_j \in E; d(\mathbf{e}_i, \mathbf{e}_j) = \sum_{h=1}^{p} Y_h(\mathbf{e}_i) \oplus Y_h(\mathbf{e}_j)$$

où

$$Y_h(\mathbf{e}_i) \oplus Y_h(\mathbf{e}_j) = \begin{cases} 1, \textit{ si } Y_h(e_i) \neq Y_h(e_j) \\ 0, \textit{ sinon} \end{cases}$$

Tableau de données ordinales

Il s'agit d'un tableau de données où les variables qui décrivent les individus sont qualitatives ordinales. Les valeurs $Y_h(\mathbf{e}_i)$ sont remplacées par leurs rangs $R_h(\mathbf{e}_i), R_h(\mathbf{e}_i) = 1, 2, \ldots, p_h$, où p_h est le nombre de valeurs distinctes de la variable Y_h. Ces valeurs sont par la suite transformées en utilisant la formule ci-dessous qui fournit une variation $Z_h(\mathbf{e}_i)$ entre 0 et 1 de $Y_h(\mathbf{e}_i)$:

$$\forall \mathbf{e}_i \in E; Z_h(\mathbf{e}_i) = \frac{R_h(\mathbf{e}_i) - 1}{p_h - 1}$$

La distance entre deux individus $\mathbf{e}_i, \mathbf{e}_j$ est ainsi calculée à partir des variations Z_h considérées comme des données numériques.

Mesure de ressemblance entre variables aléatoires

Dans certains cas, on désire analyser les variables à la place des individus. Ceci requiert la définition d'un opérateur capable d'évaluer la proximité

entre ces variables sous analyse.

L'incertitude sur une variable aléatoire Y_i (respectivement un couple de variables aléatoires (Y_i, Y_j)) peut être mesurée par l'entropie, notée $H(Y_i)$ (respectivement $H(Y_i, Y_j)$).

La quantité notée $I(Y_i : Y_j)$, appelée information mutuelle, mesure l'information transmise entre Y_i et Y_j.

$$I(Y_i : Y_j) = H(Y_i) + H(Y_j) - H(Y_i, Y_j)$$

L'indépendance entre Y_i et Y_j entraîne $I(Y_i : Y_j) = 0$ et la dépendance entre les deux variables entraîne $I(Y_i : Y_j) = H(Y_i, Y_j)$
Dussauchoy (1982) a proposé un indice de similarité normé et un indice de dissimilarité entre variables aléatoires qui s'écrivent comme suit :

$$s(Y_i, Y_j) = \frac{I(Y_i : Y_j)}{H(Y_i, Y_j)} \tag{2.3}$$

$$d(Y_i, Y_j) = \frac{H(Y_i, Y_j) - I(Y_i : Y_j)}{H(Y_i, Y_j)} \tag{2.4}$$

Il a également généralisé cette notion de dissimilarité pour mesurer la dissemblance entre deux vecteurs aléatoires.

2.4 Algorithmes de classification

Il existe plusieurs familles d'algorithmes de classification dans certains conduisent directement à des partitions tels que la méthode de classification autour de centres mobiles qui est un cas particulier de la techniques de nuées dynamiques ou des k-means (k-moyennes), ou les algorithmes ascendants (ou agglomératifs) qui procèdent à la construction des classes par agglomérations successives des objets deux à deux fournissant une hiérarchie de partitions des objets.

Domaines d'application

La classification a un rôle important à jouer dans toutes les sciences et techniques qui font appel à la statistique multidimensionnelle. Citons

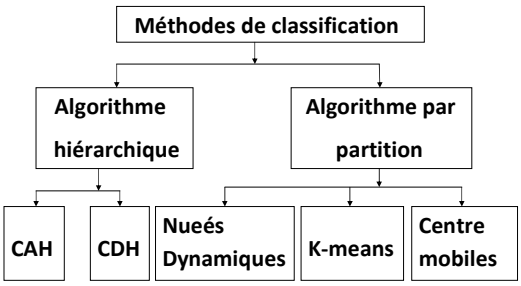

FIG. 2.2 – Méthodes de classifications

tout d'abord les sciences biologiques (botanique, zoologie, écologie,...). Ces sciences utilisent également le terme de "taxinomie" pour désigner l'art de la classification. De même, les sciences de la terre et des eaux (géologie, pédologie, géographie, étude des pollutions) font grand usage de méthodes de classifications.

La classification est fort utile également dans les sciences humaines (psychologie, sociologie, linguistique, archéologie, histoire, ...) et dans les techniques dérivées comme les enquêtes d'opinion, le marketing, ... Ces dernières utilisent parfois les mots de "typologie" et "segmentation" pour désigner la classification, ou l'une de ses innombrables variantes. Citons encore la médecine, l'économie, l'agronomie qui font beaucoup usage de ces méthodes.

2.4.1 Classification par partition (CPP)

Ce sont des méthodes qui produisent directement une partition en un nombre fixé de classes. Parmi ces méthodes, nous retrouvons :

La méthodes des nuées dynamiques [21]

L'algorithme connu sous le nom de nuées dynamiques étudié formellement par *Diday* (1971) permet de traiter des ensembles d'effectifs assez importants en optimisant localement un critère de type inertie.

Etant donnée une partition à k groupes de n points, de centres de gravités $c_1, \ldots, c_k$ et d'inerties $I_1, \ldots, I_k$. L'inertie totale I des n points autour du centre de gravité global g est :

$$I = I_B + I_W \tag{2.5}$$

avec pour

$$I_B = \sum n_{p_h} d^2(g_h, g) \tag{2.6}$$

l'inertie inter-classe des k centres de gravités et n_{p_h} le poids de la h^{ime} classe, et pour

$$I_W = \sum n_{p_h} I(p_h) \tag{2.7}$$

l'inertie intra-classe.
Un critère usuel consiste à chercher la partition telle que I_W soit minimale pour avoir des classes homogènes pour k fixé. Ce qui revient à chercher le maximum de I_B.

Remarques

1. Dans la technique des nuées dynamiques, les classes peuvent ne pas être caractérisées par un centre de gravité, mais par un noyau ayant un meilleur pouvoir descriptif que des centres ponctuels.

2. Cette méthode a l'avantage de traiter rapidement de grands ensembles d'individus. Elle fournit une solution dépendant de la configuration initiale et nécessite le choix du nombre de classes. En général, le nombre de classes est fixé par l'utilisateur et l'initialisation est faite par un tirage au hasard. Pour comparer l'individu avec les noyaux, cette méthode utilise des distances, ce qui a l'inconvénient de faire appel à des métriques.

Méthode des k-means

Cette méthode est encore appelée algorithme des centres mobiles. Ce type d'algorithme, où la classe est représentée par son centre de gravité.
Cette méthode commence par un tirage pseudo-aléatoire de centres ponctuels. Chaque réaffectation d'individus entraîne une modification de la position du centre correspondant, on peut en une seule itération trouver une partition de bonne qualité mais dépendant de l'ordre des individus.

La méthode d'agrégation autour des centres mobiles [5]

Le but de cette méthode de classification est de répartir en classes un ensemble d'objets dont on connaît les distances deux à deux. Les classes formées doivent être les plus homogènes possibles.
Cette méthode permet de traiter rapidement des ensembles d'effectifs assez élevés en optimisant localemaent un critère de type inertie.

Algorithme d'agrégation autour des centres mobiles

On supposera que les n individus à classer sont des points de $\mathbb{R}^p$ muni d'une distance euclidienne d.
Due à *Forgy* (1965), elle consiste à partir de k points $c_1, \ldots, c_k$ pris au hasard dans E (ensemble des individus), ces k points définissent une partition de E en k classes $E_1, \ldots, E_k$.
En effet, la partition de $\mathbb{R}^p$ associée à ces k centres est déterminée par les hyperplans médiateurs des centres. E_i est la classe constituée par l'ensemble des points de E plus proche de c_i que de tout autre centre.

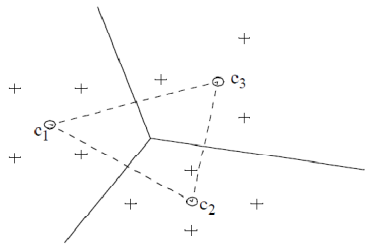

FIG. 2.3 – Principe de la méthode des centres mobiles

On remplace alors successivement les k points pris au hasard par les centres de gravités de ces classes jusqu'à ce que l'algorithme ait convergé vers la partition optimale, c'est-à-dire vers celle qui minimise l'inertie intra-classe.

Remarque

La partition optimale ainsi déterminée dépend du choix des points ayant servi de point initial de l'algorithme. En pratique, on recommence cette opération un grand nombre de fois et l'on retient la partition optimale qui présente l'inertie intra-classe minimale, ce qui revient à maximiser l'inertie inter-classe pour avoir des classes homogènes pour k fixé.

Démonstration. On va montrer que la variance intra-classe ne peut que décroitre ou bien reste stationnaire entre la partition de l'étape m et celle de l'étape $m+1$.
Des règles d'affectation permettant de faire en sorte que cette décroissance soit stricte et donc de conclure à la convergence de l'algorithme puisque l'ensemble de départ E est fini.
Supposons que les n individus de l'ensemble à classer E soient munis de poids relatifs p_i tel que $\sum_{i=1}^{n} p_i = 1$, et soit $d^2(i, g_k^m)$ le carré de la distance entre l'individu i et le centre de la classe k à l'étape m.
Nous nous intéressons à la quantité (critère) :

$$v(m) = \sum_{k=1}^{q} (\sum_{i \in C_k^m} p_i d^2(i, g_k^m)) \tag{2.8}$$

où C_k^m est une classe formée des individus les plus proches de g_k^m que de tous les autres centres, ces centres étant des centres de gravité des classes C_k^{m-1}.
La variance intra-classe à l'étape m est la suivante :

$$V(m) = \sum_{k=1}^{q} (\sum_{i \in C_k^m} p_i d^2(i, g_k^{m+1})) \tag{2.9}$$

où g_k^{m+1} est le centre de gravité de la classe C_k^m.
A l'étape $m+1$, $v(m+1)$ s'écrit comme suit :

$$v(m+1) = \sum_{k=1}^{q} (\sum_{i \in C_k^{m+1}} p_i d^2(i, g_k^{m+1})) \tag{2.10}$$

On doit montrer que :

$$v(m) \geq V(m) \geq v(m+1) \tag{2.11}$$

ce qui établira la décroissance simultanée du critère et de la variance intra-classe.
Soit $p_k = \sum_{i \in C_k^m} p_i$. D'après le théorème de *Huygens* :

$$v(m) = V(m) + \sum_{k=1}^{q} p_k d^2(g_k^{m+1}, g_k^m) \tag{2.12}$$

ce qui établit la première partie de l'inégalité.
La seconde partie découle du fait qu'entre les deux formules de $v(m+1)$ et $V(m)$ seules changent les affectations des points aux centres. Puisque C_k^{m+1} est l'ensemble des points les plus proches de g_k^{m+1} que de tous les autres centres, les distances n'ont pu que décroitre au cours de cette réaffectation. □

Exemple

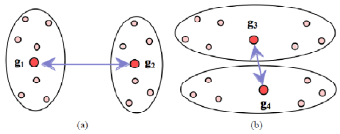

FIG. 2.4 – (a) Inerties intra-classe faible et une inter-classe élevée, (b) le contraire.

Méthode des k-medoids [9]

Dans des méthodes de k-medoids une classe est représentée par un de ses individus (médoide). C'est une méthode itérative combinant la réaffecctation des individus dans des classes avec une intervention des médoides et des autres individus. C'est une méthode simple parce qu'elle couvre n'importe quel type de variables. Quand des médoides sont choisis, les classes sont définis comme sous-ensembles des individus près des médides les plus proche par rapport à une mesure des distance choisie.

Il est alors judicieux de choisir comme centre de groupe un individu présent dans le groupe et non un individu calculé. La médoide d'un cluster est l'individu possédant la dissimilarité moyenne la plus faible avec les autres individus du cluster.

Remarque

Les méthodes non hiérarchiques permettent de traiter rapidement de grands ensembles d'individus, mais elles supposent que le nombre de classes est fixé au départ. Si le nombre de classes n'est pas connu ou si ce nombre ne correspond pas à la configuration véritable de l'ensemble d'individus, il faut presque toujours tester diverses valeurs de k, ce qui augmente le temps de calcul. C'est pourquoi, lorsque le nombre d'individus n'est pas trop élevé, on préfère utiliser les méthodes hiérarchiques.

2.4.2 Classification hiérarchique(CH)

Définition 8. *Un ensemble $\mathcal{H}$ de parties non vides de E est une hiérarchie sur E si*

- $E \in \mathcal{H}$
- $\forall e_i \in E, \{e_i\} \in \mathcal{H}$
- $\forall H, H^{'} \in \mathcal{H}, H \cap H^{'} = \emptyset$ *ou* $H \subset H^{'}$ *ou* $H^{'} \subset H$
- $\forall C_i \in \mathcal{H}; C_i = \cup_{j=1}^{n_i} C_{ij}$ *avec* $n_i = card(C_i)$

Une hiérarchie peut être vue comme un ensemble de partitions emboîtées. Graphiquement, une hiérarchie est souvent représentée par un arbre hiérarchique dit aussi *dendrogramme*.

Il existe deux grandes familles de méthodes : une descendante, dite divisive, et une ascendante, dite agglomérative. La première, moins utilisée, consiste à partir d'une seule classe regroupant tous les objets, à partager celle-ci en deux. Cette opération est répétée à chaque itération jusqu'à ce que toutes les classes soient réduites à des singletons.Mais nous nous intéressons à la seconde qui est la plus couramment utilisée.

L'interêt de ces arbres est qu'ils peuvent donner une idée du nombre de classes existant effectivement dans l'échantillion. Pourtant, déterminer le nombre exact de classes est très difficile. La visualisation du dendogramme représente un moyen mais ceci est utile seulement pour un nombre réduit de

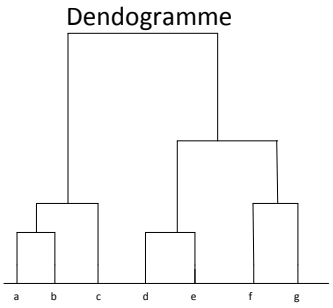

FIG. 2.5 – Dendogramme ou arbre hiérarchique

données.

Cet arbre étant obtenu de manière ascendante dans la plupart des méthodes, on regroupe d'abord les deux individus les plus proches qui forment un noeud, il obtiendra $(n-1)$ objets et on réitère le processus jusqu'à regroupement complet. Un des problèmes consiste à définir une mesure de dissimilarité entre parties.
On peut résumer les étapes de cette méthode comme suit :

Algorithme de classification ascendante hiérarchique (CAH)

Pour classifier un échantillon d'effectif n dont les individus sont numérotés $1, 2, \ldots n$, on considère cet échantillon comme la réunion de n classes à un seul élément et on regroupe progressivement les classes deux à deux selon le schéma d'un algorithme de classification ascendante hiérarchique (CAH) qui est le suivant :

1. Les classes initiales sont les individus eux-mêmes.
2. On calcule les distances entre les classes.
3. Les deux classes les plus proches sont fusionnées et remplacées par une seule.

4. Le processus reprend en 2 jusqu'à n'avoir qu'une seule classe, qui contient toutes les observations.

Voici plusieurs définitions possibles de distances entre des classes formées de plusieurs individus. Soient $C, C^{'}$ deux classes.

. *Le saut minimum/single linkage.* La distance du saut minimum entre les classes $C, C^{'}$, notée $d(C, C^{'})$ est par définition :

$$d(C, C^{'}) = min_{i \in C, j \in C^{'}}(d(i, j))$$

C'est la plus petite distance entre éléments des deux classes.

. *Le saut maximum/complete linkage.* La distance du saut maximum entre les classes $C, C^{'}$, notée $d(C, C^{'})$ est par définition :

$$d(C, C^{'}) = max_{i \in C, j \in C^{'}}(d(i, j))$$

C'est la plus grande distance entre éléments des deux classes.

. *Le saut moyen/average linkage.* La distance du saut moyen entre les classes $C, C^{'}$, notée $d(C, C^{'})$ est par définition :

$$d(C, C^{'}) = \frac{1}{|C||C'|} \sum_{i \in C} \sum_{j \in C'} d(i, j)$$

C'est la moyenne des distances entre tous les individus des deux classes.

. *Le saut barycentrique/Centroid-linkage* définit, quant à lui, la distance entre deux clusters comme la distance entre leur centre de gravité. Une telle méthode est mieux résistante aux points aberrants. Toutefois, elle est limitée aux données quantitatives numériques pour lesquelles le calcul du centre de gravité est possible.

La méthode de Ward pour distances euclidiennes

Soit E considéré comme un nuage de l'espace $\mathbb{R}^p$, on agrège les individus qui font le moins varier l'inertie intra-classe, c'est-à-dire qu'on cherche à trouver à chaque pas un minimum local de l'inertie intra-classe ou bien un maximum de l'inertie inter-classe.
L'indice de dissimilarité entre deux classes est égal à la perte d'inertie inter-classe résultant de leur regroupement.
On va quantifier maintenant cette *perte d'inertie.*

Soit g_A, g_B les centres de gravité respectifs des classes A et B, g_{AB} le centre de gravité de leur réunion, donné par :

$$g_{AB} = \frac{p_A g_A + p_B g_B}{p_A + p_B}$$

où p_A et p_B sont les poids des deux classes.

L'inertie inter-classe étant la moyenne des carrés des distances des centres de classe au centre de gravité total g, la variation d'inertie est égale à :

$$p_A d^2(g_A, g) + p_B d^2(g_B, g) - (p_a + p_B) d^2(g_{AB}, g)$$

2.5 Conclusion

Les deux modes de construction d'une partition (agglomératif et divisif) aboutissent à une classification hiérarchique indicée qui n'est pas forcément la même. Une fois qu'elle est obtenue, il peut être intéressant d'analyser et d'interpréter cette classification, afin de choisir la partition idéale et de fournir ensuite une représentation pour chaque groupe (par exemple par son centre). Il existe aussi d'autres méthodes de classification hiérarchique qui ont été développées, pour éviter certains problèmes (telle que la complexité algorithmique), et notamment pour fournir des partitions en classes de formes et tailles arbitraires.
Les résultats de la classification sont validés par l'évaluation d'indices de validité définis sur l'ensemble de données ; ceux-ci nous offrent une information sur la cohérence de la partition faite par une certaine méthode.

Chapitre 3

Comparaison de classifications

Ce chapitre est consacré à la présentation et à la définition des différents indices qui nous paraissent importants dans la validation du nombre de classes et aussi à la comparaison de classifications.

3.1 Indices de validation du nombre de classes [11]

L'objectif principal des techniques de classification est de trouver une bonne partition où les objets d'une classe doivent être semblables, c'est-à-dire que le nombre de classes soit optimal, les objets de différentes classes devraient être différents.
Une bonne classification devrait ainsi satisfaire différents critères de validité. On distingue trois catégories d'indices ou bien critères : internes, externes et relatifs.

. Les critères internes

1. Chaque classe d'une partition doit être *homogène* c'est-à-dire les objets qui appartiennet à la même classe doivent être semblables.
2. Les classes doivent être *isolées* entre elles.
3. La classification doit s'adapter aux données, autrement dit la classification doit pouvoir *expliquer la variation dans les données.*

. Les critères externes

Les classes doivent être *valides* (il doit y avoir une corrélation avec les variables externes -non utilisées pour grouper).

. ***Les critères relatifs***

La classification cherchée doit être la meilleure relativement à un certain critère.

Il existe un autre critère qui est :

. ***Stabilité***

Les classes doivent être stables, les petites modifications dans les données et dans les méthodes ne doivent pas changer d'une manière significatve les résultats.

De nombreuses procédures, appelées aussi indices de validité de clustering, ont été proposées dans le but de déterminer la meilleure partition d'un jeu de données numériques.
La présentation de ces différents indices est maintenant donnée. Pour une partition P en k classes $\{C_1, C_2, \ldots, C_k\}$ de l'ensemble d'individus $E = \{e_1, e_2, \ldots, e_n\}$, la dispersion intra-classe $s_a(C_i)$ et la séparation inter-classe $d_a(C_i, C_j)$ sont données par les formules suivantes :

$$\forall C_i \in P; s_a(C_i) = \frac{1}{|C_i|(|C_i|-1)} \sum_{u=1}^{|C_i|} \sum_{v=1}^{|C_i|} d(e_u, e_v) \tag{3.1}$$

$$\forall C_i, C_j \in P; d_a(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{u=1}^{|C_i|} \sum_{v=1}^{|C_j|} d(e_u, e_v) \tag{3.2}$$

où d représente la mesure de dissimilarité définie sur l'ensemble d'individus E et $|C_i|$ le cardinal de la classe C_i.

3.1.1 Indice de Davies-Bouldin

L'indice de Davies-Bouldin est basé sur la minimisation du rapport entre les dispersions intra-classe et la séparation inter-classe. Il est calculé comme suit :

$$DB(P) = \frac{1}{k} \sum_{i=1}^{k} max_{1 \leq j \leq k, j \neq i} \{ \frac{s_a(C_i) + s_a(C_j)}{d_a(C_i, C_j)} \} \tag{3.3}$$

On constate ainsi que ce rapport sera d'autant plus faible que les classes sont compactes et éloignées les unes des autres. Par conséquent, la partition de meilleure qualité sera celle qui minimisera l'indice de *Davies-Bouldin*.

3.1.2 Indice de Dunn

L'indice de Dunn [1974] est basé sur l'identification de clusters compacts et bien séparés.

Soit d_{min} la distance minimale entre deux objets de deux différentes classes et d_{max} la distance maximale entre deux objets de la même classe.

Alors, l'indice de *Dunn* (1974), est défini par :

$$D = \frac{d_{min}}{d_{max}} \tag{3.4}$$

L'objectif principal de cet indice est de maximiser la dissimilarité inter-classe et de minimiser la dissimilarité intra-classe. Notre but est donc de maximiser l'indice de Dunn. Une bonne classification est indiquée par des valeurs élevées de cet indice.

3.1.3 L'indice de Dunn généralisé

L'indice de *Dunn généralisé* est reconnu comme l'un des indices les plus appropriés pour l'évaluation de la qualité d'une partition donnée, car il fournit un bon compromis entre la maximisation de la dissimilarité inter-classe et la minimisation de la dissimilarité intra-classe de la partition.

$$D_g(P) = \frac{min_i\{min_{i\neq j} d_a(C_i, C_j)\}}{max_h s_a(C_h)} \tag{3.5}$$

La partition P produisant la plus grande valeur de $D_g(P)$ correspondra à la meilleure classification.

3.1.4 L'indice C_0

Soit l le nombre des paires d'individus dans une classe.
L'indice C_0 mesure la compacité des classes [Hubert et Schula 1976] et il est défini par :

$$C_0 = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{3.6}$$

où S représente la somme des distances entre toutes les paires d'individus dans une classe.

S_{min} représente la somme des l plus petites distances si toutes les paires d'individus sont considérées et S_{max} représente la somme des l plus grandes distances issues de toutes les paires d'individus. Le dénominateur sert à la normalisation, $C_0 \in [0, 1]$. Cet indice est particulièrement utilisé si les classes sont de taille similaire et il est petit si les classes est plus compacte.

3.1.5 Indice de compacité-séparabilité

L'indice de compacité-séparabilité CS est décrit dans [Forgy 1992] pour une partition floue et il utilise le même principe que l'indice DB, il tient compte à la fois de la compacité c_0 et de la séparabilité s_e des classes. Pour une classification dure il est défini par :

$$CS = \frac{c_0}{s_e} \tag{3.7}$$

où

$$c_0 = \frac{1}{k}\sum_{i=1}^{k}\sigma_i \;\; et \;\; s_e = min_{i\neq j}(d(g_i, g_j)) \tag{3.8}$$

3.1.6 L'indice de Silhouette

L'indice de silhouette est défini pour tout individu e_i de l'ensemble E par la formule suivante :

$$\forall e_i \in E; s(e_i) = \frac{b(e_i) - a(e_i)}{max(a(e_i), b(e_i))} \tag{3.9}$$

où

. $a(e_i)$ est la dissimilarité moyenne entre l'individu e_i et tous les autres individus de la classe à laquelle il appartient $C(e_i)$.

$$\forall e_i \in E; a(e_i) = \frac{1}{|C(e_i)| - 1}\sum_{e_j \in C(e_i), e_i \neq e_j} d(e_i, e_j) \tag{3.10}$$

. $b(e_i)$ est le minimum des dissimilarités moyennes entre l'individu e_i et tous les autres individus des classes de la partition P autres que $C(e_i)$.

$$\forall e_i \in E; b(e_i) = min_{C\in P, C\neq C(e_i)} d(e_i, C) \tag{3.11}$$

où $d(e_i, C) = \frac{1}{|C|}\sum_{e_j \in C} d(e_i, e_j)$

On notera que l'indice de silhouette est borné : $-1 \leq s(e_i) \leq 1$. De plus, lorsque $s(e_i)$ est proche de 1, e_i est dit bien classé dans $C(e_i)$. Quant $s(e_i)$ est proche de 0, alors e_i se situe entre deux classes. Finalement, si $s(e_i)$ est proche de -1, e_i est dit mal classe dans $C(e_i)$ et doit être rattaché à un autre cluster le plus proche.

Chaque classe est aussi représentée par une silhouette qui montre quels objets sont correctement classés à l'intérieur de cette classe et lesquels n'ont simplement qu'une position intermédiaire. Pour une classe C_i donnée, son indice de silhouette est défini par la moyenne des indices de silhouette des individus qui lui appartiennent :

$$\forall C_i \in P; s(C_i) = \frac{\sum_{e_j \in C_i} s(e_j)}{|C_i|} \tag{3.12}$$

L'indice de silhouette global de la partition P est donné par la moyenne globale des largeurs de silhouettes dans les différentes classes C_i qui composent la partition :

$$s(P) = \frac{\sum_{C_i \in P} s(C_i)}{k} \tag{3.13}$$

La meilleure partition retenue est alors celle qui permet d'obtenir un indice de silhouette global maximal.

3.2 Notations et définitions [19]

Dans cette partie, nous introduisons les notations de base, ainsi que les définitions élémentaires en classification qui seront utilisées.
P_1 et P_2 sont deux partitions des mêmes individus (ou deux variables qualitatives). X désigne le tableau de contingence associé, X_1, X_2 les tableaux disjonctifs associés à P_1 et P_2 ; on pose : $N = X_1^{'} X_2$.
Chaque partition P_k est représentée par un tableau relationnel C^k dans l'espace des individus, de dimension $n \times n$, dont le terme général $c^k_{ii'}$ est défini par :

$$c^k_{ii'} = \begin{cases} 1, \ si \ i \, et \, i^{'} \, sont \, deux \, individus \, dans \, la \, même \, classe \, de \, la \, partition \, P_k \\ 0, \ sinon \end{cases}$$

L'écriture matricielle du tableau de comparaison par paires est $C = X_1 X_1^{'}$ et son tableau complémentaire, notée $\overline{C}$ à pour terme général :

$$\overline{c_{ii'}^k} = \begin{cases} 0, \ si \ i \, et \, i^{'} \, sont \, deux \, individus \, dans \, la \, même \, classe \, de \, la \, partition \, P_k \\ 1, \ sinon \end{cases}$$

Nous posons :

$$\begin{aligned} n &= nombre \ d'individus \\ p &= nombre \ de \ classes \ de \ la \ partition \ P_1 \\ q &= nombre \ de \ classes \ de \ la \ partition \ P_2 \end{aligned}$$

Lorsque l'on croise deux partitions, on va s'intéresser aux paires $\overline{c_{ii'}^k}$ d'individus qui restent ou ne restent pas dans les mêmes classes.
On a C_n^2 paires d'individus représentées par les quatre types dans le tableau suivant :

$P_1 \setminus P_2$	**Même classe**	**Classes différentes**
Même classe	a Accord positif	d Désaccord
Classes différentes	c Désaccord	b Accord Négatif

Table 3.1 Tableau croisant les deux partitions P_1 et P_2

On notera également $A = a + b$ (nombre total d'accords) et $D = c + d$ (nombre total de désaccords). On peut aussi, au lieu de considérer les C_n^2 paires $(i, i^{'})$ considérer les n^2 paires $(i, i^{'})$ (où $(i, i^{'})$ est distingué de $(i^{'}, i)$ et où l'on comptabilise les n paires (i, i). Si $a^{'}, b^{'}, c^{'}, d^{'}$ désignent les équivalents de : a, b, c, d, on a alors :

$$a^{'} = 2a + n; b^{'} = 2b; c^{'} = 2c; d^{'} = 2d$$

Le tableau de contingence croisant P_1 et P_2 est de dimension $p \times q$. Il est caractérisé par son terme général : n_{uv} = l'effectif de la case (u, v).

Maintenant, on va citer quelques indices d'importance avérée.

3.3 Indices de comparaison de deux partitions (mêmes individus)

On peut examiner visuellement deux partitions en comparant leur dendogrammes pour différentes solutions. Cette technique est fastidieuse pour

un nombre d'individus assez grand. Pour cela, des indices et des tests statistiques existent, qui permettent de comparer les différentes classifications ; on en cite quelques-uns.

3.3.1 Indice brut de Rand [20]

Dans le but de comparer deux partitions à p et q classes respectivement, l'indice d'accord le plus utilisé est l'indice de *Rand*.

L'indice brut R de *Rand* est le pourcentage global de paires en accord :

$$R = \frac{A}{C_n^2} \tag{3.14}$$

L'indice de Rand peut être récrit sous la forme suivante dans le cas du croisement des deux partitions :

$$R = \frac{a+d}{a+b+c+d} \tag{3.15}$$

Il prend ses valeurs entre 0 et 1, il est égal à 1 lorsque les deux partitions sont identiques.

3.3.2 Indice de Rand dans sa version asymétrique [20]

On utilise l'indice de Rand asymétrique dans le cas où on a deux partitions d'un même ensemble d'individus mais avec des nombres de classes inégaux.
Soient P_1 et P_2 deux partitions de n individus dont le nombre de classes de P_1 est supérieur au nombre de classes de P_2.

P_1 est plus fine que P_2 lorsque deux éléments sont classés ensemble dans P_1 et ils le sont également dans P_2 : $\forall u = 1, \cdots, p, \exists v = 1, \cdots, q$ tel que $P_1^u \subseteq P_2^v, P^u$ (respectivement P^v) désignant la $u^{ème}$ (respectivement $v^{ème}$) classe de P_1 (respectivement P_2).

On cherche à mesurer l'inclusion de la partition P_1 dans la partition P_2.

En considérant toutes les paires d'individus, y compris celles identiques, l'écriture simple de l'indice R_A est la suivante :

$$R_A = \frac{a^{'} + b^{'} + c^{'}}{a^{'} + b^{'} + c^{'} + d^{'}} \tag{3.16}$$

Remarque

Dans le cas où les deux partitions ont même nombre de classes, l'indice de Rand asymétrique n'est pas égal à l'indice brut de Rand.

3.3.3 Indice de Jaccard [21]

L'indice de Jaccard est un coefficient d'association connu pour étudier la similarité entre individus pour des données binaires de présence-absence.
Le tableau binaire suivant représente un exemple de présence-absence de deux individus i et $i^{'}$ quelconques à p critères différents :

	y_1	y_2	$\cdots$	y_p
i	1	0		1
$i^{'}$	0	0		1

Table 3. 2 Tableau croisant les deux individus selon les p critères

On peut former alors le tableau suivant :

$i^{'} \setminus i$	1	0
1	$11_{(i,i')}$	$10_{(i,i')}$
0	$01_{(i,i')}$	$00_{(i,i')}$

Table 3.3 Tableau croisant les deux individus selon les p critères

où $11_{(i,i')} = a =$ nombre de critères ou propriétés que i et $i^{'}$ possèdent simultanément.

$01_{(i,i')} = c =$nombre de propriétés que i ne possède pas mais que $i^{'}$ possède.

$10_{(i,i')} = b =$ nombre de propriétés que $i^{'}$ ne possède pas mais que i possède.

$00_{(i,i')} = d =$ nombre de propriétés que i et $i^{'}$ ne possède pas .

L'indice J de Jaccard s'écrit de la façon suivante :

$$J = \frac{a}{a + c + d} \tag{3.17}$$

Cet indice varie entre 0 à 1 et ne tient compte que des associations positives (présences simultanées).

3.3.4 Le coefficient Kappa de Cohen [20]

Ce coefficient est destiné à mesurer l'accord entre deux variables qualitatives pour des données appariées.
Dans l'étude de l'accord entre deux variables indépendantes ayant m modalités, le coefficient kappa s'écrit :

$$k = \frac{P_o - P_e}{1 - P_e} \tag{3.18}$$

où la concordance observée P_o est la proportion d'individus classés dans les cases diagonales de concordance du tableau de contingence, soit la somme des effectifs diagonaux divisés par la taille n de l'échantillon .

$$P_0 = \frac{1}{n} \sum_{i=1}^{m} n_{ii}$$

La concordance aléatoire P_e est égale à la somme des produits des effectifs marginaux divisée par le carrée de la taille de l'échantillon.

$$P_e = \frac{1}{n^2} \sum_{i=1}^{m} n_{i.} n_{.i}$$

Le coefficient kappa est un nombre réel, compris entre -1 et 1. L'accord sera d'autant plus élevé que sa valeur est proche de 1 et l'accord maximal est atteint lorsque $P_o = 1$ et $P_e = 0.5$.
Lorsqu'il y a indépendance entre les variables, le coefficient kappa est nul et dans le cas d'un désaccord total, kappa prend la valeur -1 avec $P_o = 0$ et $P_e = 0.5$. Ceci n'est vrai que dans le cas où les marges sont égales ($n_{i.} = n_{.i}$) puisqu'il suffit de prendre les effectifs diagonaux (ceux qui expriment l'accord dans le tableau de contingence) égaux aux marges et les effectifs non diagonaux égaux à 0.

Pour des marges données, on peut determiner la valeur maximale k_m de Kappa :

$$k_m = \frac{P_m - P_e}{1 - P_e} \tag{3.19}$$

où P_m est la proportion d'accords maximaux donnée la formule suivante :

$$P_m = \frac{1}{n} \sum_{i=1}^{m} inf(n_{i.}, n_{.i})$$

3.4 Stabilité des classes

Dans cette partie, on s'intéresse à la stabilité d'une classe d'une partition, dans le but de répondre aux questions suivantes :

. Les classes des deux partitions sont-elles homogènes ?

. Les proportions des classes ont-elles changé ?

3.4.1 Test d'homogénéité du Khi-deux [21]

Ce test est appliqué pour savoir si les classes de deux partitions sont homogènes ou non, on teste alors l'hypothèse suivante :

$$\begin{cases} H_0 : \textit{les partitions proviennent de la même population} \\ \textit{vs} \\ H_1 : \textit{les partitions sont significativement différentes.} \end{cases}$$

Les k classes des deux partitions d'un même nombre n d'individus sont réparties de la façon suivante :

	$\mathbf{C_1}$	$\mathbf{C_2}$	$\cdots$	$\mathbf{C_k}$	**Total**
$\mathbf{P_1}$	n_{11}	n_{12}		n_{1k}	**n**
$\mathbf{P_2}$	n_{21}	n_{22}		n_{2k}	**n**
Total	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$		$\mathbf{n_{.k}}$	**n**

Table 3.4 Tableau de répartition des classes selon P_1 et P_2

Si n_{1h} est le nombre d'individus de la partition P_1 qui se trouvent dans la classe h, on a :
$n = \sum_{h=1}^{k} n_{1h} = \sum_{h=1}^{k} n_{2h}$ = taille de la population pour les deux partitions.

$n_{.h} = n_{1h} + n_{2h}$ = nombre total d'individus se trouvant dans la classe h pour les deux partitions.

Pour l'hypothèse H_0, les probabilités $p_1, p_2, \ldots, p_k$ d'être dans les classes $C_1, C_2, \ldots, C_k$ sont rarement connues. Il s'agit alors de comparer les effectifs observés n_{1h} ou n_{2h} aux effectifs espérés np_h qui ne doivent pas en différer de beaucoup (si l'hypothèse d'homogénéité est plausible).
La statistique-test est donnée par :

$$D^2 = \sum_{h=1}^{k} \sum_{i=1}^{2} \frac{(n_{ih} - np_h)^2}{np_h}$$

qui suit, sous H_0, un $\mathcal{X}^2_{k-1}$, k étant le nombre de classes dans chacune des deux partitions.
On estime les $\hat{p}_h = \frac{n_{.h}}{2n} = \frac{n_{1h}+n_{2h}}{2n}$, ce qui fait $(k-1)$ estimations indépendantes (et donc un nombre de degrés de liberté de $(k-1)$).
D'où

$$D^2 = \sum_{h=1}^{k} \sum_{i=1}^{2} \frac{(n_{ih} - \frac{nn_{.h}}{2n})^2}{\frac{nn_{.h}}{2n}} = \sum_{h=1}^{k} \sum_{i=1}^{2} \frac{(n_{ih} - \frac{n_{.h}}{2})^2}{\frac{n_{.h}}{2}}$$

D^2 est un $\mathcal{X}^2$ à $(k-1)$ degrés de liberté, pour deux partitions homogènes.

Remarque

L'inconvénient de ce test est qu'il est utilisé sur les distributions marginales lignes, il n'exploite pas le fait que les individus n'aient pas changé de classes.

3.4.2 Test de Mac Nemar [21]

Pour étudier la stabilité des classes, on teste si les proportions des classes de deux partitions ont changé. Ce test non-paramétrique mesure, pour des données dichotomiques, l'égalité des proportions des classes.

On utilise le test généralisé de Mac Nemar qui étudie la variation des pourcentages sur un ensemble d'individus pour des classes des deux partitions. Il est utilisé pour tester si la probabilité qu'un individu classé dans

(i, j) est la même que la probabilité qu'un individu classé dans (j, i).

Pour deux partitions P_1 et P_2 formées de k classes chacune, le tableau de contingence est représenté comme suit :

$\mathbf{P_1 \backslash P_2}$	$\mathbf{C_1}$	$\mathbf{C_2}$	$\cdots$	$\mathbf{C_v}$	$\mathbf{C_k}$	**Total**
$\mathbf{C_1}$	n_{11}	n_{12}		n_{1v}	n_{1k}	$\mathbf{n_{1.}}$
$\mathbf{C_2}$	n_{21}			n_{2v}	n_{2k}	$\mathbf{n_{2.}}$
$\vdots$						
$\mathbf{C_u}$	n_{u1}	n_{u2}		n_{uv}	n_{uk}	$\mathbf{n_{u.}}$
$\mathbf{C_k}$	n_{k1}	n_{k2}		n_{kv}	n_{kk}	$\mathbf{n_{k.}}$
Total	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	$\cdots$	$\mathbf{n_{.u}}$	$\mathbf{n_{.k}}$	$\mathbf{n}$

Table 3.5 Tableau de contingence de P_1 et P_2

Avec n_{uv} = nombre d'individus qui sont dans la classe u de P_1 et dans la classe v de P_2.
On teste donc :

$$\begin{cases} H_0 : n_{u.} = n_{.u} \ \forall u \in k \\ vs \\ H_1 : \exists u^{'} \ telque \ n_{u^{'}.} \neq n_{.u^{'}}. \end{cases}$$

La statistique du test de Mac Nemar dans le cas de notre tableau s'écrit alors :

$$T = \sum_{u \neq v} \frac{(n_{uv} - n_{vu})^2}{n_{uv} + n_{vu}} \tag{3.20}$$

On rejette H_0 au niveau de confiance α, si T dépasse le quantile $(1 - \alpha)$ de la loi de khi-2 de degré de liberté égal à $k(k-1)/2$ où k est le nombre de classes de chacune des deux partitions. Sinon, on accepte à $\alpha\%$ que les proportions des classes n'ont pas changé dans les deux partitions.

Remarque

Ce test a un avantage sur le test précédent car il tient en compte du fait que ce sont les mêmes individus.

3.5 Conclusion

On a vu dans ce chapitre l'importance de quelques indices pour appliquer une procédure de validité du nombre de classes ainsi que pour comparer une partition par rapport à une autre dans le but de choisir la meilleure des deux. On a vu deux tests statistiques (test d'homogénéité du $\mathcal{X}^2$ et test de Mac Nemar) pour vérifier la stabilité des classes.

Il existe aussi un indice du Khi-deux qui n'a pas été utilisé car il sert à vérifier l'indépendance et non la concordance.

Chapitre 4

Application [3]

Notre démarche consiste à traiter un tableau de données en appliquant les méthodes de classification par partitions(CPP) et hiérarchique.
Plusieurs logiciels sont utilisés lors du traitement des données par exemple :

. **Splus** est utilisé pour appliquer la méthode des k-means et exécuter les algorithmes qui ont été proposés,

. **SPAD** est utilisé pour la méthode de la classification ascendante hiérarchique, pour choisir le nombre de classes des partitions et leur description (les classes),

. et le logiciel **R** dont on se servira pour notre application.

4.1 Le Logiciel R

R est un système d'analyse statistique et graphique crée par *Ross IHAKA* et *Robert GENTLEMAN*. C'est à la fois un logiciel et un langage qualifié de dialecte du langage **S** créé par AT et T Bell Laboratories.
R est un logiciel libre, clone d'un autre logiciel très célèbre dans la communauté statisticienne : $\mathbf{S}^{+}$. Il peut être téléchargé gratuitement sur ***www.r-project.org***. Sur ce site, se trouvent également des documentations très complètes, notamment le manuel du chercheur-biologiste *Emmanuel PARADIS*.

Comme tous les logiciels libres, le développement et l'amélioration de **R** peuvent être effectués par tout un chacun. Les développeurs insistent sur le fait que **R** permet un calcul vectoriel lui ouvrant des applications dans

d'autres domaines que la statistique.
Il permet notamment de :

. manipuler des données (stocker, manipuler des tables de données de grande dimension, y extraire des données, les résumer ...) ;

. effectuer du calcul matriciel et autres opérations mathématiques complexes ;

. réaliser des analyses statistiques -des plus simples aux plus complexes ;

. réaliser des graphiques -du plus simple au plus élaboré ;

. programmer de façon simple et efficace ;

. se lier avec d'autres logiciels ou langages (possibilités de lien avec **C**, **C**$^{++}$ et **Fortran** pour compiler de lourdes tâches).

Packages utilisés

Toutes les analyses statistiques de notre application sont faites, comme précédemment indiqué, sous le logiciel **R**, grâce à différents **packages** qui ont été téléchargés :

1. ade4
2. FactoMiner
3. clusterSim

Notre but dans cette application est de calculer le nombre de classes optimal, application des méthodes de classification ascendante hiérarchique et k-means, et de calculer la valeur de l'indice de Davies-Bouldin et la valeur de l4indice de Rand pour comparer la meilleure méthode parmi les deux.

4.2 Traitement des données

Le tableau de données **olympic** contient les performances de trente-trois (33) athlètes aux jeux olympiques de en 1988 dans les épreuves du décathlon.
Les 10 colonnes numériques correspondent aux épreuves suivantes :
100 : course du 100 mètres,
long : saut en longueur,
poid : lancer du poids,
haut : saut en hauteur,
400 : course du 400 mètres,
110 : course du 110 mètres-haies,

disq : lancer du disque,
perc : saut à la perche,
jave : lancer du javelot,
et
1500 : course du 1500 mètres.

Le tableau de données (*olympic$tab*) est fourni avec la librairie **ade4**, en utilisant les commandes suivantes :

```
> library(ade4)

Attachement du package : 'ade4'


        The following object(s) are masked from package:base :

         within

> data(olympic)
> olympic$tab
> plot(olympic$tab)
> olympic$tab
```

On a fait une représentation graphique des données pour mieux voir la corrélation entre les dix (10) variables. Par exemple, on voit que les deux variables "poid" et "disq" sont corrélées, par contre, la variable "jave" n'est pas corrélée avec la variable "1500".

Création de la matrice des distances

Puisque les données ne sont pas exprimées dans la même unité, on procède à leur centrage et leur réduction. La distance qu'on va utiliser est la distance euclidienne grâce à la commande :

```
> don<-scale(olympic$tab, center = TRUE, scale = TRUE)
```

On calcule le tableau des distances par :

```
> dc<-dist(don,method ="euclidean",diag=FALSE,upper=FALSE)
> dc
```

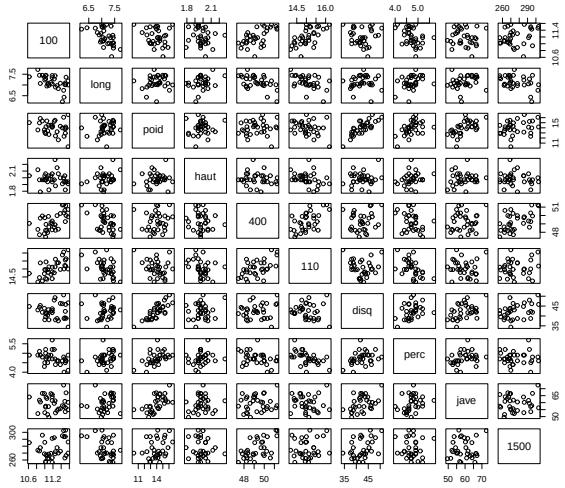

FIG. 4.1 – La représentation graphique de tableau des données

Choix du critère d'agrégation

Le choix du critère d'agrégation est important. Selon le critère choisi, on aboutira à une partition claire ou illisible. Par exemple, dans le cas où le critère d'agrégation est "single", on aboutit à une classification illisible. Pour cela, on va la comparer avec le critère d'agrégation "Ward", en utilisant les commandes suivantes :

```
> par(mfrow=c(1,2))
> hier<-hclust(dc,"single")
> hier

Call: hclust(d = dc, method = "single")

Cluster method   : single
Distance         : euclidean
> hier<-hclust(dc,"ward")
> plot(hier,hang=-1)
```

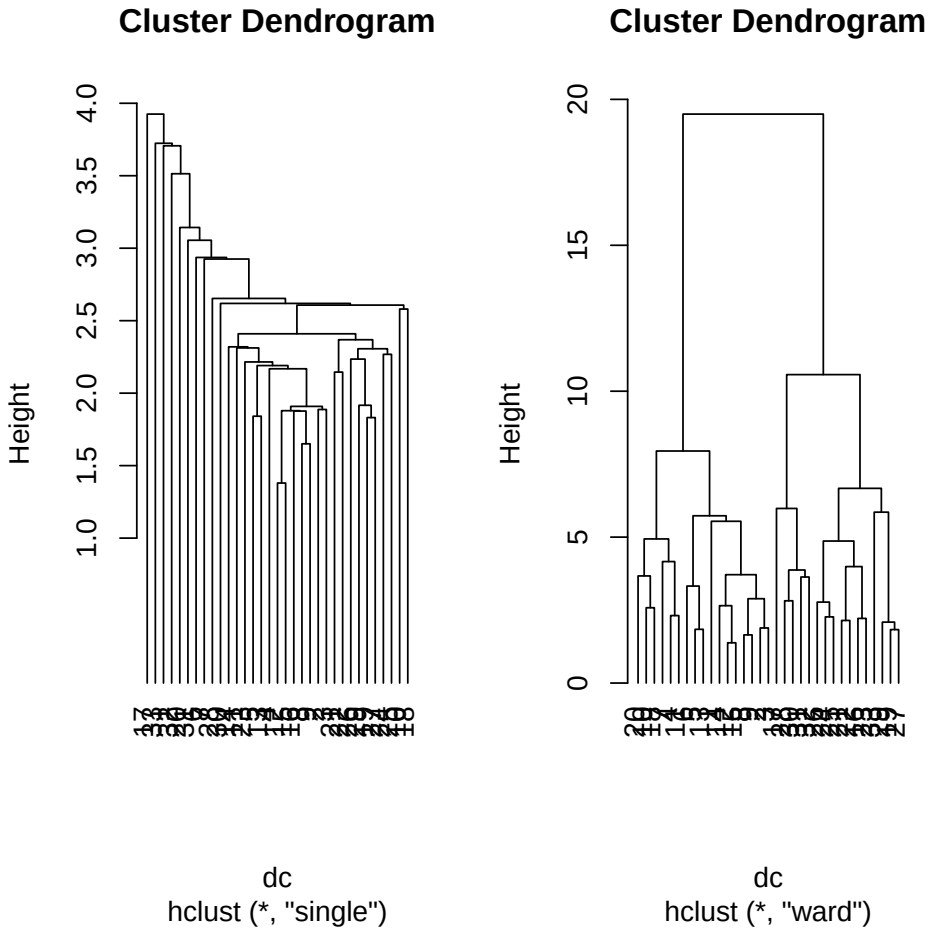

FIG. 4.2 – Dendogramme héirarchique avec le critère d'agrégation "Ward" et "Single"

Remarque

Si on choisit le critère "Ward", on aboutit à une classification plus claire que l'autre.

On effectue une classification hiérarchique par le critère de Ward pour connaître le nombre de classes indispensable pour appliquer la méthode des k-means.

Le nombre de classes dans notre application est peut être compris entre 2 et 32, pour cela on va utiliser les commandes suivantes qui peuvent nous donner un nombre de classes qui s'avère meilleur pour notre situation :

```
> library(FactoMineR)
Le chargement a nécessité le package : ellipse

Le chargement a nécessité le package : lattice

Le chargement a nécessité le package : cluster

Le chargement a nécessité le package : scatterplot3d

Attachement du package : 'FactoMineR'

        The following object(s) are masked from package:ade4 :

        reconst

> res.pca <- PCA(olympic$tab, scale.unit=TRUE)
> HCPC(res.pca,conso=0)
$data.clust
     100 long  poid haut   400   110  disq perc  jave   1500 clust
33 11.57 7.19 10.27 1.91 50.71 16.20 34.36  4.1 54.94 269.98     1
32 11.47 6.43 12.33 1.94 50.30 15.00 38.72  4.0 57.26 293.72     1
31 11.43 6.22 13.98 1.91 51.25 15.88 46.18  4.6 57.84 294.99     1
30 11.50 7.09 12.94 1.82 49.27 15.56 42.32  4.5 53.50 293.85     1
28 11.51 7.01 14.17 1.94 51.16 15.18 45.84  4.6 56.28 303.17     1
29 11.26 6.90 12.41 1.88 48.24 15.61 38.02  4.4 52.68 272.06     1
17 11.46 6.75 16.07 2.00 51.28 16.06 50.66  4.8 72.60 302.42     2
```

```
26 11.33 6.83 11.63 2.06 48.37 15.39 37.52  4.6 55.42 270.07     1
24 11.30 6.97 13.23 2.15 49.98 15.38 38.72  4.6 54.34 277.84     1
22 11.49 7.02 13.80 2.03 50.60 15.22 39.08  4.7 60.92 262.93     1
21 11.52 7.36 13.93 1.94 49.99 15.64 38.82  4.6 67.04 266.42     1
27 11.10 6.98 12.69 1.82 48.63 15.13 38.04  4.7 49.52 261.90     1
23 11.38 7.08 14.31 2.00 50.24 14.97 46.34  4.4 55.68 272.68     1
25 11.00 7.23 13.15 2.03 49.73 14.96 38.06  4.5 52.82 285.57     1
18 11.57 7.00 16.60 1.94 49.84 15.00 46.66  4.9 60.20 286.04     2
19 11.07 7.04 13.41 1.94 47.97 14.96 40.38  4.5 51.50 262.41     1
20 10.89 7.07 15.84 1.79 49.68 15.38 45.32  4.9 60.48 277.84     2
16 11.09 7.08 14.51 2.03 49.89 14.78 43.20  4.9 57.18 268.54     1
15 11.03 7.45 14.20 1.97 48.94 15.44 41.66  4.7 64.00 267.48     2
12 11.18 7.34 14.48 1.94 49.02 15.11 42.76  4.7 65.84 256.74     2
9  11.15 7.12 14.52 2.03 49.15 14.66 42.36  4.9 66.46 269.62     2
13 11.02 7.29 12.92 2.06 48.23 14.94 39.54  5.0 56.80 257.85     2
14 10.99 7.37 13.61 1.97 47.83 14.70 43.88  4.3 66.54 268.97     2
10 11.23 7.28 15.25 1.97 48.60 14.76 48.02  5.2 59.48 292.24     2
8  11.05 6.95 15.34 2.00 48.21 14.36 41.32  4.8 63.00 265.86     2
11 10.94 7.45 15.34 1.97 49.94 14.25 41.86  4.8 66.64 295.89     2
7  11.18 7.05 14.12 2.06 49.34 14.39 41.68  5.7 61.60 291.20     2
1  11.25 7.43 15.48 2.27 48.90 15.13 49.28  4.7 61.32 268.95     2
3  11.18 7.44 14.20 1.97 48.29 14.81 43.66  5.2 64.16 263.20     2
5  11.02 7.43 12.92 1.97 47.44 14.40 41.20  5.2 57.46 256.64     2
4  10.62 7.38 15.02 2.03 49.06 14.72 44.80  4.9 64.04 285.11     2
6  10.83 7.72 13.58 2.12 48.34 14.18 43.06  4.9 52.18 274.07     2
2  10.87 7.45 14.97 1.97 47.71 14.46 44.36  5.1 61.76 273.02     2


Call: agnes(x = X, diss = FALSE, metric = metric, stand = FALSE,
method = method)

Cluster method   : ward
Distance         : euclidean
Number of
objects: 33
```

On voit que le nombre de classes est égale à 2. Le R nous donne le nombre optimal de clusters.

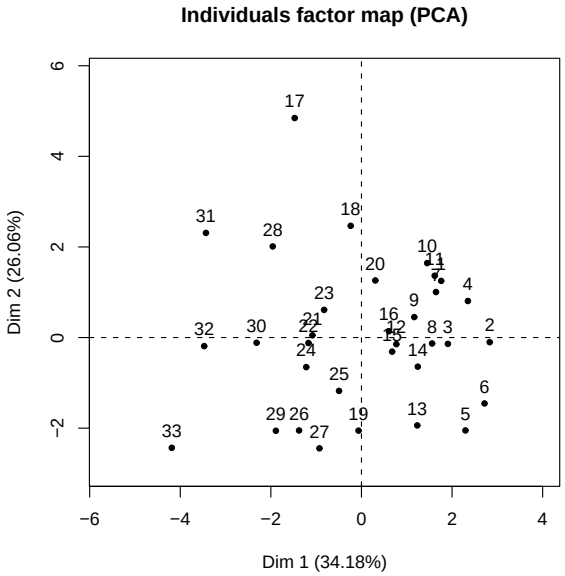

FIG. 4.3 – Représentation des individus

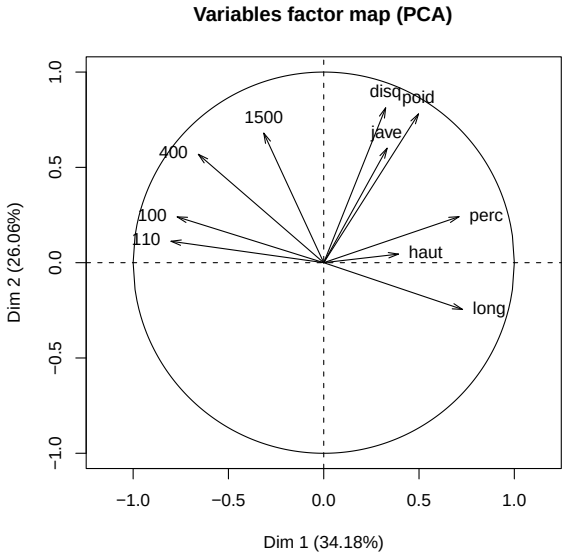

FIG. 4.4 – Représentation des variables

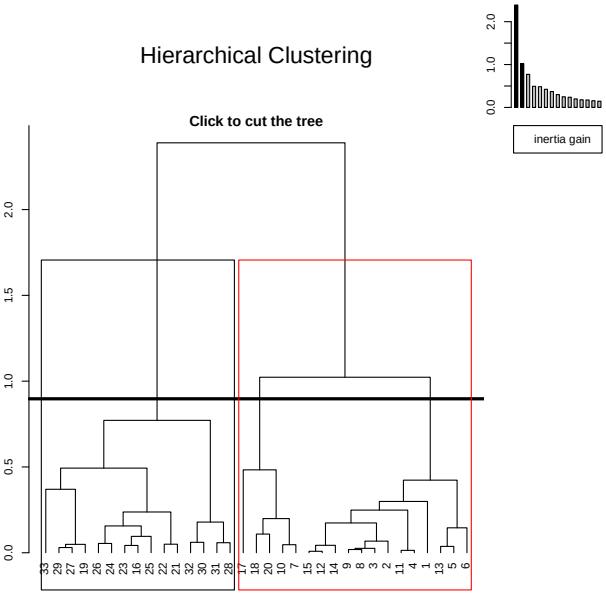

FIG. 4.5 – Arbre hiérarchique

On voit que dans la représentation graphique de l'arbre hiérarchique, il y a aussi le graphique des gains d'inertie intra-classe obtenus à chaque itération de l'algorithme CAH. On remarque que le gain d'inertie intra-classe diminue en fonction du nombre d'itérations.

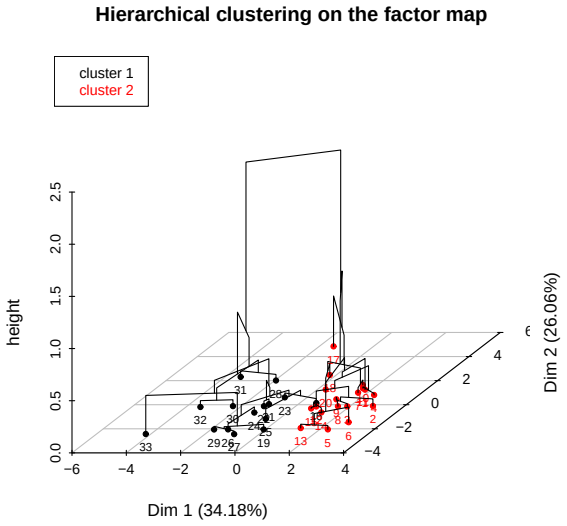

FIG. 4.6 – Représentation 3D de l'arbre hiérarchique sur le premier plan factoriel

L'arbre hiérarchique est représenté en trois dimensions sur le plan principal de l'ACP (Fig. 4.6). La fonction ayant proposé un découpage en deux classes, les individus sont colorés en fonction de leur classe d'appartenance.

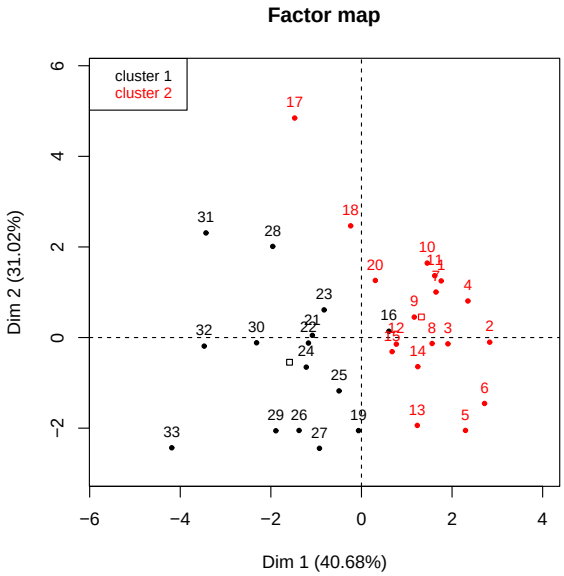

FIG. 4.7 – Représentation des clusters

Méthode de k-means

Pour la méthode de k-means, on va utiliser la commande suivante :

```
> km=kmeans(olympic$tab, centers=2, iter.max=100, algorithm="Forgy")
> km
K-means clustering with 2 clusters of sizes 22, 11

Cluster means:

    100  long     poid     haut      400      110     disq

1 11.16227 7.205 13.73864 1.991818 48.92545 15.03864 41.49455

2 11.26455 6.990 14.45182 1.964545 49.97909 15.06909 44.07273

  perc      jave      1500
1 4.722727  59.05545   267.5032

2 4.772727  60.20545  293.1091

 Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
 1  1  1  2  1  1  2  1  1  2  2  1  1  1  1  1  2  2  1  1  1

22 23 24 25 26 27 28 29 30 31 32 33
 1  1  1  2  1  1  2  1  2  2  2  1

Within cluster sum of squares by cluster:
[1] 1706.1441  902.9649

Available components:

[1] "cluster"  "centers"  "withinss" "size"
```

Elle nous donne la taille de chaque classe, comment les individus sont classés et aussi les variables.

La valeur de l'indice de Davies et Bouldin

Dans le cas de la méthode de k-means, la valeur de l'indice de Davies-Bouldin est donnée par les commandes suivantes :

```
> library(clusterSim)
> data(olympic)
> cl1 <- pam(olympic$tab, 2)
>d<-dist(olympic$tab)
> print(index.DB(olympic$tab, cl1$clustering,d, centrotypes="centroids"))
$DB [1] 0.6926743

$d
          1         2
1   0.00000    25.79364
2   25.79364   0.00000
```

Dans le cas de la méthode de k-medoids, la commande qu'on utilise pour calculer la valeur de l'indice de Davies-Bouldin est la suivante :

```
> print(index.DB(olympic$tab, cl1$clustering,d, centrotypes="medoids"))
$DB [1] 0.8233158

$d
          1         2
1    0.00000    23.16417
2  23.16417    0.00000
```

Remarque

Ces commandes, nous donnent aussi la matrice de distance entre les clusters.
On voit que la valeur de l'indice de Davies-Bouldin dans le cas k-means est plus petite que celle de cas k-medoids, la partition de meilleure qualité sera celle qui minimisera l'indice de Davies-Bouldin, donc la meilleure méthode c'est celle de k-means.

Comparaison de deux partitions ascendantes hiérarchiques

Pour comparer deux partitions ascendantes hiérarchiques sur même individus mais avec différentes choix de critères d'agrégation et aussi nombre de classes. On voit le changement de la valeur de l'indice de Rand.
1ère cas :

```
> library(clusterSim)
>  data(olympic)
>  z<-data.Normalization(olympic$tab,type="n0")
>  d<-dist(z)
>  h<-hclust(d, method="ward")
>  t<-cutree(h,k=2)
> yy<-hclust(d, method="single")
> cmn<-cutree(yy,k=2)
> print(comparing.Partitions(t,cmn,type="rand"))
[1] 0.5643939
```

2ème cas :

```
> h<-hclust(d, method="ward")
> t<-cutree(h,k=2)
> v<-hclust(d, method="ward")
> ml<-cutree(v,k=3)
> print(comparing.Partitions(t,ml,type="rand"))
[1] 0.7727273
```

L'indice de Rand dans le 2^{me} cas vaut 0.772, cette valeur relativement proche de 1 ne suffit pas pour dire que les deux partitions sont proches, en effet cet indice donne la même importance aux couples d'individus qui sont ou non dans la même classe (accord global).

Chapitre 5

Conclusion Générale

Le problème de la clssification des données prend une large part dans les analyses statistiques que les chercheurs de divers domaines s'emploient à traiter pour répondre à des questions concrètes relatives à leurs préoccupations.

Dans tout l'arsenal des méthodes existantes, nous avons synthétisé plusieurs d'entre elles pour résoudre et répondre aux questions du problème général de la classification non-supervisée. Elles diffèrent par les mesures de proximité, ou de dissemblance/ressemblance utilisées, la nature des données sous analyse et leur objectif final. Chaque méthode possède des points forts et des "faiblesses".
Les méthodes ascendantes hiérarchiques sont utilisées pour des échantillons de petite taille car la complexité (des algorithmes sous-jacents impliquant des matrices de distances de grandes dimensions) est très élevée. Si au contraire, des problèmes de temps d'exécution se posent, on préférera alors la méthode des k-means.

Dans notre application, on s'est intéressés à la comparaison, sur un ensemble de données de variables numériques, de diverses méthodes et de certains indices pour décider de la partition optimale.

Malgré le nombre imporatnt de méthodes connues jusque-là, plusieurs problématiques restent encore ouvertes dans le cadre de la classification. Un problème très souvent rencontré concerne la difficulté de fixer les paramètres des méthodes à utiliser par l'utilisateur (le praticien dans quelque domaine qui a des données à traiter). La complexité des méthodes mises en œuvre reste une problématique dans certains cas (de tableaux de données de tailles

assez élevées).

Bibliographie

[1] A.Belhedi, *Statistique et analyse des données*, Université de Tunis, 2010.

[2] M.Boubou, *Contribution aux méthodes de classification non supervisée via des approches prétopologique et d'agrégation d'opinions*, Université Claude Bernard-Lyon1, 2007.Thèse de Doctorat

[3] A.Bouchier, *L'analyse des données multivariées à l'aide du logiciel*, Montpellier, 2010.

[4] A.Boulemnadjel, Partitionnement neuronal et validité des classes Application à la segmentation d'images, Université Mentouri-Constantine, 2009. Thèse de Doctorat

[5] F.Dazy, J.Le Barzic, F.Lavallard et G.Saporta, *L'analyse des données évolutives méthodes et application*, Technip, 1996.

[6] A.Da Silva, Y.Lechevallier, *Analyse exploratoire des indices de détermination du bon nombre de clusters application aux aonnées évolutives*, INRIA Paris, 2009.

[7] S.Déjean, *Analyse statisque de données d'expression*, Université de Toulouse et CNRS, 2008.

[8] Béatrice De Tilière, Analyses statistiques multivariées, 2008.
http ://www.proba.jussieu.fr/ detiliere/Cours/polycop_bio.pdf

[9] H.ELGHAZEL, Classification et prévision des données Hétérogènes : Application aux Trajectoires et Séjours Hospitaliers, Université Claude Bernard Lyon1, 17 décembre 2007.Thèse

[10] Équipe De Mathématique Appliquées, *Applications linéaires et matrices*, UTC, 2009.
http ://tice.utc.fr/moodle/course/view.php ?id=335

[11] S.Gadat, *Positionnement multidimensionnel-classification.*
www.math.univ-toulouse.fr/∼gadat/Ens/· · · /03-m1-classif.pdf

[12] C.Lazar, *Méthodes non supervisées pour l'analyse des données multivariées*, l'Université de Reims Champagne Ardenne, 2008.Thèse de Doctorat

[13] B.Liquet, R.Drouilhet et P.Lafay de Micheaux, *Le logiciel R maîtriser le langage*, Springer-Verlag France, 2011.

[14] M.Paegelow, *Expression (Carto-)Graphique*, 2000
http ://sigfrance.free.fr/ressources/filebrowser/downloads/Cours%20 cartographie/pdf/lc10.pdf.

[15] D.Poidevin, *Les méthodes de discrétisation*, Ellipses, 1999.

[16] F.Rossi, *Classification automatique*, TELECOM ParisTech, 2009.

[17] G.Saporta, *Problèmes posés par la comparaison de classifications dans des enquêtes différentes*, France, 1997.

[18] G.Saporta, Probabilités et Analyse des Données et Statistique, Technip, Paris, 2006.

[19] G.Saporta, G.Youness, *Une méthodologie pour la comparaison de partitions*, NUMDAM, 2004.

[20] G.Saporta, G.Youness, *Concordance entre deux Partitions : quelques propositions et expériences*, Institut des Sciences Appliquées et Economiques CNAM-Université Libanaise et CEDRIC.
http ://cedric.cnam.fr/index.php/publis/article/view ?id=316

[21] G.Youness, *Contributions à une méthodologie de comparaison de partitions*, 2004. Thèse de Doctorat
www.info.univ-angers.fr/∼ gh/wstat/discr.php.

Résumé

Le seul moyen de faire une méthode instructive et naturelle, est de mettre ensemble les choses qui se ressemblent et de séparer celles qui diffèrent les unes des autres.

Georges BUFFON, Histoire naturelle, 1749.

Cette phrase du célèbre naturaliste et écrivain *Georges BUFFON* peut servir de définition générale à un modèle de classifcation.

Les modèles les plus classiquement utilisés en classiffication sont les partitions et les hiérarchies de parties. Dans les deux cas, les objets qui se ressemblent sont regroupés en classes (ou clusters). Pour les partitions, les classes sont deux à deux disjointes ; pour les hiérarchies, elles peuvent être emboîtées. Dans les deux cas, elles ne sont pas empiétantes, dans le sens où l'intersection de deux d'entres elles n'en produira jamais de troisième.

La notion de classification est essentielle en science car elle permet aux scientifiques de mettre de l'ordre dans l'information et les connaissances qu'ils ont sur le monde. Aussi, depuis longtemps des scientifiques et des chercheurs de divers bords ont essayé de classer des espèces (animales ou autres), et -plus généralement- des données de diverses natures.

De nombreuses classifications ont été créées. Face à celles-ci, les scientifiques sont souvent incapables de désigner la meilleure d'entre elles, c'est à dire celle qui a une prévalence sur toutes les autres pour tout ensemble de données ; car chacune présente un intérêt, au moins légèrement supérieur, par rapport à d'autres et en fonction de la tâche considérée.

Le terme "*classification*" est associé à la notion d'abstraction. En effet, une classification permet de synthétiser des informations dans des groupes ou ensembles de données très généraux ; c'est une forme d'abstraction dans le sens où l'on va mettre de côté les descriptions exactes des objets et ne faire ressortir que les traits particuliers que certains d'entre eux ont en commun.

L'importance de la classification dans les sciences se reflète dans la grande variété des domaines où tant leur nature que leur construction ont fait l'objet de recherches ; on citera à ce propos SOKAL [1963], BONNER[1964], FORGY [1965], Mac QUEEN[1967], LANCE et WILLIAMS [1967], DIDAY [1971], BENZECRI [1973], GORDON [1987], CELEUX et al. [1989], ...

Dans le cadre d'un problème de classification, on dispose d'un ensemble de données qui représente une collection d'individus (objets) Les classes sont encore inexistantes. L'objectif est alors d'obtenir des classes d'objets homogènes, en favorisant l'hétérogénéité entre ces différentes classes.

Pour cela, la définition de la "*classification*" amène à se poser les questions suivantes :

- Comment les objets à classer sont-ils définis ?
- Comment définir la notion de ressemblance (dissemblance) entre objets ?
- Comment sont structurées les classes (clusters) ?
- Comment préférer une classification par rapport à une autre ?

Dès le départ il est nécessaire de différencier la classification non supervisée et la classification supervisée ou analyse discriminante. La classification supervisée consiste à construire des règles de décision en se basant sur un ensemble de données pour lesquelles les étiquettes des classes sont connues a priori. Le but de la classification non supervisée est de trouver une organisation des données cohérente et valide, qui puisse mettre en évidence les vraies structures dans un ensemble de données sans aucune connaissance a priori sur les données traitées.
Parmi les différentes méthodes, on peut considérer deux grands types d'approches :

1. **Non-paramétriques :** Les approches dites non-paramétriques (classification hiérarchique, méthode des centres mobiles) ne considèrent qu'une seule hypothèse : plus deux individus sont proches, plus ils ont

de chance de faire partie de la même classe.

2. **Probabilistes :** Les approches dites probabilistes utilisent une hypothèse sur la distribution des individus à classifier. Par exemple, on peut considérer que les individus de chacune des classes suivent une loi normale. Le problème qui se pose alors est de déterminer quels sont les paramètres de la loi et à quelles classes les individus ont le plus de chances d'appartenir.

Dans notre travail, nous nous intéressons *exclusivement* aux méthodes non-paramétriques qui sont des méthodes de classification automatique qu'on appelle aussi classification non supervisée.

Ce mémoire est organisé en quatre (4) chapitres de la manière suivante :

Dans le premier chapitre, nous définissons et donnons quelques types de déscrétisations ; on en donnera certaines caractéristiques, et l'on rappellera quelques méthodes de calcul du nombre de classes (méthodes de la racine carrée, de SCOTT, ... etc).

Dans le deuxième chapitre, on présentera les méthodes classiques de classification automatique utilisées en analyse de données dans le cadre non-paramétrique ; nous reviendrons -au préalable- sur quelques rappels concernant des notions telles que : tableaux de données ou de contingence, tableau disjonctif, matrices de distance, de variance-covariance et de corrélation, inerties et mesures de ressemblance.

Le troisième chapitre étudie en détail quelques indices de validation des classes (clusters), ainsi que d'autres (indices) qui serviront à comparer des partition sur un même ensemble d'individus.

Dans le chapitre quatre, après une brève présentation du logiciel R, on fait une (re)présentation des données ; la représentation graphique en dendogramme nous aidera à sélectionner le critère d'agrégation utilisé. Après une première analyse (une ACP), on appliquera la méthode ascendante hiérarchique qui nous donne un nombre optimal de clusters auxquels on applique la méthode des k-means ; après calcul de quelques indices, on détermine après comparaison la partition optimale.

Enfin, on va terminer avec une conclusion générale et essayer d'indiquer quelques perspectives et voies de recherche dans le vaste domaine que reste la classification en analyse statistique des données.