N° d'ordre : 07/2008-E/MT

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ DES SCIENCES ET DE LA TECHNOLOGIE
« HOUARI BOUMEDIENNE »
FACULTÉ DES MATHÉMATIQUES

Thèse
Présentée pour l'obtention du diplôme de Doctorat D'Etat
EN : MATHEMATIQUES
Spécialité: Probabilités & Statistiques
Par : REBBOUH Amar

Sujet

Analyse de l'information apportée par les éléments constitutifs d'une structure de tableau de données

Soutenue publiquement le 13 décembre 2008 devant le jury composé de:

Mr	BERRACHEDI	Abdelhafid	Professeur	USTHB	**Président**
Mr	DJEDOUR	Mohamed	Professeur	USTHB	**Directeur de thèse**
Mr	AIDER	Meziane	Professeur	USTHB	**Examinateur**
Mr	ANES	Ouali	M/C	INPS, Alger	**Examinateur**
Mr	IBAZIZEN	Mohamed	Professeur	UMTO	**Examinateur**
Mr	HOUACINE	Amrane	Professeur	USTHB	**Examinateur**

Table des matières

Chapitre
Introduction générale

Cette thèse s'inscrit dans le cadre de la classification automatique et plus précisement dans la classification des éléments constitutifs d'une structure de tableaux multiples. L'analyse des éléments constitutifs d'une juxtaposition de tableaux multiples reste un large domaine d'investigation. L'objectif principal de la classification automatique et de l'analyse des données est de définir des indices de similarité entre individus ou entre variables qui permettent d'associer à un couple d'objets ou de variables une valeur numérique. Ces indices de mesures sont choisis en fonction des données et de la structure de classification désirée.

Nous apportons dans cette thèse quelques nouveaux éclairages qui permettent d'appréhender l'information apportée par les éléments constitutifs d'une structure de juxtaposition de tableaux multiples. Une structure de juxtaposition de tableaux multiples peut être obtenue dès que l'on est conduit à observer plusieurs fois chaque objet pour les paramètres qui le décrivent. Aussi, nous présentons plusieurs approches qui permettent :

- *De comparer globalement des différentes variables*
- *D'étudier les différents individus.*
- *De faire des typologies d'individus suivant des critères adaptés à la structure de données se présentant sous la forme d'une juxtaposition de tableaux à observations multiples.*

La première partie de la thèse est une rétrospective des techniques classiques du type factorielles pouvant s'adapter à ce type de données. Dans la seconde partie, nous définissons trois indices qui permettent sous certaines conditions, de mesurer la similarité entre les éléments constitutifs de la structure. Dans la troisième partie, nous introduisons la notion d'objets aléatoires et de systèmes physiques associés pour modéliser les observations de chaque individu pour les variables qui les décrivent. La quatrième partie est consacrée à la classification d'objets décrits par des données entachées d'erreurs de mesure. Enfin, la dernière partie de la thése regroupe certaines perspectives, ouvertures et développements futurs.

Si les algorithmes de classification automatique existants se distinguent pour la plupart par le choix de l'indice mesurant la ressemblance ou la dissemblance entre objets. Lermam[62] note que le problème de ce choix est délicat et est loin d'être résolu. Cette remarque reste hélas vraie à l'heure actuelle. Aussi, dans cette thèse, nous avons défini 5 indices qui sont

- L'indice basé sur le produit scalaire de *Hilbert-Schmidt* et sur la définition d'objets représentatifs de chaque élément constitutif de la structure. Cet indice a été adapté à une structure de tableaux de mesure ($TofM$) et un algorithme de type k-means McQueen[65]a été validé pour classifier les clients domestiques d'une entreprise de distribution d'électricité.

- L'indice basé sur des techniques factorielles, adapté dans le cas d'une structure de juxtaposition de tableaux de mesure.
- L'indice basé sur les techniques utilisées en théorie de l'information (système physique, entropie,...) et sur les travaux de Kullback-Leibler[55].
- La distance classique euclidienne entre sommets des pavés qui décrivent les individus dans le cas de la classification de données avec erreurs de mesure et si les variables ayant participé à la description des individus sont non corrélées.
- Si les variables sont corrélées donc quelconques, nous utilisons la distance pondérée de *Mahannalobis.*

Dans la troisieme partie, nous introduisons la notion d'objets aléatoires. Les objets aléatoires sont une modélisation des observations de chaque individu pour les variables qui les décrivent. Cette approche utilise la notion d'entropie comme mesure de l'incertitude des états du système associé à chaque élément de la structure. Le concept de système physique est celui proposé par Shannon-Weinner[90].Cette notion fait encore l'objet de discussions entre scientifiques de différentes spécialités (chimie, théorie du signal, statistique, ...)

Depuis 1824, date à laquelle **Carnot** dans "r*éflexions sur la force motrice du feu*" a ébauché les principes fondamentaux d'une nouvelle science : **La thermodynamique**. Cette science est basée sur l'existence d'une grandeur physique fondamentale baptisée : **entropie**.

Le principe que chacun des "systèmes physiques" possède une entropie qui lui est propre et qui représente son degré de "*désordre*" a été admis depuis. Cependant, c'est sur la définition d'un système physique que les discussions entre scientifiques demeurent.

Le fait de considérer qu'une observation est une réalisation d'une variable aléatoire conduit à raisonner sur l'ensemble des états ou valeurs que peut prendre cette variable. Aussi, si la variable est discrète et le nombre d'états fini, chaque état sera mesuré par sa fréquence. Il s'agit d'étudier la probabilité que la variable ou le système associé d'être dans une des configurations possibles ou états. On se ramène donc à la notion de système physique définie par Ludwig Boltzman [1872]

Cette approche qui consiste en la modélisation des éléments constitutifs ou observations répétées de la structure par des systèmes physiques s'étend à des classes d'individus donc à un ensemble beaucoup plus vaste que celui étudié. Nous signalons que ces notions de systèmes physiques et d'entropie ont été développées de manière plus générale par Shanonn[90].

Le formalisme proposé donne une explication satisfaisante de la distance introduite par Kullback -Leibler. De plus, l'indice adapté permet de construire une hiérarchie indicée sur les éléments constitutifs d'une structure de juxtaposition de tableaux multiples et s'étend à de multiples applications et structures dans le cas le plus général.

Lerman [62] a proposé l'indice du lien de vraisemblance maximum qu'il a adapté à une structure de juxtaposition de tables de contingence mais cet indice et les nombreux indices proposés dans la littérature ne s'adaptent pas à notre structure. Parmi ces indices, on peut citer: l'indice de Ward, l'indice de Sokal & Sneath, l'indice d'Anderberg, l'indice de Cover,...Pour plus de précisions voir [7]

En résumé, cette thèse est découpée en 5 chapitres.

Dans le chapitre I nous introduisons les types de structure de données qui nous intéressent et nous présentons l'originalité et l'intérêt pratique.

Dans le chapitre II, nous étudions les différentes approches de type factorielles qui peuvent s'y adapter. Le cas d'une structure de juxtaposition de tableaux de mesure est largement étudié.

Dans le chapitre III, nous définissons 3 indices de ressemblance adaptés aux éléments constitutifs de la structure et détaillons les différentes étapes de l'algorithme des moyennes mobiles ou k-means[40]. L'algorithme utilisant la distance induite par le produit scalaire de *Hilbert-Schmidt* est programmé en MATLAB 5.3 et utilisé pour classifier les clients domestiques d'une entreprise de distribution d'électricité. Ce chapitre a fait l'objet d'une publication dans la revue *communications in statistics*.

Le chapitre IV de cette thèse constitue une nouvelle approche tant au niveau du formalisme que des concepts définis. Ainsi, nous introduisons la notion d'objets aléatoires et adaptons de manière naturelle un indice de distance. Cet indice est apparu, de manière presque inattendue, comme une nouvelle version de la distance de *Kullback-Leibler* définie pour d'autres applications. Les différentes étapes de l'algorithme de classification hiérarchique sont développées, un exemple déroulé manuellement et une application sur données réelles complétent ce chapitre dont une synthése a fait l'objet d'une publication dans *" Journal of Applied Mathematics and Decision Science"*

Le chapitre V est une extension des résultats établis dans le chapitre III vers une application à un problème classique et important . Il s'agit de la classification d'objets dont les descriptions sont entachées d'erreurs de mesure . Des approches sont proposées et validées sur des exemples.

Une conclusion, certaines perspectives de développement et une bibliographie commentée terminent ce travail.

Chapitre I
Structures de juxtaposition de tableaux multiples

Ces structures de données peuvent être obtenues dés que l'on est conduit à effectuer plusieurs observations sur les variables qui décrivent chaque individu. Les individus peuvent ne pas être décrits par le même groupe de variables et le nombre d'observations peut différer d'un individu à l'autre ou d'une variable à l'autre.

1 Structure de données à observations répétées

C'est une structure de données particulière de tableaux multiples fréquente en pratique. Elle se schématise comme suit

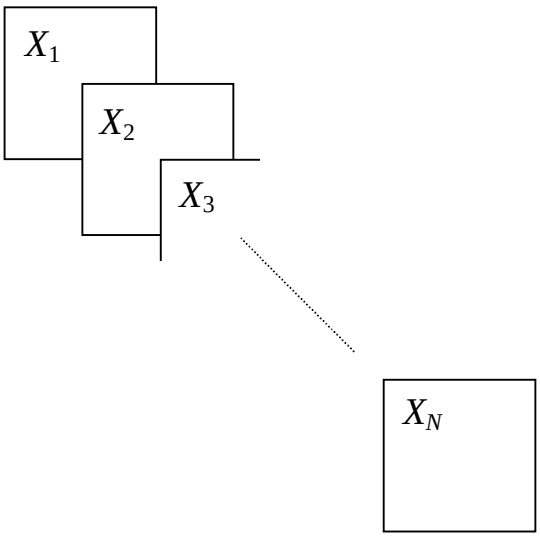


fig 1

1.1 Structure d'une juxtaposition de tableaux de mesure (TofM)

Soit Ω un ensemble fini de N individus, le tableau

$$X = [X_1, \ldots, X_N],$$

est constitué de la juxtaposition de N sous tableaux où

$$X_i = \left[x_{il}^j\right]_{j=1,\ d_i=\|P_i\|;l=1,\ L}\ ,$$

X_i est le sous tableau de dimension $L \times d_i$ contenant les observations répétées de l'individu $\omega_i \in \Omega$ pour le groupe de variables quantitatives (continues) P_i de cardinal d_i qui le décrivent x_{il}^j est la $l^{ième}$ observation de l'objet ω_i pour la variable V^j. Une observation θ_i est une ligne du tableau X. À cette observation dite « moyenne » est associée N individus $\{\hat{\imath}; i = 1, N\}$dits partiels correspondants aux divers sous tableaux X_i.

1.2 Structure de juxtaposition de tableaux de données qualitatives

Soit Ω un ensemble fini de N individus, T un descripteur de la forme

$$T = [T_1, \ldots, T_N]\,,$$

où

$$T_i = \left[m_{il}^j\right]_{\ j=1,\ d_i=\|P_i\|;\ l=1,L}$$

T est un tableau constitué de la juxtaposition de N sous tableaux $\{T_i;\ i = 1, n\}$,

T_i est le sous tableau de dimension $L \times d_i$ contenant les observations répétées de l'individu $\omega_i \in \Omega$ pour le groupe de P_i variables qualitatives qui le décrivent, m_{il}^j est la $l^{ième}$ modalité prise par l'objet ω_i pour la variable V^j.

Le tableau T_i regroupe les modalités prises par l'individu ω_i pour les variables qui le décrivent.

Cette structure est une structure de données brutes non exploitable sous cette forme pour son analyse, on adopte la démarche classique utilisée dans l'analyse factorielle des correspondances multiples qui consiste à associer à chacun des sous tableaux, le tableau disjonctif complet correspondant. Nous obtenons ainsi, une structure de juxtaposition de tableaux de données catégorielles. L'analyse de cette structure fera l'objet du chapitre IV de cette thèse.

2 Quelques exemples pratiques d'une structure de juxtaposition de tableaux multiples

Les exemples de structure de juxtaposition de tableaux de ce type sont nombreux en médecine où X_1,..., X_N sont des dossiers de malades regroupant les bilans effectués en

mesurant les différents paramètres et symptômes qui décrivent la maladie. Les tableaux n'ont pas la même dimension car le choix des paramètres et leurs mesures dépendent de l'état général du malade ou de la maladie.

En écologie végétale, on rencontre souvent cette structure de données particulière. En effet

Si $R_1,\ldots,\ R_L$ sont L relevés de la région étudiée, e_i^l l'espèce végétale i présente dans le relevé R_l , v_j^l le paramètre j mesuré sur son cortège floristique.

La structure de données dont on dispose en entrée se présente comme suit

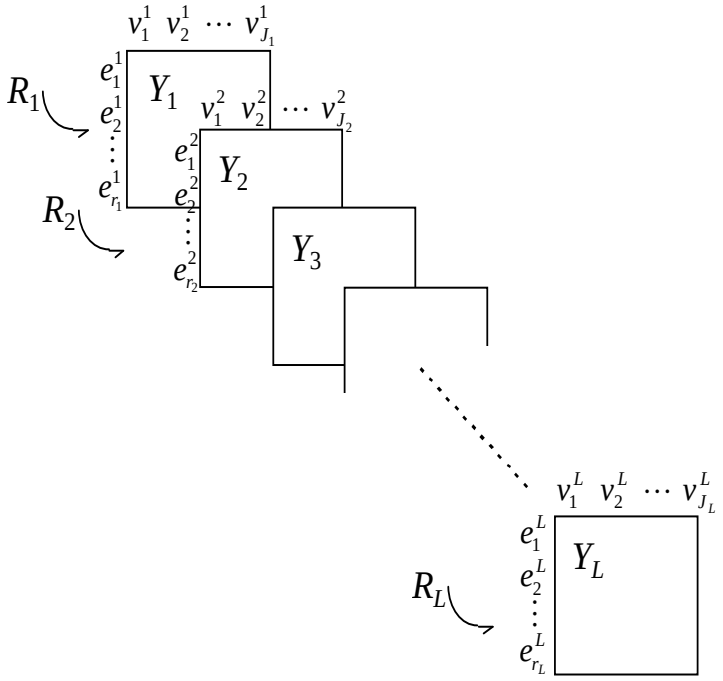


fig 2

Chaque relevé R_i est décrit par

$$Y_i \wedge V_i$$

où Yi est le tableau contenant les mesures faites sur son cortège floristique, est le vecteur contenant les mesures des paramètres édaphiques ou paramètres propres au relevé.

 # Un des problèmes posé est l'étude des liens pouvant existerentre le cortège floristique et les paramètres édaphiques donc entre les tableaux Y_i et les vecteurs V_i #.

Les paramètres édaphiques sont hétérogènes (qualitatifs et quantitatifs). Dans le cas où l'on pourrait les recoder en un paramètre qualitatif avec plusieurs modalités, ce problème devient un problème type que l'analyse discriminante peut résoudre.

En effet, soit

$$V = \begin{bmatrix} V_1 \\ \vdots \\ V_L \end{bmatrix}$$

le tableau contenant les descriptions des paramètres édaphiques des différents relevés.

Une typologie des relevés construite à partir de V conduit à recoder les paramètres par une variable qualitative dont les modalités sont les classes obtenues. Dans ce cas, on retrouve les conditions d'application de l'analyse factorielle discriminante. Cependant, cette situation n'est pas toujours possible et les résultats sont difficilement interprétables. De plus, dans la pratique, nous n'avons pas les mêmes espèces dans les différents relevés et les relevés sont caractérisés par d'autres paramètres tels que les caractérsistiques du sol: calcium, ensoleillement, pentes, etc... De plus, les variables ne sont pas les mêmes pour chaque tableau

$$\{Y_l \ ; l = 1, \ L\}$$

A partir de cette structure, on peut construire une structure du type précédent. En effet

Si $e_1,\ldots,\ e_N$ sont les différentes espèces rencontrées, on construit les tableaux $X_1,\ldots,$ X_N caractérisant respectivement chacune des espèces de la manière suivante.

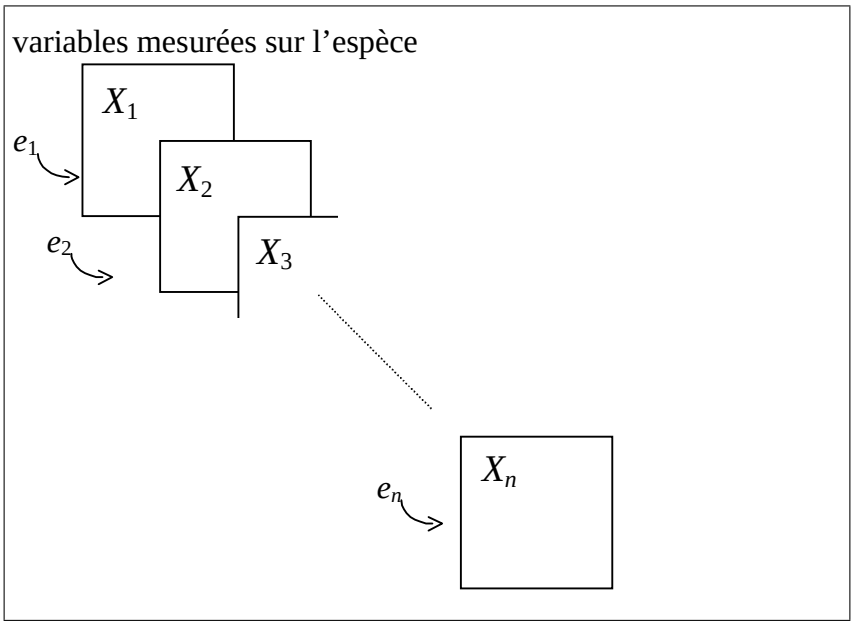


fig 3

Comme les espèces ne sont pas toutes présentes dans les différents relevés, les tableaux $X_1,\ldots,X_N$ n'ont pas la même dimension.

La structure de données obtenue est une structure de juxtaposition de tableaux multiples. Si le problème de l'analyse de tableaux multiples par des techniques factorielles

reste un vaste domaine de recherche Lebart, Morineau et Pion[58], plusieurs modèles et approches sont proposés pour comparer les tableaux, décrire leur structure commune et appréhender leur différence.

3 Quelques travaux antérieurs

Le fait que le théorème d'Eckart et Young[35][1935] de décomposition en éléments singuliers permettant de reconstituer partiellement ou totalement la structure de départ n'ait pas de généralisation dans le cas de tableaux multiples, a conduit certains chercheurs à développer des modèles particuliers aux disciplines et à la nature de leurs données.

On peut citer les premiers travaux sur l'analyse de tableaux multiples par les techniques factorielles qui remontent à Tucker[97][1964] et [98][1966], Harschman[45][1970]. On cite dans le même répertoire les travaux de Koomerberg[53][1983]; Carlier[14][1985]; de Von Der Heijder[101] ; Carlier et Coll[15][1988]; ...

Cuttman[26][1941] ; Burt[12][1950] et Hayashi[48][1956] sont les précurseurs de l'ACM " Analyse des correspondances multiples ". Des extensions ont été développées par Escofier et Cordier[1965]; Benzekri[7][1973]; Horst[49][1961]; Kettenring[50][1971];... L'analyse des correspondances multiples a été également développée sous le nom de " *Homogeneity analysis* " et sous le nom de " *Dual Scaling* " par Nishisato[70][1980]. La première application de l'*ACM* à un tableau disjonctif complet est l'œuvre de JP Nakache[69].

Les résultats et propriétés connus actuellement ont été mis en forme et programmés par Lebart et Tabard[1973]. Enfin, un exposé synthétique de certaines techniques a été réalisé par Tenenhaus et Young[1985]; Lebart, Morineau et Piron[58].

Dans le cas où les tableaux $X_1,\ldots,\ X_N$, regrouperaient les mêmes individus décrits par des groupes de variables différentes, l'AFC peut répondre au problème posé.

Escofier et L'hermier des Plantes[63] ont introduit la méthode STATIS " Structure des Tableaux À Trois Indices de la Statistique ". Développée par la suite par Lavit[56]. Cette méthode a pour but d'analyser un ensemble d'individus décrits par plusieurs groupes de variables ou d'analyser un ensemble de variables mesurées sur plusieurs groupes d'individus.

Dans le cas où notre structure se présente sous la forme

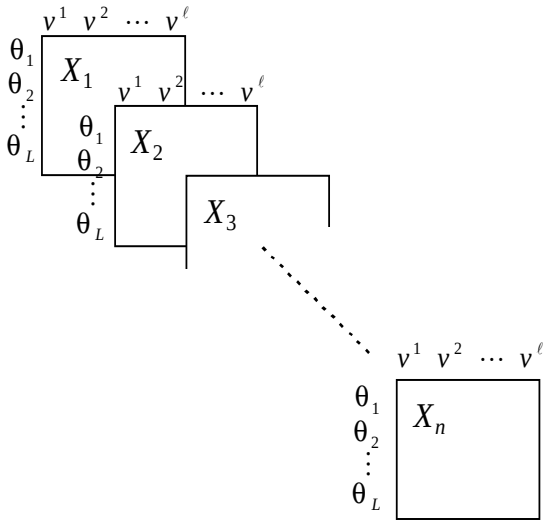


fig 4

où $\theta_1,\ldots,\theta_L$ sont considérés comme un ensemble de L individus décrits respectivement par les tableaux $X_1,\ldots,X_N$ de même dimension, la méthode STATIS permet effectivement d'étudier l'évolution des différents $\{\theta_i;\ i=1,L\}$, d'obtenir les liaisons entre les variables $\{V^j; j=1,\ p\}$ et de trouver une structure commune aux différents tableaux.

Le chapitre II de ce travail est consacré à des approches factorielles qui peuvent être adaptées et étendues aux données écologiques si $\theta_1,\ldots,\theta_L$ représentent les relevés $X_1,\ldots,X_N$ les tableaux décrivant les espèces. Ce sera une étude particulière qui nécessite que les mêmes espèces soient présentes dans les différents relevés, ce qui n'est pas la réalité qu'on désire appréhender.

Une autre approche factorielle peut être utilisée dans le cas où dans la structure toutes les variables seraient quantitatives et les tableaux de même dimension.

Cette approche DACP " *Double Analyse en Composantes Principales* " introduite par Bouroche[10] s'adapte parfaitement mais l'interprétation des résultats n'est pas évidente ce qui limite considérablement son application. De plus, elle se repose essentiellement sur la recherche du référentiel commun de représentation, qui demeure, dans l'optique des techniques proposées, un problème assez ouvert.

Cette approche s'articule autour de 3 phases distinctes.

La première étape consiste à l'analyse du phénomène d'évolution globale. Cette évolution est étudiée par une analyse en composantes principales des centres de gravité des nuages associés aux L tableaux $Y_1,\ldots,Y_N$. Elle correspond à ce qui est appelé: *l'étude de l'interstructure.*

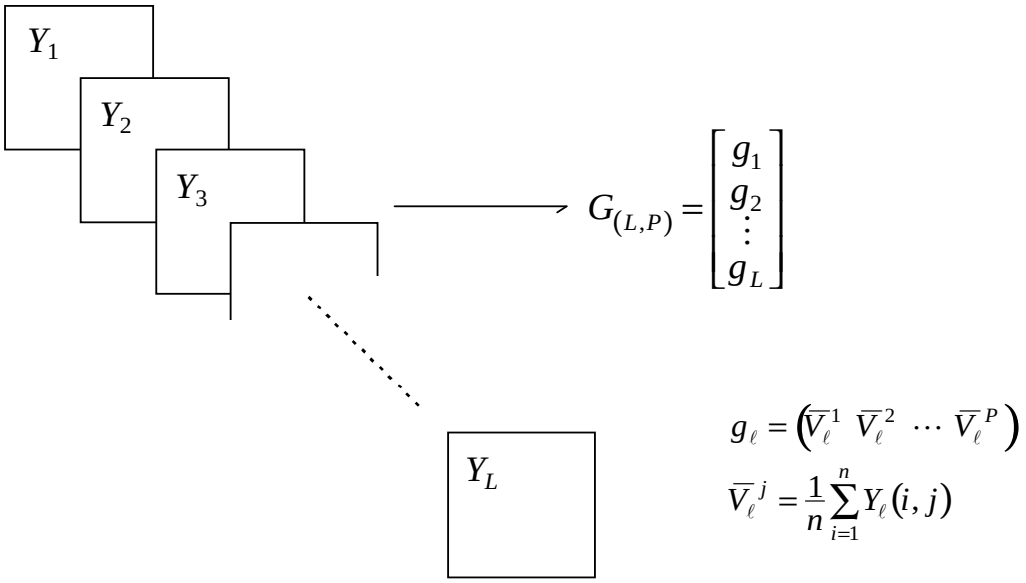


fig 5

Chaque tableau est résumé par le vecteur contenant le centre de gravité du nuage associé. L'*ACP* du tableau G permet d'étudier l'évolution des relevés (dans le cas de données écologiques) par l'intermédiaire des centres de gravité.

La seconde étape consiste à étudier la déformation des nuages autour de leur centre de gravité. Pour ce faire, Bouroche[10] propose d'effectuer L analyses en composantes principales (*ACP*) des L nuages de points associés aux tableaux centrés. Ainsi, cela permet d'éliminer le phénomène d'évolution globale.

La troisième phase consiste à rechercher un espace de représentation des espèces commun Bouroche [10]propose 4 procédures heuristiques permettant de déterminer un référentiel commun.

Cette étape correspond à ce qui est appelé : *étude de l'intrastructure*.

Outre les inconvénients propres aux différentes approches pour analyser des tableaux multiples la structure de données dont on dispose n'obéit pas, dans le cas général, aux hypothèses qui sous-tendent ces techniques. Les tableaux juxtaposés n'ont ni le même nombre de lignes, ni le même nombre de colonnes.

Cette structure de données en entrée n'a pas eu non plus, l'attention en classification automatique, à la hauteur de sa fréquence dans les cas pratiques.

Plusieurs structures de classification automatique existent, les plus courantes sont les partitions, hiérarchies, arbres et pyramides. Pour obtenir une partition à K classes (K fixé), la pratique courante consiste à optimiser un critère basé sur l'inertie. L'espace où sont définis les objets est muni d'une métrique M. Une solution est obtenue par un algorithme itératif du type k-means ou centres mobiles, Thordike[96][1953]; Forgy[40][1965]; Ball and Hall[5][1967] ; Macqueen[65]; Diday[31], ...

Il existe plusieurs variantes de ces algorithmes qui se distinguent principalement par certains aspects tels que : les noyaux, centres ou représentants des classes et les métriques sur l'espace où sont définis les objets à classifier Benzecri[7][1973] ; Anderberg[2][1973]; Diday[31][1980]; ...

Dans le cas où chaque tableau X_i de la structure regrouperait les valeurs multiples et non prévisibles à priori, prises par les variables qui décrivent l'objet correspondant et si les objets sont décrits par le même groupe de variables, une modélisation et une approche classification par des partitions dont le critère est basé sur la distance induite par le produit scalaire de *Hilbert-Schmidt* sont proposées dans le chapitre III . Un programme informatique en MATLAB 5.3 est déroulé pour classifier les clients d'une entreprise de distribution d'électricité.

La structure de données se présente dans ce cas

A l'objet ω_i correspond le tableau X_i		V^1	V^2	$\cdots$	V^P	
	θ_1	$\times$	$\times$	$\cdots$	$\times$	$\rightarrow E_1^i$
$\rightarrow$	θ_2	$\times$	$\times$	$\cdots$	$\times$	$\rightarrow E_2^i$
	$\vdots$					
	θ_{L_i}	$\times$	$\times$		$\times$	$\rightarrow E_P^i$

tab 1

Cette modélisation exige que les variables soient observées le même nombre de fois pour chaque objet qu'elles décrivent. Les tableaux $X_1, \ldots, X_n$ ont donc le même nombre de lignes.

Ce cas particulier, même s'il ne s'étend pas au cas général, est proche de la réalité car les contraintes exigées (même groupe de variables pour les objets) sont fréquentes dans toutes les techniques classiques d'estimation. Les lignes de chacun des tableaux, correspondent aux observations des objets pour les variables qui les décrivent. Ces observations doivent être prises dans les mêmes conditions si l'on veut que les comparaisons aient un sens.

Enfin, la distance de Kullback-Leibler s'adapte au cas d'une structure où les tableaux n'ont pas le même nombre de lignes mais doivent avoir le même nombre de colonnes càd qu'il est nécessaire que les individus soient décrits par les mêmes variables.

4 Problèmes posés

Lerman[60] [62] a proposé l'algorithme du lien de vraisemblance qu'il a adapté pour traiter les éléments constitutifs d'une structure de données sous forme d'une juxtaposition de tables de contingence mais la structure dont on dispose, nouvelle par sa forme ne pas être transformée pour être appréhendée par l'approche de Lerman[62].

Il s'agit donc de définir une autre manière de décrire les objets. La description devrait tenir compte

1. *De la variabilité des observations.*
2. *Du caractère aléatoire des observations.*
3. *Du fait que les objets ne sont pas décrits par le même groupe de variables.*
4. *Et enfin du fait que les objets ou variables ne sont pas observés le même nombre de fois.*

- Il faut de plus, que cette description offre des possibilités graphiques et de simulation.

Chapitre II

Approches factorielles pour analyser les éléments constitutifs d'une juxtaposition de n tableaux à observations multiples

1 Analyse après réduction

On suppose dans ce paragraphe que les objets sont décrits par le même groupe de variables et observés le même nombre de fois. Soit X la structure de données suivante

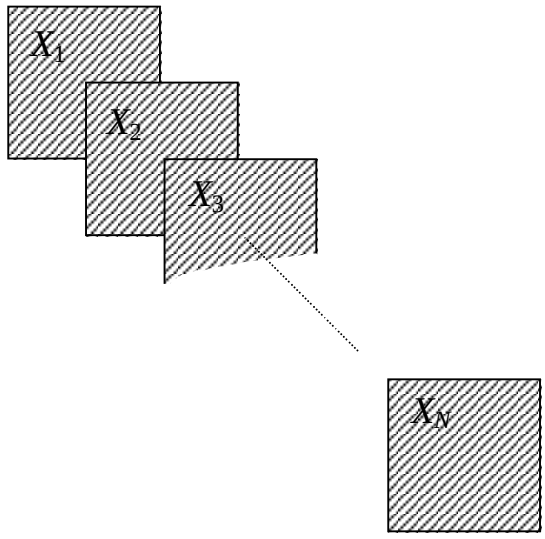


fig 6

où les n blocs de sous tableaux ont la même dimension.
Pour tout $i = 1,\ldots, n$; X_i est un tableau de dimension $(L \times d)$ tel que

$$X_i = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ \ell \\ \vdots \\ L \end{array} \begin{array}{c} \begin{array}{cccccc} V^1 & V^2 & \cdots & V^k & \cdots & V^d \end{array} \\ \begin{bmatrix} & & & & & \\ & & & \vdots & & \\ & & \cdots & x_{i\ell}^k & & \\ & & & & & \\ & & & & & \end{bmatrix} \end{array}$$

tab 2

x_{il}^k est la $l^{ième}$ observation de l'objet ω_i pour la variable, L est le nombre d'observations commun à toutes les variables qui décrivent les individus. d est le nombre de variables dans chaque groupe décrivant chaque individu.

On a

$$M = n \times d$$

Suivant le type de variables qui interviennent dans la description des individus plusieurs approches du type factoriel peuvent être adaptées pour comparer les différents tableaux, dégager leur structure commune et appréhender leurs différences.

Il est à noter que les observations qui constituent les lignes des sous-tableaux de la structure peuvent jouer le rôle d'individus. Les techniques factorielles multiples permettent à l'exploration simultanée des tableaux qui décrivent le même ensemble d'individus.

Souvent les observations ne jouent pas le rôle d'objets en pratique.

2 *Les L observations de chaque individu pour chaque variable qui le décrit sont résumées par une seule valeur*

2.1 Cas continu

Pour tout $i = 1, n$ le tableau X_i contenant les L observations répétées de l'individu ω_i pour les variables quantitatives (continues) qui le décrivent

$$X_i = \left[x_{il}^j\right]_{\substack{l=1,\,L \\ j=1,\,d}} \;\; ; \;\; x_{il}^j \in \mathbb{R}$$

Soit $x_{i.}^{j}$ la $jième$ colonne de cette matrice, on désire résumer cette colonne par une seule valeur. Deux approches sont possibles pour répondre à ce problème

- Faire un ajustement colonne par colonne pour chaque tableau.
- Faire un ajustement global du nuage associé aux d colonnes par un sous-espace affine ou un sous-espace vectoriel de dimension 1 si le nuage est centré.

Cette dernière technique est la première étape de la méthode $DACP$ (Double Analyse en Composantes Principales) introduite par Bouroche[10] et consiste à l'analyse du phénomène d'évolution globale des différents tableaux.

2.2 Ajustement point par point

Dans ce cas, $X_{i.}^{j}$est une série statistique de L observations de la variable V^j. Cette série peut être représentée géométrique par un vecteur de $\mathbb{R}^L$. La technique d'usage pour résumer la série par un seul point nécessite le choix d'une métrique sur et la résolution du problème d'optimisation lié à cette métrique

$$Min\left\{D^2\left(X_{i.}^{j},\ t.J_L\right),\ t\in\mathbb{R}\right\}$$

J_L est le vecteur de $\mathbb{R}^L$ dont toutes les composantes valent 1. Si D est la métrique euclidienne classique, l'optimum t_{ij}^{*} est atteint pour la moyenne empirique de la série.

D'autres métriques conduisent à la médiane, au mode de la série comme solution de ce problème . En général, ce problème n'admet pas toujours de solution.

Le choix de la métrique repose sur la forme de la distribution ou de la fonction de répartition représentée graphiquement par un histogramme ou par un diagramme en bâtons. La découpe en classes de la série pose le problème fondamental qui est "*l'effet partition*": un changement de la partition de l'espace des observations peut modifier le graphe de l'histogramme, gommer ou faire apparaître des symétries ou des modes... De plus, il est difficile de mesurer la qualité de l'ajustement. Enfin, cet ajustement ne permet pas de reconstituer même partiellement la série (les données de départ) car l'application qui associe à la série sa valeur centrale n'est pas injective.

2.3 Ajustement du nuage global par une droite

Le tableau X_i peut être considéré comme un nuage de d points de $\mathbb{R}$

$$X_i=\left[X_{i.}^{1},\ldots,\ X_{i.}^{d}\right]$$

.Sans restreindre la généralité, supposons que les variables sont centrées et que $\mathbb{R}^L$ est muni de la métrique euclidienne classique. On a

$$\overline{X}_{i.}^{j}=\frac{1}{L}\sum_{t=1}^{L}x_{it}^{j}=0$$

Le problème posé est un problème de réduction numérique. Autrement dit, un problème de compression des données. Parmi les critères d'ajustement d'un nuage de d points par un sous-espace vectoriel, celui retenu et qui conduit probablement aux calculs analytiques les plus simples est le critère classique des moindres carrés. Il consiste à rechercher dans notre cas, la droite d'allongement maximum du nuage de points et donc à rendre minimale la somme des carrés des écarts

$$\sum_{j=1}^{d} M_j H_j,$$

M_j et H_j sont deux points $\mathbb{R}^L$de tels que

$$\overrightarrow{OM_j} = X_{i.}^{j},$$

H_j est la projection orthogonale de M_j sur Δ_u # l'axe engendré par le vecteur unitaire $\overrightarrow{u}$# et passant par l'origine centre de gravité du nuage "*condition nécessaire pour qu'une droite satisfasse le critère des moindres carrés.*"

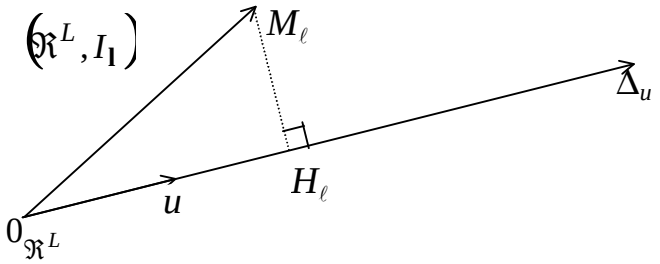


fig 7

La droite Δ_u qui réalise le meilleur ajustement du nuage des d points de $\mathbb{R}^L$au sens des moindres carrés est la droite passant par l'origine de $\mathbb{R}^L$ engendrée par le vecteur propre unitaire U_1^* associé à la plus grande valeur propre $\lambda_1$de la matrice carrée d'ordre d,

$$X_i'.X_i.$$

La qualité d'ajustement se mesure par

$$T_1 = \frac{\lambda_1}{\sum_{t=1}^{d} \lambda_t} = \frac{\lambda_1}{tr\left(X_i'.X_i\right)}$$

Dans ce cas, chaque vecteur $X_{i.}^{j}$ pour tout $j = 1, d$ est résumé par sa projection sur la droite

$$X_{i.}^{j} \longrightarrow y_i^j = \left(X_{i.}^{j}\right)'.u \in \mathbb{R}.$$

Plusieurs travaux ont été menés pour établir la loi de la matrice

$$S = X_i'.X_i$$

dans le but d'expliciter la loi de ses valeurs propres. Sous certaines hypothèses, il a été montré que S suit la loi de *Wirshart*. Fisher[39][1939] a établi la densité de probabilité des valeurs propres issues d'une matrice de *Wirshart*. Dans Anderson[3] sont données la densité de probabilité de *Wirshart* et de certaines lois dérivées.

Il est apparu que l'utilisation du taux d'inertie T_1 comme outil d'évaluation globale de la qualité de représentation est très délicat. T_1 représente une part de la variance brute initiale qui n'est pas une mesure de référence adéquate. Benzecri[7] utilise un taux d'inertie corrigé dans le cas de l'analyse des correspondances d'un tableau disjonctif complet. Ce taux est donné sous la forme

$$\tau\left(\lambda\right) = \left(\frac{d}{d-1}\right)\left(\lambda - \frac{1}{d}\right),$$

pour $\lambda \succ \frac{1}{d}$ où d est le nombre de variables actives.

L'ajustement global du nuage par une droite fait appel à une forme quadratique définie positive et à un ajustement par minimisation d'un critère lié à cette forme.

D'autres approches sont possibles en modifiant le type de distance, la nature du sous espace vectoriel ou les deux en même temps. D'autres critères peuvent être aussi utilisés; à la méthode des moindres carrés, basée sur la norme L^2, on peut substituer celles des moindres valeurs absolues basée sur la norme L^1 Van Cutsen[99][1994]; ...

2.4 L'une des variables est qualitative

Si V^j est une variable qualitative à r_j modalités $\{m_l;\ l = 1,\ r_j\}$ décrivant l'individu ω_i. La colonne j du tableau X_i est de la forme

$$X_i = \left[x_{i.}^{1}, \ldots,\ x_{i.}^{d}\right],$$

où $x_{i.}^{j}$ est l'une des modalités.

Le tableau X_i est un tableau hétérogène donné sous forme de codage brut. Il est inexploitable sous cette forme.

Devant un tel tableau, nous adoptons l'attitude des factorialistes qui consiste à transformer les variables continues en variables qualitatives dont les modalités sont les classes de l'histogramme correspondant. Les techniques qui peuvent être utilisées pour ce recodage sont nombreuses. Suivant la forme de la dispersion des points sur l'axe des réels, on peut adapter, si le nombre de classes est fixé à priori, la méthode hiérarchique ascendante à une variable et qui consiste à

1. *Choisir un indice d'agrégation*
2. *Dérouler un algorithme du type hiérarchique jusqu'à agréger les L valeurs en k classes.*

$$X^j = \begin{bmatrix} x_{i1}^j \\ \vdots \\ x_{iL}^j \end{bmatrix} \longmapsto \widetilde{X^j} = \begin{bmatrix} y_{i1}^j \\ \vdots \\ y_{ik}^j \end{bmatrix} \text{ où } y_{il}^j = m_t^j \text{ si } x_{il}^j \in \text{ à la classe } t.$$

Ainsi le tableau X_i peut être transformé en un tableau disjonctif complet Z_i ou en un tableau de Burt B_i.

L'ajustement de X_i se fait à partir de Z_i ou B_i.

2.5 Ajustement global du nuage

2.5.1 Tableau disjonctif complet associé à X_i

Soit L le nombre d'observations répétées sur chaque variable décrivant les individus $\{\omega_i; i\}$. Posant

$$P = \sum_{j=1}^{d} r_j,$$

r_j nombre de modalité ou classes de la variable V^j.

On construit à partir de X_i le tableau Z_i à L lignes et P colonnes décrivant les réponses des L observations pour les d variables par un codage binaire. Le tableau Z_i est la juxtaposition de d sous-tableaux

$$Z_i = \left[Z_{i.}^1, \ldots, \ Z_{i.}^{d_i} \right],$$

$$
Z_i^j = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ \ell \\ \vdots \\ L \\ \text{marge en} \\ \text{colonnes} \end{array}
\begin{array}{c}
\begin{array}{cccccccccc} m_1^j & m_2^j & \cdots & & \cdots & m_k^j & & \cdots & m_{r_j}^j \end{array} \\
\begin{bmatrix}
0 & 0 & \cdots & 1 & 0 & 0 & \cdots & \cdots & 0 \\
\vdots & & & & & & & & \\
0 & 0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 \\
& & & & & & & &
\end{bmatrix}
\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{array} \\
\begin{array}{cccccc} \left(Z_i^j\right)_{\bullet 1} & \cdots & \left(Z_i^j\right)_{\bullet k} & \cdots & \left(Z_i^j\right)_{\bullet r_j} \end{array}
\end{array}
$$

Z_i^j est un tableau de dimension $(L \times r_j)$.

Le sous-tableau Z_i^j est tel que sa lième ligne contient $r_j - 1$ fois la valeur 0 et une fois la valeur 1 dans la colonne correspondante à la modalité de la variable V^j observée à la lième observation.

$$\left(Z_i^j\right)_{.\,k} = \sum_{t=1}^{L} \left(Z_i^j\right)_{ttk},$$

où $\left(Z_i^j\right)_{ttk}$est le terme général du tableau Z_i^j.

Le tableau $Z_i = \left[Z_{i.}^1, \ldots, Z_{i.}^{d_i}\right]$ est donc de la forme

$$Z_i = \left[Z_{i.}^1\right] \ldots \left[Z_{i.}^{d_i}\right].$$

Soit t_{lk}^r le terme général du tableau Z_i ; $t_{lk}^r = 0$ ou 1 pour tout $l = 1$, L et $k = 1$, P.

Les marges en lignes du tableau Z_i (disjonctif complet) sont constantes et égales à d.

Les marges en colonnes sont données par

$$\left(Z_i\right)_{.j} = \sum_{k=1}^{L} t_{lk}^r$$

Cette marge correspond au nombre de fois que la modalité où la classe j a été observée pour l'individu ω_i.

Pour chaque sous-tableau $Z_{i.}^j$ l'effectif total est donné par

$$\sum_{k=1}^{r_j} \left(Z_i^j\right)_{.k} = \sum_{t=1}^{L} \sum_{k=1}^{r_j} \left(t_{lk}^r\right) = L.$$

L'effectif total du tableau disjonctif complet est

$$z_i = \sum_{t=1}^{L} \sum_{k=1}^{P} (t_{lk}^r) = L \times d.$$

2.5.2 Tableau de contingence de Burt associé

On construit à partir de Z_i, le tableau symétrique B_i d'ordre P qui rassemble les croisements deux à deux de toutes les variables qui décrivent l'individu ω_i.

$$B_i = (Z_i)' \, Z_i.$$

Le terme général de B_i s'écrit

$$b_{jj'}^i = \sum_{l=1}^{L} z_{jj'}^l.$$

B_i est une juxtaposition de tableaux de contingence.

- Les marges sont pour tout $j \leq P$ (identiques pour les lignes et les colonnes de par la symétrie de B_i) données par

$$b_{j''}^i = \sum_{j=1}^{P} b_{jj'}^i = d.\,(Z_i)_{.j'}$$

- L'effectif total b_i de B_i vaut

$$b_i = \sum_{j=1}^{P} \left(\sum_{j'=1}^{p} b_{jj'}^i \right) = d. \sum_{j'=1}^{p} (Z_i)_{.j'} = d \left(. \sum_{j'=1}^{p} \sum_{k=1}^{L} Z_{kj}^i \right)$$

$$= d \left(. \sum_{k=1}^{L} \underbrace{\sum_{j=1}^{P} Z_{kj}^i}_{d} \right) = d \left(\sum_{k=1}^{L} d \right) = d^2.L$$

$$Z_i = \left[\left[Z_{i.}^1 \right] \ldots \left[Z_{i.}^{d_i} \right] \right] \Rightarrow B_i = (Z_i)' \, Z,$$

le tableau B_i est formé de d^2 blocs. Le bloc$(Z_i^q)'.Z_i^{q'}$ indicé par q et $q\prime$ est une matrice rectangulaire de dimension $r_q.r_{q'}$ qui est le tableau de contingence croisant les variables V^q et $V^{q'}$. Si $q = q\prime$; $(Z_i^q)'.Z_i^q$ est un bloc carré de dimension r_q obtenu par le croisement de la variableV^q par elle même.

C'est donc une matrice diagonale de dimension r_q qu'on note W_i^q

$$W_i^q = (Z_i^q)'.Z_i^q,$$

Les éléments diagonaux sont donnés par

$$(W_i^q)_l = (Z_i)_{.l}\,.$$

Soit D^i la matrice diagonale d'ordre P ayant les mêmes éléments diagonaux que B_i

$$d_{ll}^i = (D_i)_{ll} = b_{ll}^i = (B_i)_{ll} = (Z_i)_{.l}\,.$$

La matrice D^i peut être considérée comme une matrice à d^2 blocs où seules les d matrices diagonales $\left\{D_q^i; q-1, d\right\}$ ne sont pas nulles. On a

$$D_q^i = (Z_i^q)'.Z_i^q = W_i^q.$$

2.5.3 Critère d'ajustement et distance du χ^2.

Les observations sont toutes affectées d'une masse $m_l = \frac{1}{L}$. Chacune des modalités m_j est affectée du poids $\frac{1}{dL}\ (Z_i)_{.j}$. La distance de χ^2 appliquée à un tableau disjonctif complet conserve un sens. En effet, dans $\mathbb{R}^L$, la distance entre modalités s'écrit

$$d^2\left(j, j'\right) = \sum_{l=1}^{L} L\left[\frac{Z_{lj}^i}{(Z_i)_{.j}} - \frac{Z_{lj'}^i}{(Z_i)_{.j'}}\right]^2,$$

ainsi, deux modalités obtenues par les mêmes observations coïncident. Par ailleurs les modalités de faible effectif sont éloignées des autres modalités. La distance entre deux observations θ_l et $\theta_{l'}$ s'exprime par

$$d^2\left(l, l'\right) = \frac{1}{d_i}\sum_{l=1}^{L}\frac{L}{(Z_i)_{.j}}\left[z_{lj}^i - z_{l'j}^i\right]^2,$$

ainsi deux observations sont proches si les résultats sont les mêmes modalités, elles sont éloignées sinon.

Soient

- $F_i = \frac{1}{dL} Z_i$ de terme général $f^i_{lj} = \frac{Z^i_{lj}}{d.L}$,
- $D^i_P = \frac{1}{dL} D^i$ de terme général $f^i_{.j} = \delta_{lj} \frac{(Z_i)_{.j}}{dl}$
- $D^i_L = \frac{1}{L}.I_L$(I_L est la matrice identitée d'ordre L) de terme général $f^i_{.j} = \frac{\delta_{lj}}{L}$

Pour trouver les axes factoriels, on diagonalise la matrice

$$S_i = (F_i)' \left(D^i_L\right) F_i.D^i_P = \frac{1}{d} (Z_i)' Z_i \left(D^i\right)^{-1},$$

Le terme général de S_i est donné par

$$\left(S^i\right)_{j_{j_{\grave{u}}}} = \frac{1}{d (Z_i)_{.j'}} \sum_{l=1}^{L} z^i_{lj}.z^i_{l'j},$$

Dans $\mathbb{R}^P$, l'équation du $\alpha^{ième}$axe factoriel U_α est donnée par

$$\frac{1}{d} (Z_i)' .Z_i.D_i^{-1}.U_\alpha.$$

L'équation du $\alpha^{ième}$ facteur $\varphi_\alpha = D_i^{-1}.U_\alpha$ vérifie

$$\frac{1}{d} (Z_i)' .Z_i.\varphi_\alpha = \lambda_\alpha.$$

De même que l'équation du $\alpha^{ième}$facteur Ψ_α dans $\mathbb{R}^L$s'écrit

$$\frac{1}{d} Z_i D_i^{-1}. (Z_i)' .\Psi_\alpha = \lambda_\alpha \Psi_\alpha.$$

Les facteurs φ_α et Ψ_α représentent les coordonnées des points lignes et les points colonnes sur l'axe factoriel Δ_{U_α}. Les relations de transition entre facteurs sont

$$\begin{cases} \Psi_\alpha = \frac{1}{d\sqrt{\lambda_\alpha}} Z_i.\varphi_\alpha \\ \varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} D_i^{-1} (Z_i)' .\Psi_\alpha \end{cases}$$

Les coordonnées de l'observation θ_l sur l'axe factoriel Δ_{U_α}sont données par

$$(\Psi_\alpha)_l = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{P} \frac{Z_{lj}^i}{(Z_i)_{.j}} (\varphi_\alpha)_j = \frac{1}{d\sqrt{\lambda_\alpha}} \sum_{j \in P(l)}^{P} (\varphi_\alpha)_j$$

où $P(l)$ désigne l'ensemble des modalités obtenues à l'observation θ_l
La coordonnée de la modalité j sur l'axe factoriel Δ_{U_α} est donnée par

$$(\varphi_\alpha)_j = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{l=1}^{L} \frac{Z_{lj}^i}{(Z_i)_{.j}} (\Psi_\alpha)_l = \frac{1}{(Z_i)_{.j}\sqrt{\lambda_\alpha}} \sum_{j \in I(j)}^{P} (\Psi_\alpha)_l$$

où $I(J)$ désigne l'ensemble des observations ayant été la modalité j.

2.5.4 Analyse du tableau de Burt: équivalence avec l'analyse du tableau disjonctif complet

L'analyse des correspondances appliquée à un tableau disjonctif complet Z_i est équivalente à l'analyse du tableau de Burt [12] B_i correspondant et produit les mêmes facteurs. En effet

L'analyse des correspondances du tableau B_i (symétrique d'ordre $P \times P$) se ramène à l'analyse d'un nuage de P points-modalités dans $\mathbb{R}^P$. Les marges de ce tableau, en ligne et comme en colonne sont les éléments diagonaux de la matrice D_i.

Compte tenu de l'équation donnant le $\alpha^{iéme}$ facteur φ_α de l'analyse du tableau disjonctif complet Z_i

$$\frac{1}{d} D_i^{-1} (Z_i)' . Z_i . \varphi_\alpha = \lambda_\alpha \varphi_\alpha (*) .$$

La matrice à diagonaliser dans le cas l'AC de Z_i est

$$S_i = \frac{1}{d} D_i^{-1} (Z_i)' . Z_i = \frac{1}{d} D_i^{-1} B_i .$$

Pour l'analyse du tableau B_i associé à Z_i, le tableau des fréquences relatives F_i s'écrit

$$F_i = \frac{1}{L.d} B_i ; D_P^{-1} = D_L^{-1} = \frac{1}{L.d} D_i^{-1} .$$

On diagonalise la matrice

$$S_i^* = \frac{1}{d_i} D_i^{-1} B_i . D_i^{-1} B_i \, (*) \Rightarrow S_i^* = (S_i)^2 .$$

En multipliant les deux membres de$(*)$ par $\frac{1}{d_i} D_i^{-1} B_i$ on obtient

$$\frac{1}{d^2} D_i^{-1} B_i . D_i^{-1} B_i \varphi_\alpha = \lambda_\alpha^2 \varphi_\alpha .$$

Les facteurs des deux analyses sont donc colinéaires dans $\mathbb{R}^P$ mais les valeurs propres associées sont différentes. On a la relation

$$\lambda_{B_i} = \lambda_{Z_i}^2 ,$$

au sens où les facteurs issus de l'analyse de Z_i, représentent les coordonnées factorielles des modalités ont pour norme λ_{Z_i}. Alors le facteur correspondant de l'analyse de B_i, noté φ_{B_i} a pour norme

$$\lambda_{B_i} = \lambda_{Z_i}^2 .$$

Relations liant les deux systèmes de coordonnées factorielles

$$\varphi_{B_i \alpha} = \varphi_\alpha \sqrt{\lambda_\alpha} .$$

Pour résumer le tableau X_i par une droite, on choisira la droite qui donne la plus grande inertie, c'est à dire celle engendrée par le premier axe factoriel dans ce cas

$$\begin{array}{ccc} & Z_i & \\ \nearrow & & \searrow \\ X_i & \longrightarrow & Y_1 \in \mathbb{R}^P \end{array}$$

dont les coordonnées sont données par

$$(Y_1)_j = (\varphi_1)_j = \frac{1}{(Z_i)_{.j} \sqrt{\lambda_\alpha}} \sum_{j \in I(j)}^{P} (\Psi_1)_l$$

où Ψ_1 est le vecteur propre associé à la plus grande valeur λ_1 non triviale de la matrice

$$\frac{1}{d}\left(Z_i\right)'.Z_i.D_i^{-1}$$

Cette étude préliminaire à l'analyse globale d'une juxtaposition de tableaux à observations multiples se résume ainsi

Cas continu

$$\boxed{X'_{(M=n\times d,L)} = \begin{bmatrix} X'_1 \\ \vdots \\ X'_n \end{bmatrix}} \xrightarrow{\text{aprés réduction}} Y'_{(n,d)} = \begin{bmatrix} Y''_1 \\ \vdots \\ Y'_n \end{bmatrix}$$

$$\xleftrightarrow{\text{analyse par des méthodes classiques}}$$

Cas qualitatif

$$Z'_{(nP,L)} = \begin{bmatrix} Z'_1 \\ \vdots \\ Z'_n \end{bmatrix}$$

$$\nearrow \qquad \downarrow \qquad \searrow$$

$$X'_{(M=\mathrm{v}n\times d,\mathrm{v}L)} = \begin{bmatrix} X'_1 \\ \vdots \\ X'_n \end{bmatrix} \quad \downarrow Y'_{(n,\,d)} = \begin{bmatrix} Y'_1 \\ \vdots \\ Y'_n \end{bmatrix}$$

$$B'_{(nP,L)} = \begin{bmatrix} B'_1 \\ \vdots \\ B'_n \end{bmatrix} \nearrow$$

Cette approche n'est efficace que si l'inertie expliquée par le premier axe est suffisante. Ce qui n'est pas toujours le cas en pratique. Comme elle procède par double ajustements, elle peut servir comme étape préliminaire à l'analyse de grands tableaux.

3 Analyse sans réduction

3.1 Comparaison globale des tableaux : l'interstructure.

Dans cette partie, nous étudions les liens pouvant exister entre les tableaux décrivant les individus et les comparer globalement. On est conduit pour ce faire à

- Choisir des objets représentatifs des tableaux
- Choisir une métrique entre objets représentatifs
- Trouver une image euclidienne des objets représentatifs.

3.1.1 Objet représentatif du tableau X_k décrivant l'individu ω_k

Soit W_k la matrice des produits scalaires internes au tableau X_k qui décrit l'individu ω_k, on a

$$W_k = \underbrace{(X_k\ )}_{(l,\, r_k)}.\underbrace{(X_k)'}_{(r_k,\, L)} \text{ou} \quad W_k = X_k.Q_k\,(X_k)'$$

W_k est une matrice carrée d'ordre L, Q_k est une métrique interne aux observations de l'individu ω_k pour le groupe de variables qui le décrit Q_k peut être la métrique des poids affectés à chacune des observations.

L'objet W_k est représentatif de l'individu ω_k car il regroupe les produits scalaires entre les observations de l'individu pour les P_k variables qui le décrivent. W_k contient donc tous les liens inter observations. W_k est une matrice symétrique.

W_k peut être considérée comme un point de $\mathbb{R}^{L^2}$ en empilant ses L colonnes. Si $V^{j1}, \ldots, V^{jr_k}$ sont les r_k variables qui décrivent l'individu ω_k. On a

$$X_k = \left[x^k_{t\ js}\right]_{t=1,\ L;\ \ s=1,\ r_k}$$

.

Soient W^1_k,$\ldots, W^L_k$ les L colonnes de la matrice W_k.

Proposition 1

La $t^{ième}$coordonnée du vecteur W^l_kest donnée par

$$\left(W^l_k\right)_t = \sum_{m=1}^{r_k} \left(x^k_{t\ jm}\right)\left(x^k_{t\ jm}\right)$$

Si $\widehat{W_k}$ est le vecteur de $\mathbb{R}^{L^2}$obtenu en empilant les L colonnes de la matrice $\mathbf{W}_k$

$$\widehat{W_k} = \begin{bmatrix} [W^1_k] \\ \vdots \\ [W^L_k] \end{bmatrix} \text{ où } W^l_k = \begin{bmatrix} W^k_{1i} \\ \vdots \\ W^k_{Li} \end{bmatrix} \in \mathbb{R}^L.$$

Le tableau

$$W_{L^2} = \left[\widehat{W_1}, \ldots, \widehat{W_n}\right],$$

est le tableau contenant les n vecteurs de $\mathbb{R}^{L^2}$ en colonnes , W_{L^2} est un tableau de dimension $L^2 \times n$. On défini ainsi, un nuage de n points tableaux représentés par des vecteurs de $\mathbb{R}^{L^2}$et représentatifs des tableaux caractérisant les n individus .

3.1.2 Choix de la métrique sur $\mathbb{R}^{L^2}$

Afin de représenter graphiquement les n points tableaux représentatifs des n objets $\{\omega_1, \ldots, \omega_n\}$, il est nécessaire de définir une distance entre éléments représentatifs.

Produit scalaire de Hilbert-Schmidt Définiton 1

Le produit scalaire de Hilbert-Schmidt entre deux objets représentatifs $\widehat{W_k}$ et $\widehat{W_{k'}}$ est défini par

$$\prec \widehat{W_k}, \widehat{W_{k'}} \succ = tr\left(D.W_k, DW_{k'}\right),$$

où $\widehat{W_k}$ et $\widehat{W_{k'}} \in \mathbb{R}^{L^2}$; W_k et $W_{k'}$ sont des matrices symétriques d'ordre L, D : matrice diagonale d'ordre L contenant les poids affectés aux observations. trA est la trace de la matrice A.

Norme et distance induites par le produit scalaire de Hilbert-Schmidt. La norme d'un élément représentatif est donnée par

$$\left\|\widehat{W_k}\right\|_{HS} = \prec \widehat{W_k}, \widehat{W_k} \succ_{HS} = tr\left(D.W_k, DW_k\right)$$

Proposition 2

Si $\lambda_k^{(l)}$est la $l^{ième}$ valeur propre de la matrice $\mathbf{W}_k.D$. Alors

$$\left\|\widehat{W_k}\right\|_{HS}^2 = \sum_{l=1}^{L} \lambda_k^{(l)},$$

Preuve: Soient $\lambda_1, \ldots, \lambda p$ les p valeurs propres de la matrice A diagonalisable de dimension P. on a

$$A = Q^{-1} D_\lambda Q,$$

où $Q = [U_1, \ldots, U_p]$ avec U_i vecteur propre associé à λ_i de A, D_λ matrice diagonale dont les éléments sont les valeurs propres $\{\lambda_i;\ i = 1,\ p\}$. On a

$$tr\left(A\right) = \sum_{l=1}^{p} \lambda_l,$$

d'une part et d'autre part

$$A^2 = A \times A = \left(Q^{-1} D_\lambda Q\right)\left(Q^{-1} D_\lambda Q\right) = Q^{-1}\left(D_\lambda\right)^2 Q$$

or

$$\left(D_\lambda\right)^2 = \begin{pmatrix} \lambda_1^2 & & \\ 0 & . & 0 \\ & & \lambda_p^2 \end{pmatrix},$$

donc A_2 est aussi diagonalisable et

$$tr\left(A^2\right) = \sum_{l=1}^{p} \left(\lambda_l\right)^2$$

La distance entre deux éléments représentatifs $\widehat{W_1}$ et $\widehat{W_2}$ des individus ω_1 et ω_2 est définie à partir du produit scalaire de Hilbert-Schmidt de la manière suivante

$$d_{HS}\left(\widehat{W_1}, \widehat{W_2}\right) = \left\|\widehat{W_1} - \widehat{W_2}\right\|_{HS} = \sqrt{\left\|\widehat{W_1}\right\|^2 + \left\|\widehat{W_2}\right\|^2 - 2 \prec \widehat{W_1}, \widehat{W_2} \succ_{HS}}$$

Remarque 1

En pratique, il est fréquent que les éléments $\widehat{W_k}$ aient des normes très différentes les unes des autres. C'est une situation très délicate quant à l'interprétation des résultats car il est évident que les tableaux de norme assez élevée influent sur les résultats. On est donc conduit à normaliser les éléments représentatifs $\widehat{W_k}$,.

$$\widehat{W_k} \to \frac{\widehat{W_k}}{\left\|\widehat{W_k}\right\|_{HS}}.$$

L'analyse générale du tableau W_{L^2} en se plaçant sur $\mathbb{R}^n$ où l'espace vectoriel $\mathbb{R}^{L^2}$est muni de la métrique M_{HS} induite par le produit scalaire de Hilbert-Schmidt et $\mathbb{R}^n$est muni de la métrique euclidienne I_n conduit à la diagonalisation de la matrice carrée S d'ordre n. Le terme général de S est défini par

$$s_{kk'} = tr\left(W_k . W_{k'}\right).$$

Si les éléments représentatifs$\left\{\widehat{W_k}; k = 1,\ n\right\}$ ne sont pas normés et

$$s_{kk'} = \frac{tr\left(W_k.W_{k'}\right)}{\left\|\widehat{W_k}\right\|_{HS}\left\|\widehat{W_{k'}}\right\|_{HS}},$$

Si ces éléments sont normés

On se place dans $(\mathbb{R}^n; I_n)$, au tableau $\mathbf{W}_{L^2}$ correspond n points tableaux représentés dans $\mathbb{R}^{L^2}$par les colonnes et L^2 points vecteurs de $\mathbb{R}^n$ qui sont les lignes du tableau . On cherche à ajuster le nuage des L^2 points de $\mathbb{R}^n$ par des sous espaces vectoriels de dimension $1,\ldots, r$ qui satisfassent au critère des moindres carrés.

Remarque 2

Si les variables qui décrivent les différents individus sont centrées. Cette hypothèse implique les éléments représentatifs sont centrés et justifie l'ajustement du nuage par des sous espaces vectoriels au lieu de sous espaces affines. En effet

V^j centrée pour tout $j \in Q_i$ et pour tout $i = 1,\ n$, On a

$$\frac{1}{L}\sum_{h=1}^{L}\left(x_{hj}^i\right) = \overline{x_{.j}} = 0,$$

Or

$$\overline{W_t^i} = \frac{1}{L}\sum_{h=1}^{L}\left(W_{ht}^i\right) = \frac{1}{L}\sum_{h=1}^{L}\left(\sum_{m=1}^{r_t}\left(x_{hjm}^i\right)\left(x_{tjm}^i\right)\right) \Rightarrow$$

$$\overline{W_t^i} = \sum_{m=1}^{r_t} x_{tjm}^i\left(\frac{1}{L}\sum_{h=1}^{L}\left(x_{hjm}^i\right)\right) = \sum_{m=1}^{r_t} x_{tjm}^i.0 = 0$$

3.1.3 Recherche du sous espace vectoriel de dimension un.

On dispose d'un nuage de L^2 points de $(\mathbb{R}^n;\ I_n)$ qui représentent les lignes du tableau W_{L^2}.

Soit W^i la $i^{ème}$ ligne du tableau W_{L^2}; $W^i \in (\mathbb{R}^n;\ I_n)$

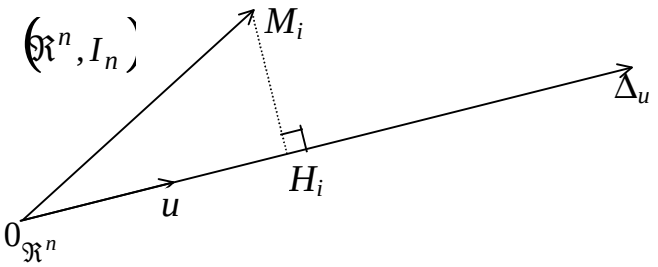


fig 8

H_i projection orthogonale de M_i sur Δ_u; $\overrightarrow{OM} = (W^i)'$. On a

$$\overline{OH_i} = W^i.u$$

En notant par

$$\mathbf{U} = \begin{bmatrix} \overline{OH_1} \\ \vdots \\ \overline{OH_n} \end{bmatrix} = \begin{bmatrix} W^1.u \\ \vdots \\ W^n.u \end{bmatrix} = W_{L^2}.u \in \left(\mathbb{R}^{L^2}, M_{HS}\right)$$

Il s'agit de rechercher u^* tel que l'on ait

$$\underset{u,\|u\|=1}{Max} \|U\|_{HS},$$

avec $\|U\|_{HS} = U'.M_{HS}.U$

Posant $M_{HS} = M$ pour faciliter les notations. On doit résoudre le problème d'optimisation suivant :

$$\underset{u,\|u\|=1}{Max} \left(u'.\left(W_{L^2}\right)'.M.\left(W_{L^2}\right).u\right),$$

le Lagrangien s'écrit

$$\varphi(u) = u'.\left(W_{L^2}\right)'.M.W_{L^2}.u - \lambda\left(u'.u\right),$$

condition nécessaire

$$\frac{\partial \varphi}{\partial u} = 0 \Leftrightarrow (W_{L^2})'.M.W_{L^2}.u = \lambda.u,$$

λ est donc valeur propre de la matrice $S = (W_{L^2})'.M.W_{L^2}.u$. Le maximum est obtenu en choisissant la plus grande valeur propre de S. L'analyse générale du tableau W_{L^2} conduit à la diagonalisation de la matrice S. Le terme général de la matrice à diagonaliser S est

$$(S)_{kk'} = s_{kk'} = \left(\widehat{W_k}\right)'.M.\left(\widehat{W_{k'}}\right) = \prec \widehat{W_k}, \widehat{W_{k'}} \succ_{HS} = tr\left(W_k.W_{k'}\right) \ ,$$

en effet

$$W_{L^2} = \left[\widehat{W_1}, \ldots, \widehat{W_n}\right] \Rightarrow (W_{L^2})' = \begin{bmatrix} \left(\widehat{W_1}\right)' \\ \vdots \\ \left(\widehat{W_n}\right)' \end{bmatrix} \Rightarrow$$

$$(W_{L^2})'.M.(W_{L^2}) = \left[\left(\widehat{W_k}\right)'.M.\widehat{W_{k'}}\right]_{k=1,L;k'-1,L} \Rightarrow s_{kk'} = tr\left(W_k.W_{k'}\right)$$

Si l'on norme des éléments représentatifs alors

$$s_{kk'} = \frac{tr\left(W_k.W_{k'}\right)}{\left\|\widehat{W_k}\right\|_{HS}\left\|\widehat{W_{k'}}\right\|_{HS}} (*)$$

$(*)$ est le coefficient Rv de [*Robert et Escofier*, 1976].

Cette analyse du tableau W_{L^2} permet de représenter les n points tableaux dans un espace de plus faible dimension et de comparer globalement les tableaux entre eux et par conséquent les individus qu'ils représentent.

Si tous les tableaux sont voisins, ils seront concentrés autour d'un point dans l'espace et le premier axe joindra l'origine à ce point. On pourrait au contraire voir les tableaux s'échelonner le long de cet axe et mesurer ainsi sur l'axe une sorte d'adéquation du tableau au modèle moyen.

Propriété et interprétation des coefficients Rv Distance entre deux éléments représentatifs normés $\widehat{W}_k$ et $\widehat{W}_{k'}$.

Proposition 3

$$d_{HS}\left(\widehat{W}_k, \widehat{W}_{k'}\right) = \sqrt{2\left(1 - Rv\left(k, k'\right)\right)}$$

où

$$Rv\left(k, k'\right) = s_{kk'} = \frac{tr\left(W_k.W_{k'}\right)}{\left\|\widehat{W_k}\right\|_{HS}\left\|\widehat{W_{k'}}\right\|_{HS}}$$

En effet (distances et produits scalaires calculés à partir de la définition d'Hilbert-Schmidt)

$$Rv\left(k, k'\right) = \prec \frac{\widehat{W_k}}{\left\|\widehat{W_k}\right\|_{HS}}, \frac{\widehat{W_{k'}}}{\left\|\widehat{W_{k'}}\right\|_{HS}} \succ_{HS}$$

et

$$d^2_{HS}\left(\widehat{W_k}, \widehat{W_{k'}}\right) = \prec \frac{\widehat{W_k}}{\left\|\widehat{W_k}\right\|_{HS}} - \frac{\widehat{W_{k'}}}{\left\|\widehat{W_{k'}}\right\|_{HS}}, \frac{\widehat{W_k}}{\left\|\widehat{W_k}\right\|_{HS}} - \frac{\widehat{W_{k'}}}{\left\|\widehat{W_{k'}}\right\|_{HS}} \succ$$

En utilisant la bilinéarité du produit scalaire, on obtient

$$d_{HS}\left(\widehat{W_k}, \widehat{W_{k'}}\right) = \sqrt{2\left(1 - Rv\left(k, k'\right)\right)}$$

Propriétés remarquables

$$Si\ Rv\left(k, k'\right) = 1 \Rightarrow d_{HS}\left(\widehat{W_k}, \widehat{W_{k'}}\right) = 0 \Rightarrow \frac{\widehat{W_k}}{\left\|\widehat{W_k}\right\|_{HS}} = \frac{\widehat{W_{k'}}}{\left\|\widehat{W_{k'}}\right\|_{HS}} = c_k$$

Cette égalité se traduit en pratique par le fait que les tableaux X_k et $X_{k'}$ sont équivalents dans le sens où les observations de l'objet $\omega_{k'}$ se déduisent des observations de l'individu ω_k par une homothétie de rapport c_k.

-Si $Rv\left(k, k'\right) = 0 \Rightarrow$ les variables qui décrivent l'objet ω_k sont non corrélées avec les variables qui décrivent $\omega_{k'}$

En effet,

pour l'individu $\omega_{k'}$, la variable est identifiée par le vecteur $X^{k'}_{.j}$de $\mathbb{R}^L$où

$$X_{k'} = \begin{bmatrix} ... \begin{bmatrix} X_{1i}^{k'} \\ \vdots \\ X_{Li}^{k'} \end{bmatrix} .... \end{bmatrix}$$

.

Pour l'individu ω_k , la variable est identifiée par le vecteur $X_{.j}^{k}$ de $\mathbb{R}^L$ où

$$X_k = \begin{bmatrix} .... \begin{bmatrix} X_{1i}^{k} \\ \vdots \\ X_{Li}^{k} \end{bmatrix} .... \end{bmatrix}$$

On suppose que $D = I_L$ (matrice identité d'ordre L), on ne privilégie donc aucune observation.

On a

$$W_k = \left[W_{t\ j}^{k}\right]_{t=1,\ L;\ j=1,L}; W_{k'} = \left[W_{t\ j}^{k'}\right]_{t=1,\ L;\ j=1,\ L}$$

$$\Rightarrow tr\left(W_k.W_{k'}\right) = \sum_{u=1}^{L}\left(\sum_{v=1}^{L} W_{uv}^{k} W_{uv}^{k'}\right)$$

or

$$W_{uv}^{k} = \sum_{m=1}^{r_k}\left(X_{u\ j_m}^{k'}\right)\left(X_{v\ j_m}^{k'}\right) \text{ et } W_{uv}^{k'} = \sum_{t=1}^{r_{k'}}\left(X_{u\ j_t}^{k''}\right)\left(X_{v\ j_t}^{k\prime}\right) \Rightarrow$$

$$tr\left(W_k.W_{k'}\right) = \sum_{u=1}^{L}\left(\sum_{v=1}^{L}\left(\sum_{m=1}^{r_k}\left(X_{u\ j_m}^{k'}\right)\left(X_{v\ j_m}^{k'}\right) \times \sum_{t=1}^{r_{k'}}\left(X_{u\ j_t}^{k''}\right)\left(X_{v\ j_t}^{k\prime}\right)\right)\right)$$

$$\Rightarrow tr\left(W_k.W_{k'}\right) = \sum_{t=1}^{r_{k'}}\sum_{m=1}^{r_k}\underbrace{\sum_{u=1}^{L}\sum_{v=1}^{L}\left(\left(X_{u\ j_m}^{k'}\right)\left(X_{v\ j_m}^{k'}\right) \times \left(X_{u\ j_m}^{k''}\right)\left(X_{v\ j_m}^{k\prime}\right)\right)}_{*}$$

$$(*) = \sum_{u=1}^{L}\left(\left(X_{u\ j_m}^{k'}\right)\left(X_{u\ j_m}^{k'}\right) \times \sum_{v=1}^{L}\left(X_{v\ j_m}^{k''}\right)\left(X_{v\ j_m}^{k\prime}\right)\right) = \left(\prec X_{..jm}^{k}, X_{.j_m}^{k'} \succ\right)^2 \Rightarrow$$

$$tr\left(W_k.W_{k'}\right) = \sum_{t=1}^{r_{k'}}\sum_{m=1}^{r_k}\left(\prec X_{..jm}^{k}, X_{.j_m}^{k'} \succ\right)^2 \Rightarrow Si\ tr\left(W_k.W_{k'}\right) = 0 \Rightarrow$$

$$\prec X_{jm}^{k}, X_{j_m}^{k'} \succ = 0 \Rightarrow$$

les variables décrivant l'individu ω_k sont non corrélées avec les variables décrivant l'individu $\omega_{k'}$.

Interprétation géométrique des coefficients Rv On a

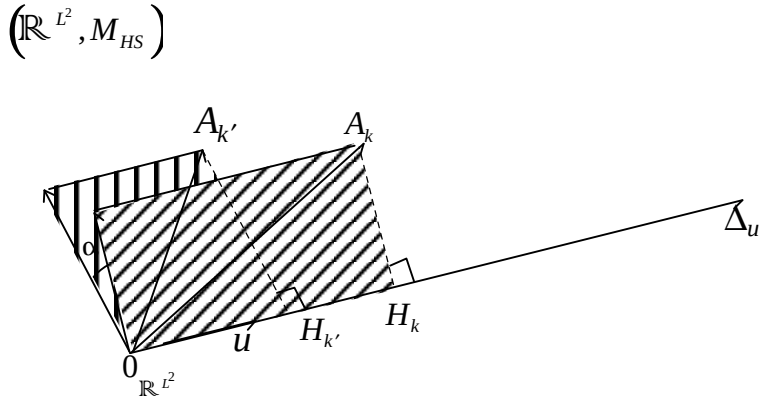


fig 9

$$\overrightarrow{OA_k} = \widehat{W_k} \ ; \ \overrightarrow{OA_{k'}} = \widehat{W_{k'}} \ \text{ et } \quad \left(\overrightarrow{OA_k}; \overrightarrow{OA_{k'}}\right) = \alpha$$

H_k et $H_{k'}$ sont respectivement les projections M_{HS} orthogonales de $\overrightarrow{OA_k}$ et $\overrightarrow{OA_{k'}}$ sur l'axe Δ_u. On a

$$\prec \widehat{W_k}, \widehat{W_{k'}} \succ = \left\|\widehat{W_k}\right\| \left\|\widehat{W_{k'}}\right\| \cos\alpha$$

$$\Rightarrow \cos\alpha = \frac{1}{\left\|\widehat{W_k}\right\| \left\|\widehat{W_{k'}}\right\|} \prec \widehat{W_k}, \widehat{W_{k'}} \succ = Rv\left(k, k'\right)$$

,

ainsi les coefficients Rv représentent les cosinus des angles entre les éléments représentatifs de l'image euclidienne.

3.2 Le nuage moyen ou compromis : l'intrastructure

L'analyse de l'interstructure a mis en évidence, sans les expliquer, les ressemblances (ou dissemblances) entre les différents tableaux étudiés.

Il est utile et intéressant de construire un nuage moyen qui soit un compromis entre les différents nuages associés aux tableaux$\{X_k;\ k = 1,\ n\}$ par leurs éléments représentatifs $\left\{\widehat{W_k}; k = 1,\ n\right\}$. Ce compromis doit être de même nature que les $\widehat{W_k}; k = 1, n$.

3.2.1 Construction d'un compromis

Le compromis peut être défini de différentes façons en fonction de la nature des données et des connaissances dont on dispose. Le bon sens veut que l'on définisse comme la moyenne pondérée des éléments représentatifs $\widehat{W_k}$ ou $\frac{\widehat{W_k}}{\left\|\widehat{W_k}\right\|_{HS}}$. Soit c le compromis en question. Alors

$$c = \sum_{t=1}^{n} \alpha_t \widehat{W_k} \text{ ou } \sum_{t=1}^{n} \beta_t \frac{\widehat{W_k}}{\left\|\widehat{W_k}\right\|_{HS}},$$

Le problème est donc la détermination des coefficients$\{\alpha_t; t\}$ ou $\{\beta_t; t\}$.

Sous les conditions : $(c1)$ et $(c2)$ où

$(c1)$: Le compromis c est l'objet représentatif le plus corrélé au sens de Hilbert-Schmidt avec les objets W_t.

$(c2)$: c est de même nature que les $\widehat{W_k}; k = 1, n$.

$(c2)$ se traduit par

$$\|c\|_{HS} = \sum_{t=1}^{n} |\alpha_t| \left\|\widehat{W_k}\right\|_{HS}$$

Si les éléments représentatifs $\widehat{W_k}$ sont normés.

$$\|c\|_{HS} = \sum_{t=1}^{n} |\alpha_t|$$

3.2.2 L'expression du compromis

Soit u^1 le vecteur propre de la matrice S associé à la plus grande valeur propre λ_1. (u^1 est choisi de telle sorte que toutes ses composantes soient positives) ce qui est possible grâce au théorème de Frobenus *

$$u^1 = \begin{bmatrix} \gamma_1^1 \\ \vdots \\ \gamma_1^n \end{bmatrix}$$

Proposition 4

Les coefficients α_t et β_t sont donnés par

$$\alpha_t = \frac{1}{\sqrt{\lambda_1}} \left[\sum_{r=1}^{L} \lambda_r \sqrt{s_{rr}} \right] \lambda_t \gamma_1^t$$

$$\beta_t = \frac{1}{\sqrt{\lambda_1}} \lambda_t \gamma_1^t$$

Une démonstration est proposée par Larvit[56]
L'expression du compromis est

$$\widehat{W_1} = \sum_{t=1}^{L} \left(\left[\frac{1}{\sqrt{\lambda_1}} \left[\sum_{r=1}^{L} \lambda_r \sqrt{s_{rr}} \right] \lambda_t \gamma_1^t \right] W_t \right),$$

dans le cas où les éléments représentatifs sont$\{W_t; t = 1, L\}$

$$\widehat{W_2} = \sum_{t=1}^{L} \left(\left[\frac{1}{\sqrt{\lambda_1}} \lambda_t \gamma_1^t \right] \frac{\widehat{W_{t_t}}}{\left\| \widehat{W_t} \right\|_{HS}} \right),$$

dans le cas normé.

L'analyse du compromis revient ensuite à effectuer l'analyse en composantes principales ou l'analyse générale de $\widehat{W_1}$ ou de $\widehat{W_2}$. Cela permet de dégager la structure du nuage des observations communes aux tableaux.

Chapitre III
Classification des éléments constitutifs d'une structure de tableaux de mesure

Summary

This chapter deals with the clustering of the elements of a structure of juxtaposition of data measuring tables. One of the main issues in such problems is the selection of a one-dimensional quantity to represent the information included in the repeated observations of each variable. We propose the use of 3 different indexes to measure the distance between elements of a structure and uses the last one based on the Hilbert-Schmidt inner product for clustering purposes through an algorithmic procedure.The algorithm proposed is applied for clustering the customers of an electric company where each customer is described by a curve of load.

Keywords: Similarity, constitutive element, structure of multiple table, kullback-Leibler distance, factorial analysis, clustering, criteria, inner product of Hilbert-Schmidt

1 Introduction

L'analyse des éléments constitutifs d'une juxtaposition de tableaux multiples reste un large domaine d'investigation. L'objectif principal de la classification automatique et de l'analyse des données est de définir des indices de similarité entre individus ou entre variables qui permettent d'associer à un couple d'objets ou de variables une valeur numérique. Ces indices de mesures sont choisis en fonction des données et de la structure de classification désirée.

Dans le cas d'une structure de juxtaposition de tableaux de mesure ($TofM$), nous définissons 3 indices de similarité. Le premier indice est basé sur la distance de Kullback-Leibler, le second indice sur les techniques de l'analyse factorielle et le troisième sur le produit scalaire de Hilbert-Schmidt. Nous utilisons le troisième indice pour construire un algorithme du type k-means[65].

L'algorithme est utilisé pour classifier les clients domestiques d'une entreprise de distribution d'électricité où chaque client est décrit par des courbes de charges ou d'appels de puissance.

Cette structure de données particulière($TofM$) est obtenue dés que l'on est conduit à effectuer plusieurs observations pour chacune des variables qui décrit chaque individu.

Des exemples de ce type de structure sont nombreux en pratique, on peut citer

En médecine: Dans cas d'une maladie incurable, on observe les patients sur une longue période par des bilans réguliers des paramètres qui donnent une idée sur l'évolution de la maladie.

En écologie végétale : Dans le cas où l'on s'intéresse à un ensemble fini d'espèces végétales sur un échantillon de terre. Chaque espèce est caractérisée par un certain nombre

de paramètres écologies pour chaque relevé.

En industrie: Si les objets sont identifiés par des courbes représentatives des pics de consommation sur une période donnée à des instants réguliers.

Dans le paragraphe qui suit (section 2), nous définissons la structure de juxtaposition de tables de données de mesure et introduisons les trois indices de de dissimilarité entre éléments de la structure. Dans la section 3 et la section 4 nous construisons un algorithme, écrit en Matlab 5.3, de type k-means utilisant l'indice de distance basé sur le produit scalaire de Hilbert Schmidt. Cet algorithme, section 5 de cette partie, a permis de classifier les clients d'une entreprise de distribution d'électricité où chaque client est décrit par ses courbes de charges construites à partir d'observations régulières.

Une conclusion et certaines remarques constituent la section 6 de ce chapitre.

2 Structure de juxtaposition de tableaux de données multiples

Soit $(\Omega,\ \Delta)$ une base de connaissance où Ω est un ensemble fini d'individus ou d'objets et Δ un ensemble de descripteurs de la forme

$$\Delta = [X_1, \ldots,\ X_n]\,,$$

X_i est une matrice d'ordre $L_i \times d$, qui contient soit les observations du système des d variables $\mathrm{V}=\{V^1, \ldots, V^d\}$. Ces observations sont soit des réalisations du vecteur V considéré comme aléatoire soit des observations répétées de l'individu ω_i pour chacune des d variables qui le decrit, $L_i$est le nombre de ces observations.

$$X_i = \left[X_i^j; j = 1,\ d\right],$$

X_i^j est un vecteur avec L_i composantes contenant les observations de l'individu ω_i pour la variable V^j.

$$\left(X_i^j\right)' = \left[x_{i1}^j, \ldots,\ x_{iL_i}^j\right],$$

x_{il}^j est la lième observation de l'individu ω_i pour la variable V^J.

Si pour tout $j = 1, \ldots, d$ et pour tout $i = 1, \ldots,\ n\ ;\ \ X_i^j \in \mathbb{R}^{L_i}$ alors Δ est une structure de juxtaposition de tableaux de mesure $(TofM)$. Dans les autres cas, Δ peut être considérée comme une juxtaposition de tableaux de données catégorielles (chapitre IV) $(TofcD)$. C'est une structure de données particulière fréquente en pratique et qui se présente sous la forme

$$\omega_1 \hookrightarrow [X_1]$$

$$\omega_2 \rightarrow [X_2]$$

$$\omega_n \hookrightarrow [X_n]$$

Pour ce type de structure, habituellement nous résumons chaque colonne de chaque tableau par un nombre. La structure est transformée en un tableau classique définissant la description d'un nombre fini d'individus par un nombre fini de variables.

Cependant, cette synthèse nécessite plusieurs hypothèses quant au choix de la valeur qui résume les différentes observations de l'individu pour la variable. Cette valeur peut être obtenue par lissage typologique en utilisant le critère des moindres carrés et la norme L^2(Anderson[2](1974); Kodell and Chen[52](2001) ; . . .) ou le critère des moindres carrés et la norme L^1(Arabie and Hubert,[4](1992); Van Cutsen[99](1994);. . .).

De plus, ces techniques opèrent de manière ponctuelle, les ajustements obtenus ne permettent pas de reconstituer même partiellement les données de départ car les applications qui associent à chaque colonne une valeur centrale (moyenne, mode, médiane, quantiles, . . .) ne sont pas injectives.

Cette réduction se présente comme suit

$$(X)' = \left[(X_1)', \ldots, \ (X_n)'\right] \rightarrowtail (Y)' = \left[(Y_1)', \ldots, \ (Y_n)'\right]$$

X est de dimension (M, d) avec $M = \sum_{i=1}^{n} L_i$, tandis que Y est de dimension $(n, \ d)$.

Dans le cas où chaque tableau regroupe les réalisations indépendantes du vecteur aléatoire ou du système des d variables aléatoires $\left\{V^1, \ldots, V^d\right\}$, on a besoin de rechercher pour tout $i = 1,\ldots, n$ les estimateurs numériques de ce système qui sont les espérances mathématiques $\left\{\mu_i^j; \ j = 1, \ d\right\}$ et les composantes des matrices de variances Σ_i.

$$\Sigma_i = \begin{pmatrix} \Sigma_{11}^i & & & \\ & \Sigma_{22}^i & & \\ & & \Sigma_{kk}^i .. \Sigma_{jk}^i & \\ & & & \Sigma_{nn}^i \end{pmatrix},$$

Où $\Sigma_{ll}^i = var\left(V^l\right)$ et $\Sigma_{lk}^i = cov\left(V^j, V^k\right)$ pour $j \neq k$.

Proposition 5

Nous obtenons les estimateurs suivants

$$\forall j = 1, d; \mu_i^j = \frac{1}{L_i} \sum_{l=1}^{L_i} x_{il}^j \ ; \left(\sigma_i^j\right)^2 = \Sigma_{jj}^i = \frac{1}{L_i - 1} \sum_{l=1}^{L_i} \left(x_{il}^j - \mu_i^j\right)^2,$$

$$j \neq k; \ \ cov\left(V^j, V^k\right) = \Sigma_{lk}^i \ = \frac{1}{L_i - 1} \sum_{l=1}^{L_i} \left(x_{il}^j - \mu_i^j\right)\left(x_{il}^k - \mu_i^k\right),$$

ces estimateurs sont sans biais, consistants et convergents et ne dépendent pas du nombre d'observations.

Si les d variables sont non corrélées même si la taille des échantillons ou le nombre d'observations est relativement petit [10 ou 20], on peut construire des intervalles de confiance corresponds au seuil fixé β.

- Cas où les σ_i^j sont connues.

L'intervalle de confiance au seuil fixé β est donné par

$$I_i^j(\beta) = \left]\mu_i^j - \varepsilon_i^j(\beta) \;\; ; \;\; \mu_i^j + \varepsilon_i^j(\beta)\right[,$$

où

$$\varepsilon_i^j(\beta) = \sqrt{\frac{\sigma_i^j}{L_i - 1}} \arg\Phi^*\left(\frac{1+\beta}{2}\right),$$

$\Phi^*(y)$ est la valeur telle que la fdr de la distribution normale est y.

- Cas où $\left\{\mu_i^j \; ; \; j \succeq 1\right\}$sont connues et $\sigma_i^j = 0 \; ; \forall j \succeq 1$ et les tableaux X_i sont centrés.

De la même manière, les intervalles de confiance approximés de la variance au seuil fixé β sont donnés par

$$I_i^j(\beta) = \left]\sigma_i^j - t_\beta.D_i^j \;\; ; \;\; \sigma_i^j + t_\beta.D_i^j\right[$$

où

$$D_i^j = \sqrt{\frac{0.8.L_i + 1/2}{L_i(L_i - 1)}}\sigma_i^j; t_\beta = \arg\Phi^*\left(\frac{1+\beta}{2}\right)$$

Le tableau X_i peut être transformé en un pavé de $\mathbb{R}^d$.

$$X_i \to R_i = \prod_{J=1}^{d} I_i^j(\beta)$$

où $I_i^j(\beta)$ est un intervalle de $\mathbb{R}$.

C'est le cas de données avec erreurs de mesure, R_i est un pavé de dimension d centré à l'origine de $\mathbb{R}^d$ [chapitre 5]

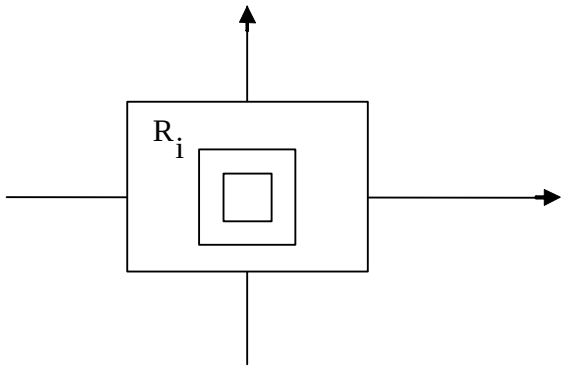


fig 10

P.Cazes, A. Chouakria, E.Diday, Y Scheketman (1997),[19] ont adapté une analyse en composantes principales pour étudier les éléments constitutifs d'une structure de ce type.

3 Indices de Distances

3.1 Indice basé sur la distance de Kullback-Leibler

3.1.1 Cas discret

Soient F_1, F_2 2 densités multidimensionnelles empiriques dont la loi de probabilité est donnée par le tableau

$F_{(.)} \rightarrow$

$$\left[\begin{array}{cccc} \left(m_1^1, m_1^2, ..., m_1^d\right) & \left(m_1^1, m_2^2, m_1^3 ..., m_1^d\right) ... & \left(m_{i_1}^1, m_{i_2}^2, ..., m_{i_d}^d\right) .. & \left(m_{r_1}^1, m_{r_2}^2, ..., m_{r_d}^d\right) \\ P_{11..1}^{(.)} & P_{12..1}^{(.)} & P_{i_1 i_2 .. i_d}^{(.)} & P_{r_1 r_2 .. r_d}^{(.)} \end{array}\right]$$

avec $\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}...\sum_{i_d=1}^{r_d}\left(P_{i_1...i_d}^{(.)}\right) = 1$; m_l^j est la $l^{ième}$ modalité ou classe de la variable V^j, r_j est le nombre de modalités ou classes.

Ce tableau peut être obtenu à partir d'une structure de juxtaposition de données catégorielles ($TofcD$) (voir Burt, [12].

Définition 2

La distance de Kullback-Leibler entre les 2 distributions est donnée par

$$d_1(F_1,F_2)=\sum_{i_1=1}^{r_1}...\sum_{i_d=1}^{r_d}\left[\left(P^{(1)}_{i_1...i_d}-P^{(2)}_{i_1...i_d}\right)\log_2\left(\frac{P^{(1)}_{i_1....i_d}}{P^{(2)}_{i_1...i_d}}\right)\right]$$

$$=K_u(F_1,F_2)+K_u(F_2,F_1)-[K_u(F_1,F_2)+K_u(F_2,F_1)]$$

avec

$$K_u(F_1,F_2)=\sum_{i_1=1}^{r_1}....\sum_{i_d=1}^{r_d}\left[\left(P^{(1)}_{i_1i_2..i_d}\right)\log_2\left(P^{(2)}_{i_1i_2,..i_d}\right)\right]$$

3.1.2 Cas continu

Soit f_1, f_2 2 densités de probabilités.

Définition 3

La distance de Kullback-Leibler entre les 2 densités est donnée par

$$d_1(f_1,f_2)=\int_{\mathbb{R}^d}(f_1-f_2)\log_2\left(\frac{f_1}{f_2}\right)d\mu$$

$$=K_u(f_1,f_2)+K_u(f_2,f_1)-[K_u(f_1,f_2)+K_u(f_2,f_1)],$$

avec

$$K_u(f_1,f_2)=\int_{\mathbb{R}^d}(f_1)\log_2(f_2)\,d\mu.$$

Après Shannon[90], Kullback[55] explique que la quantité évalue l'information moyenne perdue si on utilise la distribution f_1 alors que la vraie distribution est f_2.

Si f_1 et f_2. sont 2 densités de probabilité de la loi normale multidimensionnelle de dimension d de paramétres $\{(\mu_i,\Sigma_i);i=1,2\}$.Pour $i=1,2$

$$f_i\longrightarrow N(\mu_i;\Sigma_i)'$$
$$\Rightarrow f_i(x)=(2\Pi)^{-\frac{d}{2}}|\Sigma_i|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\|x-\mu_i\|^2_{\Sigma_i^{-1}}\right],$$

Proposition 6

On a

$$d_1(f_1, f_2) = \frac{1}{2\log(2)}\left[\left|\Sigma_1\Sigma_2^{-1}\right| + \left|\Sigma_1^{-1}\Sigma_2\right| - \|\mu_1 - \mu_2\|^2_{\Sigma_1^{-1}} - \|\mu_1 - \mu_2\|^2_{\Sigma_2^{-1}} - 2\right].$$

Hypothèse

$$Si\ d_1(f_1, f_2) = 0 \Rightarrow f_1 = f_2 \Rightarrow \omega_1 = \omega_2.$$

Dans le cas où $\mu_1 = \mu_2 \Rightarrow$

$$d_1(f_1, f_2) = \frac{1}{2\log(2)}\left[\left|\Sigma_1\Sigma_2^{-1}\right| + \left|\Sigma_1^{-1}\Sigma_2\right| - 2\right],$$

cet indice de distance peut être utilisé en associant un indice d'agrégation adéquat pour construire une hiérarchie indicée sur les éléments constitutifs de la structure de tableaux de données de mesure ou de données catégorielles.

Cependant, le choix de la distribution qui représentera une classe d'individus décrits par des distributions reste un problème posé, en conséquence un critère du type intra-classe basé sur cet indice de distance ne peut pas être défini.

3.2 Indice de distance basé sur des techniques factorielles

Cet indice est largement étudié dans Naab, (2001)[68] dans le cadre d'une thèse de Magister effectué sous ma direction. Soit

$$(X)' = \left[(X_1)', \ldots, (X_n)'\right]$$

une structure de juxtaposition de tableaux multiples. X_i est le tableau qui contient les observations de l'individu $\omega_i \in \Omega$. On peut utiliser les techniques factorielles pour réduire chacun des tableaux X_i. Chaque analyse du tableau X_i nous donne un système d'axes

$$\{E_i = \{\Delta_{u_r};\ r \succeq 1\}; i = 1,\ n\}$$

Le problème consistera à rechercher un système commun. Bouroche[10] a proposé un certain nombre de critères qui permettent de choisir ce système d'axes commun à toutes les analyses, mais les calculs sont compliqués et parfois non justifiés. Nous retenons 3 critères pour la recherche du système d'axes commun à tous les systèmes obtenus par les différentes analyses factorielles. Ces nouveaux systèmes permettent de comparer les différents éléments de la structure mais beaucoup d'informations sera perdue par les procédures proposées. Ces approches ne sont donc pas très recommandées.

Soit X un tableau de données, on suppose que les objets fictifs sont classifier en n classes

$$\{C_1, \ldots, \ C_n\}$$

d'effectifs respectifs

$$L_1, \ldots, L_n.$$

Chaque classe représente un individu.

L'analyse en composantes principales du tableau X conduit à la diagonalisation de la matrice

$$W = \sum_{i=1}^{n} (X_i)' \, D_i \, (X_i) \, ,$$

D_i est une matrice diagonale contenant respectivement les poids associés aux observations de chacun des objets. S'il n'y a pas d'observations privillègiées alors

$$D_i = \frac{1}{L_i} \Im_{L_i},$$

$\Im_{L_i}$ est la matrice identitée de dimension L_i

Soit u_r le $rième$ axe factoriel. On a

$$\|u_r\|_{\Im_d} = 1 \text{ et } u_r \underset{\Im_d}{\perp} u_{r'} \text{si } \ r \neq r'$$

Le représentant de l'individu ω_i (C_i) sur cet axe factoriel et donné par

$$(z_i)_r = \widetilde{X_i}.u_r \in \mathbb{R}^{L_i},$$

$\widetilde{X_i}$ tableau centré

Le représentant de la classe C_i est

$$T_i = \left[(z_i)_1 \, , \ldots . , (z_i)_{r \max} \right],$$

T_i est une matrice de dimension $L_i \times r \max$, $rmax$ est le nombre maximum retenu pour les valeurs propres de la matrice W.

Proposition 7

La quantité

$$d_2\left(C_1, C_2\right) = \left(\sum_{r=1}^{r\max} \left(\left\| (z_1)_r - (z_2)_r \right\|_{\Im_{r\max}} \right)^2 \right)^{\frac{1}{2}},$$

est un indice de distance entre éléments représentatifs de la structure.
Hypothèse

$$d_2\left(C_1, C_2\right) = 0 \Rightarrow C_1 = C_2 \Rightarrow \omega_1 = \omega_2.$$

La quantité d'information perdue par cette procédure est évaluée à

$$\epsilon = 100 \left(1 - 100 \left(\frac{\sum_{\alpha=1}^{r\max} (\lambda_\alpha)}{d} \right) \right).$$

3.3 Indice basé sur le produit salaire de Hilbert-Schmidt

Soit W_i la matrice symétrique des produits scalaires du tableau X_i le $i^{ème}$ élément de la structure, le teme général de cette matrice est

$$(W_i)_{lj} = \sum_{m=1}^{L_i} p^j x^i_{lm} \left(x^i_{mj} \right),$$

p^jest le poids associé à la variable V^j avec $\sum_{j=1}^{d} p^j = 1$

La matrice symétrique W_i d'ordre L_i représente l'individu ω_i car elle regroupe les produits scalaires entre les observations des variables.

Si $L_i = L$ et $p^j = \frac{1}{d}$ # *Cette hypothèse se traduit en pratique par le fait que les objets ou les individus sont observés le même nombre de fois et qu'il n' y a pas d'observations privilégiée* #

W_i peut être considérée comme un point de l'espace $\mathbb{R}^{L^2}$en empilant ses L colonnes

$$W_i \mapsto \left(\widetilde{W_i} \right)' = \left[\left(W_i^1 \right), \ldots, \left(W_i^L \right) \right],$$

avec $\rightarrowtail \left(W_i^l \right)' = \left(W_{i1}^l, \ldots, W_{iL}^l \right)$

$$\Rightarrow W_i^l \in \mathbb{R}^L \ \forall l = 1, \ldots, L \text{ et } \widetilde{W_i} \in \mathbb{R}^{L^2}$$

On défini ainsi un nuage de n vecteurs de $\mathbb{R}^{L^2}$ $\left\{ \widetilde{W_i}; i = 1, \ldots, n \right\}$ représentants les n tables $\left\{ X_i; i = 1, \ldots, n \right\}$.

3.3.1 Le produit scalaire de Hilbert-Schmidt

Proposition 8

L'application

$$\delta : M_L \times M_L \to \mathbb{R}_+$$

$$(W_1, W_2) \to \delta(W_1, W_2) = \prec W_1, W_2 \succ_{HS} = tr((DW_1.DW_2))$$

est un produit scalaire, M_L est l'ensemble des matrices symétriques d'ordre L, la matrice diagonale D d'ordre L contient les poids affectés aux observations.

Ce produit scalaire est le produit scalaire de Hilbert-Schmidt.

Si D est la matrice identitée d'ordre L alors

$$\prec W_1, W_2 \succ_{HS} = tr(W_1.W_2) = \prec \widetilde{W_1}, \widetilde{W_2} \succ_{id},$$

$\widetilde{W_1}, \widetilde{W_2} \in \mathbb{R}^{L^2}$, $\prec$, $\succ_{id}$ produit scalaire classique.

3.3.2 Norme et distance induite par le produit scalaire de Hilbert-Schmidt

La norme d'un élément représentatif W_1 est donnée par

$$\prec W_1, W_1 \succ_{HS} = tr(W_1.W_1) = \prec \widetilde{W_1}, \widetilde{W_1} \succ_{id} = \left\| \widetilde{W_1} \right\|^2,$$

La distance entre deux objets ω_1 et ω_2 est donnée par

$$d_3(\omega_1, \omega_2) = \|W_1 - W_2\|_{HS} = \left\| \widetilde{W_1} - \widetilde{W_2} \right\|_{id}.$$

$$d_3(\omega_1, \omega_2) = \sqrt{\left\| \widetilde{W1} \right\|^2 + \left\| \widetilde{W_2} \right\|^2 - 2 \left\| \widetilde{W_1} \right\| \left\| \widetilde{W_2} \right\| \cos\left(\widetilde{W_1}, \widetilde{W_2} \right)}$$

En pratique, les éléments représentatifs sont souvent normés pour éviter que les éléments qui ont une grande norme n'influent sur l'analyse. Ainsi,

Si $\left\| \widetilde{W_1} \right\| = \left\| \widetilde{W_2} \right\| = 1$, alors

$$d_3(\omega_1, \omega_2) = \sqrt{2 - R_v(1, 2)}$$

où

$$R_v(1,2) = \cos\left(\widetilde{W_1}, \widetilde{W_2}\right) =\prec \frac{\widetilde{W_1}}{\left\|\widetilde{W_1}\right\|}, \frac{\widetilde{W_2}}{\left\|\widetilde{W_2}\right\|} \succ_{id}= \frac{tr(W_1.W_2)}{\|W_1\|_{HS}\|W_2\|_{HS}}$$

R_v est donc le coefficient d'association entre les individus décrits par leur élément représentatif respectif . C'est un coefficient de corrélation vectoriel entre éléments représentatifs.

4 Critère et problème d'optimisation

Nous désirons regrouper les n individus en k classes où chaque individu est décrit par son élément représentatif. L'homogénéité des classes est mesurée par

$$H_k = \sum_{\{i\}\in P_k} d_{HS}(W_i, l_k),$$

où l_k est l'élément représentatif du noyau de la classe P_k et d_{HS} est la distance induite par le produit scalaire de *Hilbert Schmidt*. H_k mesure la dispersion des individus à l'intérieur de la classe. H_k est aussi l'inertie totale des objets de la classe autour de son noyau.

Soit P_K l'ensemble des partitions à k classes

$$P \in P_K \Rightarrow P = \{P_1, \ldots, P_K\}$$

est une partition à K classes, L_k l'ensemble des K représentations

$$(L \in L_K) \Rightarrow L = \{l_1, \ldots, l_k\}; l_t \in M_L.$$

On défini l'application

$$C_r : P_K \times L_K \to \mathbb{R}_+$$

$$(P, L) \to C_r(P, L) = \sum_{k=1}^{K} \sum_{\{i\}\in P_k} d_{HS}(W_i, l_k)$$

où

$$d_{HS}(W_i, l_k) = \left\|\widetilde{W_i} - \widetilde{L_k}\right\| = tr(W_i \times l_k)$$

$$\widetilde{W_i}; \widetilde{L_k} \in \mathbb{R}^{L^2}, W_i;\ l_k \in M_l$$

Le critère s'écrit

$$C_r(P,L) = \sum_{k=1}^{K} \sum_{\{i\}\in P_k} \left(\sqrt{\left\|\widetilde{W_i}\right\|^2 + \left\|\widetilde{L_k}\right\|^2 - 2 \prec \widetilde{W_i}, \widetilde{L_k} \succ} \right)^{1/2}$$

Ce critère exprime l'adéquation entre les éléments représentatifs des individus dans les classes avec les éléments représentatifs des noyaux des classes. Cr s'écrit

$$C_r(P,L) = \sum_{k=1}^{K} \sum_{\{i\}\in P_k} \left(2\left(1 - R_v\left(\widetilde{W_i}, \widetilde{L_k}\right)\right)\right)^{\frac{1}{2}}$$

Le coefficient R_v avec ses propriétés permet d'expliquer l'adéquation entre la partition et les représentants de ses classes. On note que plus R_v est proche de 1 plus petit est le critère et plus les individus sont en adéquation avec les noyaux des classes où ils sont affectés.

On recherche (P^*, L^*) qui réalise

$$\underset{\substack{P\in P_K \\ L\in L_K}}{Min} C_r(P,L)$$

Les algorithmes utilisés pour résoudre ce type de problème sont du type k-means. L'algorithme est basé sur la définition de la fonction de représentation g et de la fonction d'affectation f.

4.1 La fonction de représentation

$$g : P_k \to L_k$$

$$P = \{P_1, \ldots, P_k\} \to g(P) = \{L_1 \ldots, L_k\}$$

g doit vérifier

$$\underset{L\in L_k}{Min} Cr(P,L) = Cr(P, g(P))$$

Proposition 9
Ce minimum est obtenu pour

$$g(P) = \widehat{L} = \left\{\widehat{L}_1, \ldots, \widehat{L}_k\right\} = \widehat{G};\ \widehat{L}_l = \widehat{G}_l \in M_L$$

est le centre de gravité de la classe P_l. Ce centre de gravité ou noyau de la classe est donné par

$$\widehat{G}_l = \left[\left(\widehat{G}_1\right)', \ldots, \left(\widehat{G}_L\right)'\right]$$

où pour $l = 1,\ k$

$$\left(\widehat{G}_l\right)' = \left[G^l_{i1}, \ldots,\ G^l_{iL}\right] \Rightarrow \widehat{G}_l \in \mathbb{R}^L,$$

avec

$$L^l_{it} = G^l_{it} = \frac{1}{N_l} \sum_{(i)\in P_l} (W_i)_{tl}\,; N_l = |P_l|$$

4.2 La fonction d'affectation

$$f : L_k \to P_k$$

$$L = \{L_1, \ldots,\ L_k\} \to f(L) = \{P_1, \ldots,\ P_k\}$$

f doit vérifier

$$\underset{P\in P_k}{MinCr}\,(P, L) = Cr\,(f(L),\ L)$$

Proposition 10
Ce minimum est obtenu pour

$$P_l = \{\omega_i \in \Omega / d_{HS}(\omega_i,\ L_l) \preceq d_{HS}(\omega_i,\ L_t)\,; \forall t \neq l \ \text{ et } \ l \prec t \text{ en cas d'égalité}\}$$

5 L'algorithme

On choisi au hasard ou par d'autres artifices $L^{(0)}$ et on utilise alternativement les fonctions f et g pour faire décroître le critère. L'algorithme se déroule comme suit

$$L^{(0)} \xrightarrow{f} P^{(1)} \xrightarrow{g} L^{(1)} \xrightarrow{f} P^{(2)} .....L^{(*)} \xrightarrow{f} P^{(*)}$$

L'algorithme s'arrête dés que la partition obtenue ne change plus. Nous construisons ainsi 2 suites U_n et V_n telles que

$$\begin{cases} U_n = Cr\left(P^{(n)}, L^{(n)}\right) \\ V_n = \left(P^{(n)}, L^{(n)}\right) \end{cases}$$

Proposition 11[31]
La suite U_n converge en décroissant et la suite V_n est stationnaire à partir d'un certain rang donc convergente

6 Structure générale de l'algorithme : Organigramme

Déclaration des données
$T_i = [t^i_{hl}]$ tableau de dimension HxJ
$i = 1, \ldots, n$; $\varepsilon =$ seuil d'arret
$n =$ nombre d'individus, $J =$ nombre de jours
$H =$ nombre d'heures, $K =$ nombre de classes
↓
Détermination des éléments représentatifs
$i = 1, \ldots, n$; $W_i = T_i \times (T_i)'$
↓
Empilage des colonnes des éléments représentatifs
$\widetilde{W_i} = \frac{\widehat{W_i}}{\left\|\widehat{W_i}\right\|_{HS}}$
↓

Contruction de la partition initiale

↓
Partition P$^{(0)}$
Construction de P$^{(0)}$ à partir d'un tirage pseudo-aléatoire
de K nombres choisis parmis les n numéros
↓
Représentation $L^{(0)}$
Calcul des K noyaux
↓

Calcul des distances entre clients et noyaux
et **affectation** des clients aux nouvelles classes

↓

Calcul du critére
$\Sigma^{(0)} = W\left(P^{(0)}, L^{(0)}\right)$

↓

Critére
$Crit\left(P^{(m)}, L^{(m)}\right)$

↓

Partition P$^{(m)}$
Calcul du centre de gravité de chaque classe

Non
Etape $m+1$

↓

Test de convergence
$|\Sigma^m - \Sigma^{m-1}| \leq \varepsilon$

↓

fin

7 Application

7.1 Données en entrée

Ces données ont fait l'objet d'une étude dans le cadre de plusieurs projets de fin d'études. Nous les utilisons pour montrer que l'analyse faite après une étape de réduction conduit à des résultats non interprétables par les paramétres qui carractérisent les clients et que si la variabilité de la consomation d'éléctricité est assez importante il faut éviter l'étape réduction.

L'algorithme est implémenté en MATLAB 5.3 software et est appliqué pour classifier 47 individus en 5 classes. Les 47 individus constituent un panel de clients domestiques basse tension choisis pour représenter les abonnés de la zone II. Ce choix a été fait par des techniques de sondage éprouvées. La stratification des données s'est faite sur la base du critère de consommation de l'électricité de manière à ce qu'il y ait égalité de la consommation annuelle par strates (10 strates). La répartition des individus par strates s'est faite en minimisant la variance de la consommation des échantillons. Ainsi, le tirage des échantillons s'est fait par zones (9 zones). Un pas de tirage de chaque strate et de chaque zone a été défini et on a tiré dans chaque zone de chaque strate un abonné de la liste principale (900000 abonnés).

Pour ces 47 clients, la société de distribution a installé des compteurs spéciaux qui mesurent l'appel de puissance toutes les 10 minutes pendant toute la durée de l'étude (les 6 premiers mois de l'année 2000). Devant la grande masse de données relevées, l'étude s'est restreinte aux relevés du mois d'avril qui marque la fin de l'hiver et pas encore l'été (le chauffage et la climatisation ne sont pas nécessaires). Les 3 jours éliminés du mois

sont les jours de fin de semaine car certains clients sont absents du logement tandis que pour d'autres, il y a une consommation maximale durant les 3 jours, due à leur présence continuelle.

L'entreprise a choisi d'observer les pics de consommations de chaque client durant les 28 jours ouvrables du mois d'avril.

7.2 Classification aprés réduction

Pour chaque client, on a construit sa courbe de charge ou d'appels de puissance regroupant en abscisses les 24 heures de la journée et en ordonnée ses pics de demande de consommation ou de charge. Pour cette approche, on a résumé la courbe des appels de puissance pour chaque client du panel par un vecteur regroupant les moyennes des appels de puissance pour les 28 jours pour chaque heure. Ainsi,

$$\underset{(28,720)}{\widehat{T_i}} \rightarrow C_i \rightarrow \underset{((28\times 24))}{T_i} \rightarrow Y_i = \left[\overline{x_i^1}, \overline{x_i^2}, ..., \overline{x_i^d}\right] ;\ d = 28$$

$$\underset{(47,28)}{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_{47} \end{bmatrix}$$

Table 4: Meilleure partition, en 5 classes, obtenue après 10 itérations en déroulant l'algorithme k-means sur le tableau Y

classes	1	2	3	4	5
Individus dans les classes	1, 17, 18, 22 28, 29, 35, 40 44, 46	26 42 43	16 27	33 41	2, 3, 4, 5, 6, 7, 8, 0, 10, 11, 12, 13, 14 15, 19, 20, 21, 23, 24, 25, 26, 30, 31 32, 34, 36, 37.38.39.41.45.47
effectif	**10**	**3**	**2**	**2**	**30**

Pour chaque client, on a construit sa courbe de charge ou d'appels de puissance regroupant en abscisses les 24 heures de la journée et en ordonnées ses pics de demande de consommation ou de charge. Pour le mois d'étude, les observations de ω_i sont donc regroupées dans des courbes de charges de ce type

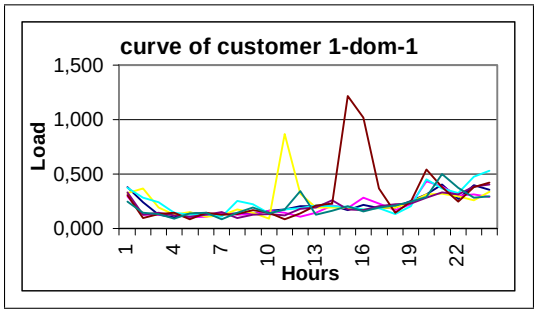


fig 11

La grande variabilité de la consommation durant les minutes et les heures de la journée (en période de pointe et en période creuse) fait que l'interprétation des classes à partir des caractéristiques de chaque client (type d'habitation, qualité des produits électriques utilisés, nombre de lampes, nombre de pièces, nombre de personnes constituant le ménage, ...) n'a pas donné de bons résultats.

Pour cet exercice, la condition de normalité n'est pas vérifiée. De plus, les courbes dépendent les unes des autres. La consommation moyenne pour chaque heure ne peut donc résumée les différentes observations.

7.3 Classification basé sur l'indice de Hilbert-Schmidt

C_i est une courbe construite à partir d'un nombre fini de points. La structure de données en entrée peut être considérée comme une juxtaposition de tableaux de mesure.

$$\underset{(28,720)}{\widehat{T}_i} \rightarrow C_i \rightarrow \underset{((28\times 24))}{T_i} = \begin{array}{c} \\ 1 \\ \vdots \\ l \\ \vdots \\ 28 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & \dots & j & \dots & 24 \end{array} \\ \begin{pmatrix} & & \dots & & & \\ \vdots & & & & & \vdots \\ & t_{il}^{j} & & & & \\ \vdots & & & & & \vdots \\ & & \dots & & & \end{pmatrix} \end{array}$$

t_{il}^{j} est la consommation de l'individu durant l'heure du jour . est un tableau de dimension Les clients sont identifiés par les nombres 1 à 47.

Table 5: partition initiale construite autour de 5 individus choisis au hasard parmi les 47 par la fonction random.

classes	1	2	3	4	5
Individus dans les classes	5, 6, 8, 9	1, 10	34	2, 3, 4, 7, 12, 13	11, 16, 17
	14, 15, 18	21.33	39	20, 22, 23, 26, 27, 28	24, 25, 31
	30, 44, 47	37, 41	43	29, 32, 35, 36, 40, 42, 47	38, 45
effectifs	**10**	**6**	**3**	**20**	**8**

Valeur du critère pour cette partition

$$Cr = 9.9455$$

Table 6: Partition optimale obtenue aprés 8 itérations

classes	1	2	3	4	5
Individus dans les classes	5, 6, 8, 9	1, 4, 10, 12, 13, 19, 21	20, 23	2, 3, 7	16, 17
	14, 15, 18	23, 25, 26, 27, 30, 33	22, 32	11, 28	24, 31
	44, 47	35, 36, 37, 40, 41, 42, 46	34	29, 43	38, 45
effectifs	**9**	**20**	**5**	**7**	**6**

Valeur du critère pour cette partition

$$Cr = 4.0498$$

7.4 Interprétation de ces résultats à l'aide des courbes représentatives des centres ou noyaux des classes

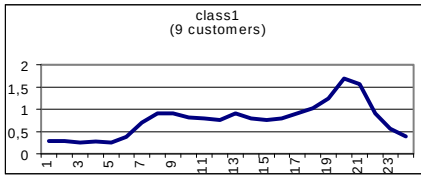


fig 12

La courbe représentant la **classe 1** regroupe 9 clients et montre un pic d'appels de puissance en période de pointe, un pic dominant en matinée et une demande uniforme en période creuse.

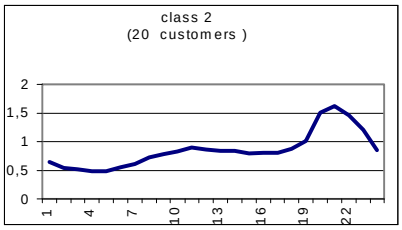


fig 13

La **classe 2** regroupe 20 clients et présente un pic d'appels de puissance identique à celui de la classe 1 mais les 2 courbes n'ont pas la même forme.

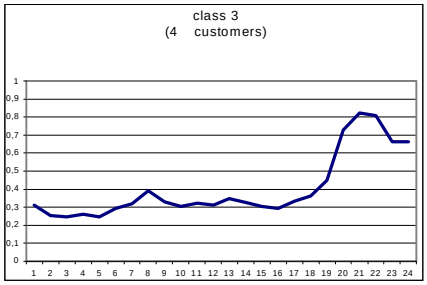


fig 14

La courbe de la **classe 3** regroupe 4 clients et montre un pic d'appel de puissance en heures de pointe équivalent à 0,8 kwh et des appels de puissance uniformes en périodes creuses.

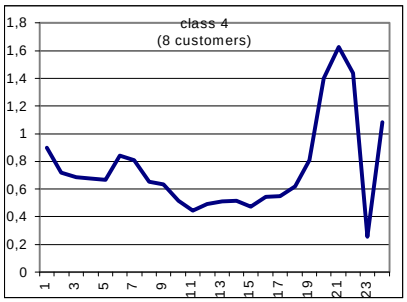


fig 15

On note que la **classe 4** regroupe 8 clients et est caractérisée par un pic d'appels de puissance dominant en soirée de 1.6 kwh et une décroissance dans les autres périodes de la journée.

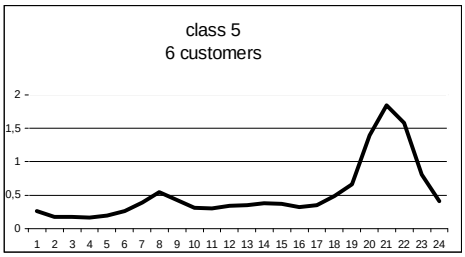


fig 16

La **Classe 5** regroupe 6 clients et est caractérisée par un pic dominant en soirée de 1.9 kw/h et une uniformité des appels de puissance dans les autres périodes de la journée avec un petit pic en matinée.

Cette application a montré que les courbes représentatives des noyaux finaux obtenues par l'algorithme présentent des formes semblables avec des niveaux différents. De plus, l'algorithme a donné des classes bien séparées. Pour ce type de données où le ratio entre les appels de puissance durant les heures de pointes et les heures creuses est supérieur à 3 , il n'est pas recommandé de réduire les données avant la classification . Enfin, nous avons pu expliquer les classes en fonction des autres caractéristiques des clients autres que la consommation d'électricité (type d'habitation, nombre de pièces, nombre de lampes utilisées, nombre de personnes ,...) ce que nous n'avons pas pu faire pour les classes obtenues après réduction.

8 Conclusion

La distance induite par le produit scalaire d'Hilbert Schmidt a permis de construire un algorithme de type k-means pour classifier les éléments constitutifs d'une structure de juxtaposition de tableaux de mesure de même dimension. Cependant, cet indice de distance ne s'étend pas au cas d'une strucuture de données catégorielles et il serait intéréssant de trouver l'équivalent comme dans le cas de la distance euclidienne et la distance du khi-deux car on ne doit pas oublier que cette distance conduit à la distance euclidienne classique ,je pense que c'est une piste intérésssante.

.

Chapitre IV
Classification des éléments constitutifs d'une structure de tableaux de données catégorielles

1 Introduction

Dans ce chapitre, on s'intéresse à la classification d'objets matriciels qui contituent les composants d'une structure des tableaux multiples et de manière particulière, à des matrices qui sont des tableaux du type disjonctifs complets donc à des matrices 0/1.Les indices de distance basées sur des trajectoires factoirielles et sur le produit scalaire de Hilbert-Schmidt proposés dans le chapitre précedent ne s'adaptent ni à ce type de structure (données qualitatives) ni au cas où les objets matriciels sont de dimensions diffèrentes *cf : application développée dans ce chapitre.*

Nous proposons un formalisme basé sur des notions simples de la statistisque descriptive et de la theorie de l'information pour construire un indice de distance entre ces objets matriciels et qui s'etend au cas de données qualitatives de manière plus générale.

Nous considérons pour ce faire, qu'une observation est une réalisation d'une variable aléatoire. Dans le paragraphe 2 de ce chapitre, nous introduisons la notion de système physique comme modèle qui régit la description de chaque objet. Le nombre des états de chaque système est fini et la probabilité que le système se trouve dans un des états est mesurée par sa fréquence. Nous utilisons l'entropie pour mesurer l'incertitude des états des systèmes Shannon[90] et enfin, nous définissons dans le paragraphe 3 une distance entre les éléments constitutifs de la structure.

Cette démarche a permis de construire une hiérarchie indicée entre les éléments constituitfs de la strucuture de juxtaposition de tableaux multiples.

Un exemple numérique, l'algorithme, une application sur données réelles et quelques perspectives de développement possibles complètent ce chapitre.

Le programme de l'algorithme écrit en visual basic sous exell est placé en annexe. Ce programme est déroulé sur l'exemple numérique.

2 Formalisme adapté

Soit $\Omega = \{\omega_1, \ldots, \omega_n\}$un ensemble fini de N objets élémentaires, et $\left\{V^1, \ldots, V^d\right\}$ d variables discrétes définies sur Ω et prennant un nombre fini de valeurs respectivement dans $D_1, \ldots, D_d$ où $D_j = \left\{m_1^j, \ldots, m_{r_j}^j\right\}$, m_l^j est la *lième* modalité de la variable V^j.

$$E_i^j = \begin{pmatrix} v_1^j \\ \vdots \\ v_{L_i}^j \end{pmatrix} \text{ où } v_l^j \in \left[m_1^j, \ldots, m_{r_j}^j\right], \forall l = 1, \ldots, L_i \; ; \forall j = 1, \ldots, d$$

E_i^j est le vecteur avec L_i composantes correspondantes aux diffèrentes observations de l' individu $\omega_i \in \Omega$. E_i^j contient par exemple, les observations répetées de l'objet ω_i pour la variable discréte V^j qui le décrit.

La structure de juxtaposition de tableaux de données catégorielles ($TofcD$) est

$$\Delta = [\Delta_1, \ldots, \ \Delta_N] \ \text{ avec } \quad \Delta_i = \left[\Delta_i^1, \ldots, \ \Delta_i^d\right],$$

Δ_i est une matrice de dimension $L \times M$, $M = \sum_{j=1}^{d} r_j$ et Δ_i^j est un tableau de données catégorielles défini comme suit

$$\Delta_i^j = \begin{bmatrix} \overbrace{m_1^j m_2^j \ldots m_{r_j}^j}^{V^j} \\ 00000100 \\ 00010000 \\ \ldots z_{lt}^j \ldots \\ \ldots \\ 00001000 \end{bmatrix}$$

$$Z_{it}^j = \begin{cases} 1 & \text{si } \omega_i \text{ a choisi la modalité } m_t^j \\ 0 & \text{ailleurs} \end{cases}$$

Pour $\omega \in \Omega$, on défini l'application

$$V_\omega : \Theta \longrightarrow V(\Omega) = \{m_1, \ldots, m_q\}$$

$$(\theta_\omega)_t \longmapsto V((\ \theta_\omega)_t) = m_j$$

m_j est la *tème* observation de l'individu $\omega \in \Omega$ pour la variable V et Θ est l'ensemble qui contient les L observations de ω. On pose

$$P[V((\theta_\omega)_t) = m_j] = P[V_\omega = m_j] = f_{V_\omega}^j \ \text{ et } \ \sum_{j=1}^{q} f_{V_\omega}^j = 1$$

Défintion 4

La paire $(\Theta, P_{V_\omega}) = S_\omega = [V \hookrightarrow P_{V_\omega} = P_\omega]$ est appelé *: système physique aléatoire simple.*

$S = [V \hookrightarrow P = F]$ décrit $\omega \in \Omega$ si et seulement si la distribution de la variable V pour les observations de l'individu ω est F, le système physique S s'écrit

$$S = \underset{j=1}{\overset{q}{\Lambda}} \left[V = m_j; \Pr\left[V = m_j\right] = p_j\right],$$

où $\wedge$ est la conjonction entre événements aléatoires.
Dans le cas multidimensionnel,
Définition 5
le système physique

$$S_\omega = (\Theta, P_{V_\omega}) = \left[V \hookrightarrow P_{V_\omega} = F_d^\omega\right] = \left[V = \left(V^1, \ldots, V^d\right) \hookrightarrow F_d^\omega\right],$$

où

$$P_{V_\omega}\left(l_1, \ldots,\ l_d\right) = P\left[V_\omega = \left(m_{l_1}^j, \ldots,\ m_{l_d}^j\right)\right] = f_d^\omega\left(l_1, \ldots,\ l_d\right)$$

$$\text{et } \sum_{l_1=1}^{r_1} \sum_{l_2=1}^{r_2} \ldots \sum_{l_d=1}^{r_d} \left[f_d^\omega\left(l_1, \ldots,\ l_d\right)\right] = 1$$

est appelé: *système physique multiple aléatoire.*
Soit F_ω^j la $jième$ distribution marginale de F_d, le systéme physique aléatoire multiple associé aux distributions marginales est

$$\widehat{S_\omega} = \underset{j \in Q_\omega}{\widetilde{\wedge}} \left[S_\omega^j\right],$$

$\widetilde{\wedge}$ conjonction entre systèmes physiques.
Les systèmes physiques simples$\{S_\omega^j;\ j = 1\}$ sont donnés par

$$\forall j \in Q_\omega;\ S_\omega^j = \left[V^j \hookrightarrow F_\omega^j\right].$$

3 Distance entre systèmes physiques aléatoires multiples

3.1 Entropie comme mesure de l'incertitude des états des systèmes

Pour mesurer le degré d'incertitude des états des systèmes physiques ou de la variable aléatoire discrète, on utilise une caractéristique spéciale couramment utilisée en théorie de l'information appelée : *entropie*

Soit S système physique aléatoire

$$S = (\Theta, P_V) = [V \hookrightarrow P_V = P] = \widetilde{\bigwedge}_{i=1}^{q} \left[(S) \hookrightarrow v^j; P \ \left[(S) \hookrightarrow v^j \right] = p_j \right].$$

Le symbole $(S) \hookrightarrow v^j$ traduit le fait que le système se trouve à l'état v^j.
Formulation de Shannon [1948] **de l'entropie**
Definition 6
L'entropie du systéme est la quantité positive

$$H(S) = -\sum_{j=1}^{q} p_j \log_2 (p_j),$$

La fonction H posséde quelques propriétés élémentaires qui justifie son utilisation comme caratéristique pour mesurer le degré d'incertitude du systéme

Proposition 12

1. Si un des états est certain ($\exists\ l \in \{1, 2, ..., q\}$ tel que :$P_l = \Pr[(S) \hookrightarrow v_l] = 1$) alors $H(S) = 0$.En effet, si $p_l = 1 \Rightarrow\ p_l \log_2 (p_l) = 0 \Rightarrow \underset{p \to 0}{Lim} (p_l \log_2 (p_l)) = 0 \Rightarrow H(S) = 0$.

2. L'entropie d'un systéme physique aléatoire avec un nombre fini d'états $(m_1, m_2, .., m_q)$ est maximale si tous les états sont équiprobables: $\forall\ j \in \{1, 2, .., q\}; p_j = \Pr[S \leftrightarrow m_j] = \frac{1}{q}$. On a : $0 \preceq H(S) \preceq \log_2 (q)$.

- La caractéristique de la fonction d'entropie exprime le fait que la distribution de probabilité avec le maximum d'entropie est la plus biaisée et la plus consistante avec l'information spécifiée par les contraintes.

- Cette distribution est aussi la plus dispersée si il n'y a pas de contraintes supplémentaires que celles spécifiées dans le problème d'optimisation[90].

3.2 Entropie d'un système physique multiple aléatoire

Dans le cas multidimensionel, Soit F_d une distribution multidimensionnelle d'ordre be probability d. F_d est donnée par la table

$$F_d \rightarrow \begin{array}{cccccc} \left(m_1^1, m_1^2, ..., m_1^d\right) & \left(m_2^1, m_1^2, ..., m_1^d\right) & ... & \left(m_{l_1}^1, m_{l_2}^2, ..., m_{l_d}^d\right) & ... & \left(m_{r_1}^1, m_{r_2}^2, ..., m_{r_d}^d\right) \\ p_{11....1} & p_{21...1} & ... & p_{l_1 l_2 ... l_d} & ... & p_{r_1 r_2 ... r_d} \end{array}$$

Le système physique associé est

$$S = \overset{r_1}{\underset{l_1=1}{\wedge}} \overset{r_2}{\underset{l_2=1}{\wedge}} ... \overset{r_d}{\underset{l_d=1}{\wedge}} \left[V = \left(m_{l_1}^1, \ldots, m_{l_d}^d\right); \Pr\left[V = \left(m_{l_1}^1, \ldots., \ m_{l_d}^d\right) = p_{(l_1, \ldots, \ l_d)} \right] \right],$$

où

r_j est le nombre de valeurs prises par la variable aléatoire V^j càd: le nombre d'etats et

$$\sum_{l_1=1}^{r_1}\sum_{l_2=1}^{r_2}...\sum_{l_d=1}^{r_d}\left[p_{(l_1,\ldots,\,l_d)}\right]=1.$$

S est appelé système physique multiple

Le systéme physique multiple aléatoire associé aux distributions marginales est donné par

$$\left(\widehat{S}\right)=\widetilde{\Lambda}_{j=1}^{d}\left[S_j\right].$$

Le systéme physique simple et discrét S_j est

$$S_j=\underset{l=1}{\overset{r_j}{\Lambda}}\left[V^j=m_l^j;\Pr\left[V^j=v_l^j\right]=p_l^j\right]=\left[\Theta^j,P_{V^j}\right],$$

où

$$p_l^j=P\left[V^j=v_{_I}\right]=\sum_{l_1=1}^{r_1}\sum_{l_2=1}^{r_2}...\sum_{l_{j-1}=1}^{r_{j-1}}\sum_{l_{j+1}=1}^{r_{j+1}}...\sum_{l_d=1}^{r_d}\left[p_{l_1,l_2,...,l_d}\right];$$

$$\text{avec}\quad\sum_{l=1}^{r_j}p_l^j=1$$

Proposition 13

Si les systémes physiques simples $((S^j)\ ;j\in Q_j)$ sont indépendants, alors

$$H\left(S\right)=\sum_{j\in Q_i}H\left(S^j\right)$$

Définition 7

Le systéme physique conditionnel $S^1/\left[S^2\hookrightarrow m_l^2\right]$ est donné par

$$S^1/\left[S^2\hookrightarrow m_l^2\right]=\underset{j=1}{\overset{r_1}{\Lambda}}\left[(S)\hookrightarrow m_j^1/\ S^2\hookrightarrow m_l^2;P\left[(S)\hookrightarrow m_j^1/\ S^2\hookrightarrow m_l^2\right]=p_{j/l}\right].$$

L'entropie de ce systéme est

$$H\left(S^1/\left[S^2\hookrightarrow m_l^2\right]\right)=-\sum_{j=1}^{r_1}p_{j/l}\log_2\left(p_{j/l}\right).$$

Le systéme physique multiple aléatoire (S^1/S^2) s'écrit

$$\left(S^1/S^2\right) = \overset{r_2}{\underset{l=1}{\Lambda}} \left[\overset{r_1}{\underset{j=1}{\Lambda}} \left[(S) \hookrightarrow m_j^1/\ S^2 \hookrightarrow m_l^2; P\left[(S) \hookrightarrow m_j^1/\ S^2 \hookrightarrow m_l^2\right] = p_{j/l}\right] \right],$$

$$\Longrightarrow H\left(S^1/S^2\right) = -\sum_{l=1}^{r_2} \left[\sum_{j=1}^{r_1} p_{j/l} \log_2\left(p_{j/l}\right) \right],$$

Proposition 14

$$H(S) = H\left(S^1\right) + H\left(S^2/S^1\right) + H\left(S^3/\ S^1 \widehat{\Lambda} \dot{S}^2\right) + ... + H\left(S^d/S^1 \widehat{\Lambda}\ \dot{S}^2\ \dot{\widehat{\Lambda}} .... \dot{\widehat{\Lambda}} S^{d-1}\right),$$

Pour la démonstration se referer à [100].

Proposition 15

Les quantités

$$K(P,Q) = -\sum_{i=1}^{n} P_i \log_2\left(Q_i/P_i\right) \quad ; \quad K(Q,P) = -\sum_{j=1}^{n} P_j \log_2\left(P_j/Q_j\right),$$

sont non négatives et on a

$$\sum_{j=1}^{n} P_j \log_2\left(P_j\right) \geq \sum_{j=1}^{n} P_j \log_2\left(Q_j\right); \quad K\left(P,Q\right) = K\left(Q,P\right) \Leftrightarrow P = Q \text{ presque sûrement}$$

$K(.,.)$ n'est une fonction symétrique donc ce n'est pas une distance au sens classique du terme mais caractérise du point de vue statistique la déviation entre les distributions P et Q.

Kullback[55] explique que la quantité $K(P,Q)$ évalue la quantité d'information moyenne perdue si l'on utilise la distribution P alors que la vraie distribution est Q.

Soit S_{Π_d} l'ensemble des systémes physiques multiples aléatoires avec $\prod_{j=1}^{d} r_j$ états

$$S \in S_{\Pi_d} \Rightarrow (S) = (\Theta, P) = \overset{r_1}{\underset{l_1=1}{\widetilde{\Lambda}}} \overset{r_2}{\underset{l_2=1}{\widetilde{\Lambda}}} ... \overset{r_d}{\underset{l_d=1}{\widetilde{\Lambda}}} \left[\left[(S) \hookrightarrow \left(m_{.l_1}, \ldots, m_{l_{d_i}}\right);\ p_{l_1 l,\ldots,l_d}\right]\right].$$

Soit δ l'application définie par

$$S_{\Pi_d} \times S_{\Pi_d} \xrightarrow{\delta} \Re_+$$

$$\left((S_1), (S_2)\right) \longrightarrow \delta\left(S_1, S_2\right) = H\left(S_1\right) + H\left(S_2\right) - \left[\Psi\left(S_1/S_2\right) + \Psi\left(S_2/S_1\right)\right]$$

Où

$$\Psi\left(S_1/S_2\right) = K_d\left(P^{(1)}, P^{(2)}\right)$$

H est la fonction d'entropie, $P^{(1)}$ et $P^{(2)}$ sont des distributions multivariées de dimension d qui définissent les systémes physiques multiples aléatoires S_1, S_2 et K_d est définie comme suit

$$K_d\left(P^1,\ P^2\right) = \sum_{I_1}\sum_{I_2}....\sum_{I_d} P^{(1)}_{I_1...I_d}\log_2\left(P^{(2)}_{I_1...I_d}\right).$$

Proposition 16
δ vérifie

$$1.\ \delta\left(S_1,\ S_2\right) \geq 0$$

$$2.\delta\left(S_1, S_2\right) = 0 \Leftrightarrow S_1 = S_2$$

$$3.\ \ \delta\left(S_1,\ S_2\right) = \delta\left(S_2,\ S_1\right)\ \ \text{(symmétrie)}$$

On admet que si $\delta\left(S_1,\ S_2\right) = 0 \Leftrightarrow S_1 = S_2 \Leftrightarrow \omega_1 = \omega_2$.

δ mesure la dissimilarité entre les systémes physiques. Plus la valeur de δ est petite plus l'incertitude des systémes est grande. δ représente la quantité d'information moyenne perdue si on utilise la distribution $P^1\left(P^2\right)$ pour décrire le systéme alors que c'est l'autre distribution qui est vraie .

δ est une nouvelle version de la distance de Kullback-Leibler entre les distributions P^1 et P^2. En effet, La distance de Kullback-Leibler entre P^2et P^2 est donnée par

$$Ku\left(P^1, P^2\right) = \sum_{I_1}\sum_{I_2}...\sum_{I_d}\left(P^1_{I_1...I_d} - P^2_{I_1...I_d}\right)\log_2\left(P^1_{I_1...I_d}/P^2_{I_1....I_d}\right).$$

En développant cette expression, on retrouve δ.

4 Application numérique

4.1 Procédure pour estimer la distribution conjointe

Dans le cas où toutes les variables qui interviennent dans la description des individus sont discrétes, on donne une procédure inspirée de l'analyse factorielle pour estimer la distribution conjointe et en déduire la valeur de l'entropie des systémes multiples aléatoires correspondants.

Soit $\Delta_i = \left[\Delta_i^1, \ldots,\ \Delta_i^d\right]$ une juxtaposition de d tableaux de données catégorielles . Pour $\omega_i \in \Omega$ fixé, on a

$$P^{(i)}_{l_1 l_2 \ldots l_d} = \Pr\left[S \hookrightarrow \left(m^1_{l_1}, \ldots,\ m^d_{l_d}\right)\right] = \frac{d}{L} N^{(i)}\left(l_1, \ldots,\ l_d\right) = f^{(i)}_{l_1 \ldots l_d}$$

$N^{(i)}(.)$ est le nombre d'occurences simultannées des modalités $m^1_{l_1}$, $m^2_{l_2}$, ..., $m^d_{l_d}$.

$$\sum_{l_1=1}^{r_1} \sum_{l_2=1}^{r_2} \ldots \sum_{l_d=1}^{r_d} f^{(i)}_{l_1 \ldots l_d} = 1.$$

4.2 Exemple numérique

On considère 6 individus déscrits par 2 variables qualitatives à 2 et 3 modalités.10 observations pour chaque individu.

Table1: La juxtaposition des matrices 0/1 associées aux individus

ω_1					ω_2					ω_3				
V^1		V^2			V^1		V^2			V^1		V^2		
m^1_1	m^1_2	m^2_1	m^2_2	m^2_3	m^1_1	m^1_2	m^2_1	m^2_2	m^2_3	m^1_1	m^1_2	m^2_1	m^2_2	m^2_3
1	0	1	0	0	1	0	0	1	0	1	0	1	0	0
0	1	0	0	1	0	1	0	0	1	0	1	0	0	1
1	0	0	1	0	0	1	1	0	0	1	0	0	1	0
0	1	1	0	0	0	1	1	0	0	0	1	0	1	0
0	1	1	0	0	1	0	0	1	0	0	1	1	0	0
1	0	0	0	1	0	1	1	0	0	1	0	0	0	1
0	1	0	1	0	0	1	0	0	1	1	0	1	0	0
1	0	0	1	0	1	0	0	1	0	0	1	0	1	0
1	0	0	1	0	0	1	0	1	0	0	1	1	0	0
1	0	1	0	0	1	0	0	0	1	1	0	0	1	0

ω_4					ω_5					ω_6				
V^1		V^2			V^1		V^2			V^1		V^2		
m^1_1	m^1_2	m^2_1	m^2_2	m^2_3	m^1_1	m^1_2	m^2_1	m^2_2	m^2_3	m^1_1	m^1_2	m^2_1	m^2_2	m^2_3
1	0	1	0	0	1	0	0	1	0	1	0	1	0	0
0	1	0	1	0	0	1	1	0	0	1	0	0	0	1
1	0	0	0	1	1	0	0	0	1	0	1	0	1	0
0	1	1	0	0	1	0	0	1	0	1	0	1	0	0
0	1	0	1	0	0	1	0	1	0	0	1	0	0	1
1	0	0	1	0	1	0	0	1	0	0	1	0	0	1
1	0	0	1	0	0	1	0	0	1	0	1	1	0	0
1	0	0	0	1	1	0	0	1	0	1	0	1	0	0
0	1	0	0	1	1	0	0	1	0	1	0	0	1	0
1	0	1	0	0	1	0	1	0	0	1	0	1	0	0

4.2.1 Procédure pour construire une hiérarchie indicée sur ces objets

Table 2: Les distributions empiriques qui représentent les individus .

	(m_1^1, m_1^2)	(m_1^1, m_2^2)	(m_1^1, m_3^2)	(m_2^1, m_1^2)	(m_2^1, m_2^2)	(m_2^1, m_3^2)
F_1	0.2	0.3	0.1	0.2	0.1	0.1
F_2	0.1	0.2	0.1	0.3	0.1	0.1
F_3	0.2	0.2	0.1	0.2	0.2	0.1
F_4	0.2	0.2	0.2	0.1	0.2	0.1
F_5	0.1	0.4	0.2	0.1	0.1	0.1
F_6	0.4	0.1	0.1	0.1	0.1	0.1

Le programme écrit en visual basic déroulé sur cet exemple a donné les résultats suivants:

Table 3: Les entropies des systèmes physiques conditionnelles associés aux 6 objets

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	$2,4464$	$2,6049$	$2,5219$	$2,6219$	$2,6219$	$2,8219$
S_2	$2,6049$	$2,4464$	$2,6219$	$2,8219$	$2,8219$	$2,9219$
S_3	$2,6049$	$2,7049$	$2,5219$	$2,6219$	$2,8219$	$2,8219$
S_4	$2,7049$	$2,8634$	$2,6219$	$2,5219$	$2,7219$	$2,8219$
S_5	$2,4879$	$2,6634$	$2,6219$	$2,5219$	$2,3219$	$3,0219$
S_6	$2,6634$	$2,8634$	$2,6219$	$2,6219$	$3,0219$	$2,3219$

Etape 1. A partir de la matrice de similarité, on obtient

$$\min_{l\neq t;\ l,t=1,\ldots,5} \{dist\,(S_l, S_t)\} = dist\,(S_1, S_3) = 0,1585.$$

Alors les individus ω_1 et ω_3 sont agrégés en un individu artificiel ω_7 qui sera placé dans la derniere ligne de de la matrice la ligne et la colonne correspondantes aux objets ω_1 et ω_3 sont supprimés

Etape 2. De la nouvelle matrice de similarité, on obtient

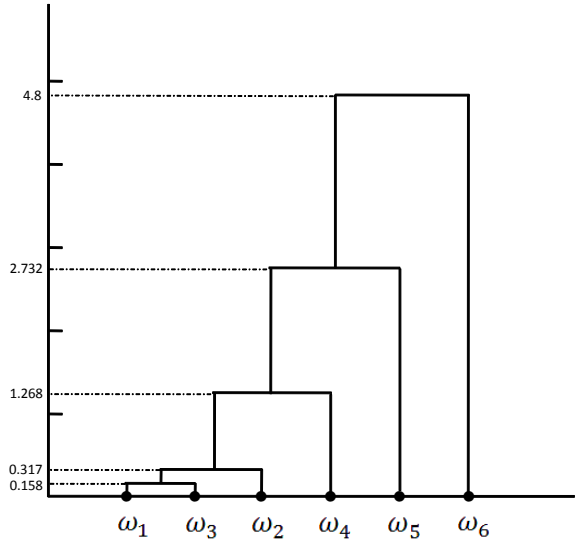
$$\min_{l\neq t;\ l,t=2,4,5,6} \{dist\,(S_l, S_t)\} = dist\,(S_2, S_7) = 0,317.$$

Les individus ω_2 et ω_7 sont agrégés en un individu artificiel ω_8 .

Etape 3. $\min_{l\neq t;\ l,t=4,5,6} \{dist\,(S_l, S_t)\} = dist\,(S_4, S_8) = 1,268.$
Les individus ω_4 et ω_8 sont agrégés en un individu artificiel ω_9 .

Etape 4. $\min_{l\neq t;\ l,t=5,6,9} \{dist\,(S_l, S_t)\} = dist\,(S_5, S_9) = 2,732.$
Les individus ω_5 et ω_9 sont agrégés en un individu artificiel ω_{10}.Les individus ω_6 et ω_{10} sont agrégés en un individu artificiel ω_{11} $dist\,(S_6, S_{10}) = 4,8.$

La hiérarchie obtenue est



- Dans cet exemple, on commence par regrouper les 2 objets les plus proches au sens de l'indice de distance entre les systémes physiques correspondants.
- Plus on monte dans la construcution de la hiérarchie plus les états des systémes mélanges obtenus sont incertains.
- L'exemple montre que l'indice de distance de Kullback Leibler et l'indice d'aggregation du saut minimum conduisent à la construcution de systemes physiques avec le maximum d'entropie donc à un systéme physique où tous les états sont équiprobables.
- Si le nombre total des modalités des variables est assez grand comparativement au nombre d'observations, la fréquence de choisir telle ou telle modalité devient petite et plusieurs fréquences seront nulles: Les modalités de fréquences nulles seront élaguées des états du systeme et n'interviennent pas dans le calcul des distances. Cela peut conduire à l'impossibilité de comparer les systèmes donc les éléments de la structure.

$$\begin{bmatrix} case\ number & cluster & dis\tan ce \\ \omega_1 & 1 & 4.201 \\ \omega_2 & 2 & 8.040 \\ \omega_3 & 1 & 3.598 \\ \omega_4 & 2 & 8.040 \\ \omega_5 & 1 & 3.324 \\ \omega_6 & 1 & 3.4118 \end{bmatrix}$$

- Si l'on ne tient pas compte de la variabilité des variables, les individus ω_2 et ω_4 sont affectés à la première classe et sont équidistants du noyau de cette classe comme le montre la partition ci dessus obtenue dans le cas où chacun des 6 individus est décrit par sa valeur la plus fréquente pour chacune des variables qui le décrit .
- Si l'on utilise la procédure que nous avons développée et en coupant la hierarchieau niveau du $3^{ième}$ palier, l'individu ω_2 et l'individu seront dans 2 classes diffèrentes.

- Dans le cas où la variabililté des observations joue un rôle important particulièrement en médecine, dans la description des individus la classification faite sans tenir compte de cette variabilité peut conduire à des résultats érronés comparativement à la réalité des données.

4.3 Détails sur le déroulement de cet algorithme

Tableau contenant les entropies conditionnelles (T)

$$T = [H(S_i/S_j)]_{\substack{i=1,\,6 \\ j=1,\,6}} \quad \text{et} \quad H(S_i/S_i) = H(S_i)\ \forall i = 1,\,2,...,6$$

$\mathbf{2,44643934}$	$2,60493559$	$2,52192809$	$2,62192809$	$2,62192809$	$2,82192809$
$2,60493559$	$\mathbf{2,44643934}$	$2,62192809$	$2,82192809$	$2,82192809$	$2,92192809$
$2,60493559$	$2,70493559$	$\mathbf{2,52192809}$	$2,62192809$	$2,82192809$	$2,82192809$
$2,70493559$	$2,86343184$	$2,62192809$	$\mathbf{2,52192809}$	$2,72192809$	$2,82192809$
$2,48794309$	$2,66343184$	$2,62192809$	$2,52192809$	$\mathbf{2,32192809}$	$3,02192809$
$2,66343184$	$2,86343184$	$2,62192809$	$2,62192809$	$3,02192809$	$\mathbf{2,32192809}$

Matrice de dissimilarité à l'étape 0

$$\Delta^0 = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S6 \end{array} \begin{array}{c} \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{pmatrix} 0 & & & & & \\ 0,3169925 & 0 & & & & \\ 0,15849625 & 0,35849625 & 0 & & & \\ 0,35849625 & 0,7169925 & 0,2 & 0 & & \\ 0,34150375 & 0,7169925 & 0,6 & 0,4 & 0 & \\ 0,7169925 & 1,0169925 & 0.6 & 0,6 & 1,4 & 0 \end{pmatrix} \end{array}$$

$$\underset{l,t=1,6}{Min}\{\delta(S_l, S_t)\} = \delta(S_1, S_3) = 0,15849625$$

Les objets 1 et 3 sont agrégés en un seul objet 7 et sont supprimés et remplacés par l'objet 7 que l'on a placé dans la dernière ligne.

Si A est une classe de systémes physiques multiples aléatoires, alors

$$\Delta(S, A) = Min\{\delta(S, H)\,; H \in A\}$$

Δ est l'indice d'aggrégation entre classes d'objets de Sneath[**?**].

Matrice de dissimilarité à l'étape 1

$$\begin{array}{c} \\ S_2 \\ S_4 \\ S_5 \\ S_6 \\ S_7 = S_1 \widetilde{\wedge} S_3 \end{array} \begin{array}{c} \begin{array}{ccccc} S_2 & S_4 & S_5 & S_6 & S_7 = S_1 \widetilde{\wedge} S_3 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 0,7169925 & 0 & & & \\ 0,7169925 & 0,4 & 0 & & \\ 1,0169925 & 0.6 & 1.4 & 0 & \\ 0,3169925 & 0.2 & 0,34150375 & 0.6 & 0 \end{pmatrix} \end{array}$$

$$\underset{l \neq t=2,4,5,,6,7}{Min}\{\delta(S_l, S_t)\} = \delta(S_2, S_7) = 0,3169925$$

Les objets 2 et 7 sont agrégés en un seul objet 8 , ils sont supprimés et remplacés par l'objet 8 que l'on a placé dans la dernière ligne.

Matrice de dissimilarité à l'étape 2

$$\begin{array}{cccccc} & S_4 & S_5 & S_6 & S_8 = S_2\widetilde{\wedge}S_7 \\ S_4 & 0 & & & \\ S_5 & 2,86797 & 0 & & \\ S_6 & 4,06797 & 5,6 & 0 & \\ S_8 & 1,26797 & 1,366015 & 2,4 & 0 \end{array}$$

$$\underset{l\neq t=4,5,6,8}{Min}\left\{\delta\left(S_4,S_8\right)\right\}=\delta\left(S_4,S_8\right)=1,26797$$

Les objets 4 et 8 sont supprimés et remplacés par l'objet 9 placé dans la dernière ligne:

Matrice de dissimilarité à l'étape 3

$$\begin{array}{cccc} & S_5 & S_6 & S_9 = S_4\widetilde{\wedge}S_8 \\ S_5 & 0 & & \\ S_6 & 11,2 & 0 & \\ S_9 = S_4\widetilde{\wedge}S_8 & 2,73203 & 4,8 & 0 \end{array}$$

$$\underset{l\neq t=5,6,8}{Min}\left\{\delta\left(S_l,S_t\right)\right\}=\delta\left(S_5,S_9\right)=2,73203$$

Les objets 5 et 9 sont supprimés et remplacés par l'objet 10
L'objet 6 est donc aggregé à l'objet 10.et $\delta\left(S_6,S_{10}\right)=4.8$

4.4 Application à des données réelles

Les données proviennent d'une étude du niveau de développement humain et économique de 48 départements

$$[R_1,\ldots,\ R_{48}]$$

d'un pays donné. L'objectif est de construire une hiérarchie indicée sur les 48 départements selon leur niveau de développement économique et humain afin de faire des comparaisons entre départements et entre sous regions à l'interieur des départements

Chaque département R_i regroupe L_i sous régions

$$\left[C_1^i,\ldots,C_{L_i}^i\right]\left(\sum_{i=1}^{n}L_i=1541\right)$$

qui constitue un territoire.

$\forall\ i=1,\ldots,\ n$ et $l=1,\ldots,\ L_i$, on mesure l'indice de développement économique et l'indice de développement humain idE et idH. Ces 2 indices composites ont été développé par les experts du PNUD et dépendent de la situation géographique et de la spécifité de la sous région

$$0\leq idE\left(C_l^i\right)\leq 1\ \text{ et }\quad 0\leq idH\left(C_l^i\right)\leq 1$$

Plus la valeur de l'indice est proche de 1 plus le développement économique ou humain de la sous régions est jugé satisfaisant. Ces indices ne sont pas calculés de la même façon selon que la sous région est classée rurale ou urbaine: Le classement fait à partir de ces valeurs n'a donc aucun sens.

Pour chaque $i = 1,\ldots, n$

$$\underset{(L_i,2)}{R_i} \mapsto \begin{bmatrix} & idE & idH \\ C_1^i & . & . \\ \vdots & \vdots & \vdots \\ C_{L_i}^i & . & . \end{bmatrix}$$

La structure de données n'est pas exploitable sous cette forme, il est néccéssaire de la transformer. Les experts du PNUD découpent les observations en 5 classes ou quintiles et affectent chacune sous région au quintile qui correspond à sa valeur pour chaque des indices.

Nous déterminons ainsi, pour les 48 séries d'observations des 2 indices

$$idE \rightarrow q_{i1}^1, \ldots, q_{i5}^1$$

$$idH \rightarrow q_{i1}^2, \ldots, q_{i5}^2$$

$\Longrightarrow$Les intervalles interquintiles

$$\begin{cases} \Im_{i1}^1, \ldots, \Im_{i5}^1 \\ \Im_{i1}^2, \ldots, \Im_{i5}^2 \end{cases}$$

Nous nous sommes pas intéressés par les valeurs des

$$\left\{q_{i1}^j, \ldots, q_{i5}^j;\ j = 1, 2\right\}$$

mais aux appartenances des sous régions aux intervalles interquintiles.

Pour $i = 1,\ldots, n$, le tableau R_i est transformé en une table de données binaires

$$T_i = \left(\begin{array}{c} \overbrace{\Im_1^1 \ \ldots \Im_5^1}^{idE} \\ 0 \ \ 0 \ 1 \ \ 0 \ \ 0 \\ 0 \ \ 0 \ \ 0 \ \ 0 1 \\ \ldots\ldots\ldots \\ 0 \ \ 1 \ \ 000 \\ \underbrace{0\ 0 \ \ 100} \end{array}\right) \left(\begin{array}{c} \overbrace{\Im_1^2 \ldots \Im_5^2}^{idH} \\ 0 \ \ 0 \ \ 01\ 0 \\ 0\ 10 \ \ 0\ 0 \\ \ldots\ldots\ldots \\ 0\ 0\ 010 \\ \underbrace{0 \ \ 010\ 0} \end{array}\right)$$

Exhiber les éventuelles disparités entre sous régions à l'interieur des départements. Les observations sont regroupées dans les tableaux $T_1, \ldots, T_n$ qui constituent une structure de juxtaposition de données catégorielles.

5 Conclusion

- Dans ce chapitre, la définition de l'entropie est celle émise par Shannon Weiner[1936]. Cette définition est encore utilisée dans la théorie de l'information et du signal.
- Le formalisme proposé donne une nouvelle explication de l'aspect pratique de la distance de Kullback-Leibler type comme un indice de distance entre les éléments représentatifs d'une strucuture de tableaux de données categorielles.
- Il est peut être possible d'étendre les résultats obtenus dans le cas d'une structure de juxtapositions de tableaux de mesure et dans le cas de données fonctionnelles.

Chapitre V
Classification des données avec erreurs de mesure

1 Introduction

- Lors des observations repétées, nous obtenons généralement des dispersions dans les résultats. Si les résultats sont aléatoires un traitement statistique permet de connaître la valeur la plus probable de la grandeur mesurée et de fixer les limites de l'incertitude.

- Ainsi, si l'on est conduit à effectuer pour chaque objet plusieurs mesures et qu'on obtienne à chaque fois un résultat diffèrent ceci est certainement dû à des phénomènes perturbateurs: dispersion statistique. Cette dispersion statistique désigne le fait que les observations sont proches les unes des autres ou au contraire éparpillées.

- Une structure de juxtaposition de tableaux de mesure peut donc être transformée en une structure de tableau de données avec erreurs de mesure. Sous l'hypothése raisonable de normalité des observations. Nous apportons un autre éclairage en proposant 2 approches pour classifier des objets matriciels selon que les variables actives soient corrélées ou non. La première approche traite le cas où les variables sont non corrélées donc indépendantes tandis que la seconde traite le cas où les variables sont quelconques.

Les procédures proposées consistent à transformer dans les 2 cas, ces objets matriciels en objets intervalles. Ainsi, la valeur la plus probable de chaque individu ω_i est donnée à l'interieur d'un sous ensemble R_i de $\mathbb{R}^d$. Les sous ensembles R_i sont des pavés, ou des ellipsoides de $\mathbb{R}^d$estimés de manière classique sous la condition que $\Pr[x_i \in R_i] \succeq 1-\beta$ avec β seuil fixé.

Si les variables ayant participé à la description des individus, sont corrélées donc quelconques, ces ensembles sont des éllipsoides. Dans chaque cas, nous adaptons un indice de distance et justifions son utilisation.

- Dans le second cas, plus général, nous avons utilisé la distance pondérée de Mahanalobis et avons construit un algorithme de type k-means.

- Nous montrons qu'une fois les clusters obtenues, leur caractérisation n'est pas entachées d'erreurs et est donnée de manière exacte . Ce qui parait tout à fait dans l'ordre des choses.

Le premier cas se ramène à la construction d'un algorithme de type k-means avec la distance classique euclidienne entre sommets des pavés qui décrivent les individus. Nous étudions plusieurs cas selon que les erreurs soient placées au niveau des observations et ou au niveau des noyaux des classes.

Habituellement, pour analyser les éléments constitutifs d'une structure de tableaux multiples, on résume chaque colonne de chaque tableau par un nombre. Les objets matriciels sont transformés en vecteurs et la structure devient un tableau classique définissant la description d'un nombre fini d'individus par un nombre fini de variables. Cependant, cette synthèse nécessite plusieurs hypothèses quant au choix de la valeur qui résume les différentes observations de l'individu pour la variable [*voir chapitre II*].

Le cas où chaque objet est décrit par une matrice est trés fréquent en pratique, il génére des problèmes mathématiques intéréssants et reste un vaste domaine de recherche et d'investigation.

Les structures de données de ce type sont trés fréquentes en pratique. On peut citer

1. Le cas de données concernant la pollution engendrée par les automobilistes au niveau d'un important carrefour de trafic routier d'une grande métropole. Les taux des principaux polluants sont mesurés chaque heure de la journée sur toute l'année. Les principaux polluants peuvent être : le taux de poussière, d'ozone, de monoxyde de carbone, de monoxyde de zinc et de dioxyde de soude etc... Chaque jour est décrit par une matrice de données quantitatives de dimension 24 fois le nombre de polluants. On désire regrouper les jours selon les seuils de pollution dont les classes obtenues peuvent être expliquées par d'autres variables concernant par exemple la météorologie.

2. En médecine, si l'on dispose d'une population finie de patients atteints d'une longue maladie où chaque malade subit des controles réguliers sur l'etat et de l'évolution de sa maladie sur une grande période. Il s'agit là aussi de regrouper les malades selon l'évolution de la maladie.

Pour ce type de données, la phase de regroupement des observations par le cumul journalier ou par la moyenne journalière pour chaque polluant dans le premier exemple peut conduire à des résultats érronnés. En effet, les pics de pollution dépendent des heures de pointe, des jours ouvrables et d'autres évènements qui peuvent soit ralentir ou augmenter la densité de la circulation donc le seuil de pollution .

De même qu'en médecine, la moyenne journalière ou le cumul journalier de la tension artérielle n'ont pas de sens et ce sont ses variations qui peuvent dénoter des cardiopathies. Aussi, il est nécéssaire de traiter l'information de la manière la plus proche de la réalité.

Cette structure a été présentée dans [1] et n'a pas eu l'attention à la mesure de sa fréquence en pratique**.** Dans Rebbouh[80], est proposé un algorithme du type k-means basé sur le produit scalaire de Hilbert-Schmidt pour classifier des d'objets matriciels de même dimension.

2 Estimation des sous ensembles R_i

Soit $\Omega = \{\omega_1, \ldots, \omega_N\}$ un ensemble fini de N individus à classifier. On suppose que pour tout $i \in \{1, \ldots, N\}$, ω_i est décrit par d variables quantitatives dont les observations sont regroupées dans le tableau T_i de dimension $N_i \times d$.

$$\omega_i \rightarrow \ T_i = \begin{bmatrix} \theta_1^i \\ \vdots \\ \theta_l^i \\ \vdots \\ \theta_{N_i}^i \end{bmatrix} \begin{bmatrix} V^1 & \dots & V^d \\ x_{i1}^1 & \dots & x_{i1}^d \\ \vdots & \vdots & \vdots \\ x_{il}^1 & \dots & x_{il}^d \\ \vdots & \vdots & \vdots \\ x_{iN_i}^1 & \dots & x_{iN_i}^d \end{bmatrix},$$

x_{il}^j est la $l^{ième}$ observation de l'individu ω_i pour la variable V^j

$$T_i = \left[\begin{array}{ccc} X_1^i & \dots & X_{N_i}^i \end{array} \right]; X_l^i \in \mathbb{R}^d,$$

T_i est un objet matriciel qui regroupe les observations répétées de l' individu ω_i pour les d variables quantitatives qui le décrivent. Ces objets matriciels peuvent être de dimensions diffèrentes. La structure

$$\Delta = [T_1, ..., T_n],$$

est une structure de juxtaposition de tableaux de mesure($TofM$).

2.1 Les d variables sont non corrélées

2.1.1 Les intervalles de confiance associés

Pour un seuil de signification β fixé, on construit les intervalles de confiance associés aux observations variées de chaque individu pour les variables qui le décrivent même si le nombre des observations diffère d'un individu à l' autre et que ces nombres soient relativement petits.

2.1.2 Premier cas σ_i^j sont connus pour $i = 1, \dots, N$ et $j = 1, \dots, d$

L'individu ω_i est observé à l'interieur de l'intervalle I_i^j de longueur $2\varepsilon_i^j(\beta)$,pour $j = 1,\dots, d$,

Proposition 17

Cet intervalle de confiance au niveau de signification β fixé est donné par

$$I_i^j(\beta) = \left]\mu_i^j - \varepsilon_i^j(\beta) \ ; \ \mu_i^j + \varepsilon_i^j(\beta)\right[,$$

où

$$\varepsilon_i^j(\beta) = \sqrt{\frac{\sigma_i^j}{N_i - 1}} \arg\Phi^* \left(\frac{1+\beta}{2}\right),$$

$\Phi^*(y)$ est la valeur pour laquelle la fonction cumulative cdf de la distribution normale est y et

$$\begin{cases} \mu_i^j = \frac{1}{N_i} \sum_{l=1}^{L_i} x_{il}^j, \\ \sigma_i^j = \left(\frac{1}{N_i - 1} \sum_{l=1}^{N_i} \left[x_{il}^j - \mu_i^j\right]^2 \right)^{1/2}. \end{cases}$$

2.1.3 Deuxième cas$\left\{\mu_i^j \ ; \ j \succeq 1\right\}$ sont connus et $\sigma_i^j \neq 0 \ ; \forall j \succeq 1$

De la même manière, nous obtenons les intervalles de confiance estimés des variances au seuil de signification β fixé sont

$$I_i^j(\beta) = \left]-t_\beta.D_i^j \ \ ; \ \ +t_\beta.D_i^j\right[$$

où

$$D_i^j = \sqrt{\frac{0.8.N_i + 1/2}{N_i(N_i-1)}}\sigma_i^j \quad \text{et} \quad t_\beta = \arg\Phi^*\left(\frac{1+\beta}{2}\right).$$

Il est à noter que ces estimateurs dépendent du nombre d'observations. Cependant, si ces matrices n'ont pas la même dimension, on peut envisager une étape de complétion. En effet, si les tableaux

$$\{T_i; \ i = 1, \ldots, \ N\}$$

n'ont pas la même dimension $\Leftrightarrow [\exists i \neq i' \text{ tel que } N_i \neq N_{i'}]$. Soit L le plus petit commun multiple des

$$\{N_i \ ; i = 1, ..., N\}$$

$$P = PPCM(N_i, \ i = 1, \ N),$$

Il existe donc r_i tel que

$$p = N_i \times r_i.$$

Maintenant, on duplique r_i fois chaque tableau T_i, on obtient ainsi, un nouveau tableau $\widehat{T_i}$ de dimension $P \times d$. Si les N_i sont des grands nombres, le plus petit commun multiple devient trés grand et la procédure conduira néccessairement à une structure de données de grands tableaux. De plus, cette technique de complétion détruit complétement la chronologie des observations. Ce n'est donc pas une technique recommandée. Il est plus raisonable de ne pas procéder à une étape de complétion avant d'entamer la phase de classification .

L'objet ω_i est donc décrit par le pavé R_i de $\mathbb{R}^d$ qui s'écrit

$$R_i = \prod_{j=1}^{d} I_i^j(\beta) \, ; I_i^j(\beta) \subset \mathbb{R}(I_i^j(\beta) \text{ est un intervalle})$$

2.2 Les d variables sont quelconques

Deux approches peuvent être envisagées:

1. La première approche consiste à utiliser une procédure heuristique basée sur les techniques factorielles de réduction. Cette procédure permet de construire de nouvelles variables non corrélées à partir des variables initiales pour chacun des objets matriciels en présence. A ce stade, deux problèmes importants sont poses. Le premier problème concerne le choix de la dimension ou du nombre de nouvelles variables à retenir. Le second problème concerne le choix du sous espace compromis ou commun qui permet de comparer les diffèrents éléments de la structure et donc de procéder à leur classification.

2. La seconde approche, plus appropriée au sens où elle evite les 2 problèmes cités précedement et qui demeurent non totalement résolus, consiste à estimer directement le couple (μ_i, Σ_i) qui résumera la description de l'individu ω_i.

2.2.1 Première approche: Choix du sous espace compromis et estimation des intervalles de confiance

- L'analyse en composantes principales déroulée sur chaque tableau T_i conduit à la construction de r_i axes factoriels orthogonaux sur lesquels les projections des N_i observations de ω_i donnent de nouvelles variables non corrélées qui sont les composantes principales . Nous obtenons ainsi, en deroulant N analyses, N systèmes d'axes
$$\left\{\left\{\Delta_{u_1^{(i)}},\ \ldots,\ \Delta_{u_{r_i}^{(i)}}\right\};\ i=1,\ \ldots,N\right\}$$

- Dans le but de comparer les objets et de pouvoir les regrouper en classes, on doit les représenter dans un même référentiel. Ce problème fondamental de la recherche de système compromis ou commun des N systèmes, fait encore l'objet de recherche en mathématiques et intéresse particulièrement la géométrie différentielle, on cite les travaux de Orlick.P and Tera.H[74][1992] . Dans[10] sont proposés des critères pour le choix de ce referentiel mais les procédures proposées sont difficilement justifiables et ne constituent pas l'objet de notre étude.

- La procédure heuristique que nous proposons s'inspire des techniques utilisées dans le cas de la représentation simultannée des individus et des variables dans un même referentiel. Il s'agit de construire d^* axes pour l'analyse du tableau T_1. Ces d^* axes, dont le nombre est choisi par une procédure simple voir $(*)$, constituerons le repère de référence et on plongera les autres axes dans ce système. On effectuera donc $N-1$ rotations de centre $0_{\mathbb{R}^d}$ et d'angles
$$\left\{\theta_l^{(1)}; l=1,\ldots,N-1\right\}$$

La projection de la lième observation de l'individu ω_i sur le rième axe factoriel est donnée par
$$W_i^r = T_i.u_r,$$

où u_r est le vecteur propre normè associè à la rième plus grande valeur propre λ_r de la matrice d'inertie Σ_i. Nous supposons que le tableau T_i est centré

$$\Sigma_i = \frac{1}{N_i}(T_i)'(T_i),$$
$$d^* = Min\ \{r_i,\ i = 1, \ldots,\ n\}\ (*)$$

où r_i est le plus petit nombre positif tel que

$$\frac{\lambda_1^i + \lambda_2^i + \ldots + \lambda_{r_i}^i}{tr(\Sigma_i)} \geqslant a,$$

a est fixé $(a = 0.7)$, λ_l^i est la lième plus grande valeur propre de la matrice d'inertie Σ_i . Les d^* nouvelles variables sont centrées et données par

$$\begin{cases} W_{(i)}^l = T_i.u_l \\ \overline{W_i^l} = 0 \end{cases} \quad \forall l = 1, \ldots, d^*.$$

Les intervalles de confiance I_i^j qui décrivent l'individu ω_i sont estimés à partir des observations contenues dans la matrice

$$Y^{(i)} = \begin{array}{c} \begin{array}{cccccc} W^1 & \ldots & W^j & \ldots & W^{n*} \end{array} \\ \begin{bmatrix} y_{i1}^1 & \ldots & y_{i1}^j & \ldots & y_{i1}^{n*} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{il}^1 & \ldots & y_{il}^j & \ldots & y_{il}^{n*} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{iN_i}^1 & \ldots & y_{iN_i}^j & \ldots & y_{iN_i}^{n*} \end{bmatrix} \end{array}$$

$$I_i^j(\beta) = \left]-t_\beta.D_i^j\ \ ;\ \ +t_\beta.D_i^j\right[$$

où

$$D_i^j = \sqrt{\frac{0.8.N_i + 1/2}{N_i(N_i - 1)}}\sigma_i^j;\ t_\beta = \arg\Phi^*\left(\frac{1+\beta}{2}\right).$$

où

$$\sigma_i^l = \left(\frac{1}{N_i - 1}\sum_{t=1}^{L_i}\left[(T_i).(u_l)_t\right]^2\right)^{1/2} = \left(\frac{1}{N_i - 1}\sum_{t=1}^{L_i}\left(y_{it}^j\right)^2\right)^{\frac{1}{2}}$$

2.2.2 Seconde approche: Estimation des couples (μ_i, Σ_i)

Cette approche fera l'objet d'une étude détaillée car elle nous parait plus générale de par ses conditions d'application. Les espérances mathématiques $\left\{\mu_i^j;\ j=1,\ \ldots,d\right\}$ et les composantes des matrices de variances covariances estimées Σ_i sont données par

$$\mu_i = \begin{pmatrix} \mu_i^1 \\ \vdots \\ \mu_i^d \end{pmatrix} \text{ avec } \mu_i^j = \frac{1}{L_i}\sum_{l=1}^{L_i} x_{il}^j$$

$$(\Sigma_i)_{jk} = Cov(V_{(i)}^j, V_{(i)}^k) = \Sigma_{jk}^i = \frac{1}{N_i - 1}\sum_{l=1}^{L_i}\left(x_{il}^j - \mu_i^j\right)\left(x_{il}^k - \mu_i^k\right),$$

Nous assurons dans tous les cas que la condition de normalité des données est vérifiée et que

$$P\left[x_i \in R_i\right] \succeq 1-\beta$$

De plus, ces estimateurs sont sans biais, convergents et consistants même si le nombre N_i d'observations est assez petit. Il faut que $N_i \succ d\ ; \forall i = 1, N$ pour que les matrices de variance covariances estimées soient régulières. De plus, pour que la normalité des observations soit envisageable, il faut que ces nombres dépassent 30.

3 Choix de l'indice de distance entre éléments représentatifs

3.1 Les individus sont décrits par des pavés de $\mathbb{R}^d$

Soit $\Re_{V_d}$ l'ensemble des pavés droits de $\mathbb{R}^d$

$$R \in \Re_{V_d} \Longleftrightarrow R = \prod_{j=1}^{d} \Im^j \text{ où } \Im^j \subset \mathbb{R}\ ; \Im^j = [x_{j-}s_j;\ x_{j+}s_j]$$

Proposition 18

Le nombre de sommets présents dans un un pavé droit de $\mathbb{R}^d$ qui représente l'ordre de ce graphe vaut 2^d.

A chaque

$$R = \prod_{j=1}^{d} [[x_{j-}\sigma_j\ ;\ x_{j+}\sigma_j]]$$

On lui associe les sommets $S_t^{(.)} \in \mathbb{R}^d$

$$\left(S_t^{(.)}\right)' = \left[a_1^t, ..., a_d^t\right] \ \in \mathbb{R}^d \text{ avec } a_l^t = \begin{cases} x_{l-}s_l \\ \text{ou} \\ x_{l+}s_l \end{cases}$$

3.1.1 Indice de distance entre objets intervalles

Proposition 19

L'application δ définie par

$$\delta : \Re_{V_d} \times \Re_{V_d} \longmapsto \mathbb{R}^+$$

$$(R_1, R_2) \longmapsto \delta\ (R_1, R_2) = \frac{1}{2^d} \sum_{t=1}^{2^d} \left[\left(\left\| S_t^{(1)} - S_t^{(2)} \right\| \right) \right],$$

où $S_t^{(1)}$, $S_t^{(2)} \in \Re^d$et $\|.\|$ est la norme euclidienne, δ est une distance entre objets intervalles.

On suppose que

$$\delta\ (R_1, R_2) = 0 \Leftrightarrow \omega_1 = \omega_2$$

3.1.2 Indice d'aggrégation entre objets et pavés de $\mathbb{R}^d$

L'application D

$$D : \mathbb{R}^d \times \Re_{V_d} \longmapsto \mathbb{R}* :$$

$$(X_i, R) \longmapsto D\,(X_i, R) = \frac{1}{2^d} \sum_{t=1}^{2^d} \left\| X_i - S_t^{(.)} \right\| = \frac{1}{2^d} \sum_{t=1}^{2^d} \left(\sum_{j=1}^{d} \left((X_i)_j - \left(S_t^{(.)} \right)_j \right)^2 \right)^{\frac{1}{2}}$$

Où $S_t^{(.)} \in \mathbb{R}^d$ est le $t^{ième}$ sommet du pavé R , $(X_i)_j$ et $\left(S_t^{(.)} \right)_j$ sont respectivement la $j^{ième}$ coodonnée de X_i et $S_t^{(.)}$. D mesure la distance moyenne entre le pavé R et le vecteur $X_i \in \mathbb{R}^d$.

3.2 Les individus sont décrits par des ellipsoides de $\mathbb{R}^d$

3.2.1 Distance de Mahannalobis pondérée.

Définition 8

Soit $M_d\,(\mathbb{R})$ l'ensemble des matrices réelles définies positives, l'application d_{ma} définie par

$$\left(\mathbb{N}\times\mathbb{R}^d\times M_d\left(\mathbb{R}\right)\right)\times\left(\mathbb{N}\times\mathbb{R}^d\times M_d\left(\mathbb{R}\right)\right)\to\mathbb{R}$$

$$\left(\ T_i,T_l\right)\to d_{ma}\left(\ T_i,T_l\right)=\frac{N_i\left\|\mu_i-\mu_l\right\|_{\Sigma_i^{-1}}+N_l\left\|\mu_i-\mu_l\right\|_{\Sigma_l^{-1}}}{N_i+N_l}$$

est la distance pondérée de Mahannolobis .

- Cette distance tient compte non seulement de la moyenne des observations et de leur nombre mais aussi de leur variabilité.
- La distance de Mahanalobis (nommée d'après **Prasantra Chandra Mahanalobis**[1936] scientifique indien, chercheur en statistique) permet de comparer deux échantillons . En effet, si leur précision est grande, alors une différence entre leurs moyennes est significative, et on peut les considérer comme différents. En revanche, si la précision est faible, donc la dispersion est grande, alors une différence entre leurs moyennes n'est pas significative, elle est peut-être uniquement dûe au manque de précision.
- Comme toujours en statistique, on peut parfois affirmer que deux quantités diffèrent mais jamais qu'elles sont vraiment égales : on peut parfois dire non, jamais oui.
- Mathématiquement, cette distance quadratique est la différence entre les deux moyennes au carré, pondérée par leur précision (plus précisément, divisée par la somme de leur variance).

On suppose de plus que

$$1.d\left(\ T_i,T_l\right)=0\Leftrightarrow\omega_i=\omega_l$$

$$2.N_i+N_l\succ d.$$

Cette seconde hypothése assure que les matrices Σ_i et Σ_l soient singulières.

4 Classification d'objets décrits avec erreurs de mesure

4.1 Objets décrits par des pavés de $\mathbb{R}^d$

Le modèle des données est de la forme

$$\Delta=\left[R_1,\ldots,\ R_N\right],$$

ou R_i est un pavé de $\mathbb{R}^d$.

Les valeurs prises par l'individu ω_i pour les d variables quantitatives sont dans un pavé R_i de $\mathbb{R}^d$.

Si les variables sont non correlées, on peut utiliser des techniques de l'analyse factorielle pour réduire chaque tableau T_i et construire ainsi de nouvelles variables non corrélées, on retrouve ainsi le précédent cas. En général, les techniques de l'analyse factorielle nécéssitent des critéres pour le choix d'un espace commun qui permettent de comparer les éléments réduits pour chacun des tableaux . Dans notre cas, ces réductions conduisent à des pavés centrés à l'origine de $\mathbb{R}^d$ ce qui les rend comparables et cela nous permet de comparer les éléments de la structure des données.

4.1.1 Procédure

La procédure suivante montre comment construire ces nouvelles variables non corrélées

Soit T_i le tableau contenant les observations de l'individu ω_i

$$\omega_i \longrightarrow T_i = \begin{bmatrix} V^1 & . & V^j & V^d \\ x_{i1}^1 & . & x_{i1}^j & \dots x_{i1}^d \\ \vdots & \dots & \vdots & \dots \vdots \\ x_{iL_i}^1 & \dots & x_{iL_i}^j & \dots x_{iL_i}^d \end{bmatrix} \longrightarrow T_i = \begin{bmatrix} X_i^1 \\ \vdots \\ X_i^{L_i} \end{bmatrix}; \left(X_i^l\right)' \in \left(\Re^d, \jmath_d\right)$$

Proposition 20

La projection de la *lième* observation de l'individu ω_i sur le *rième* axe factoriel est donnée par

$$W_i^r = T_i.u_r,$$

où u_r est le vecteur propre normalisé associé à la *rième* plus grande valeur propre λ_r de la matrice d'inertie Σ_i, T_i est un tableau centré et

$$\Sigma_i = \frac{1}{L_i} \left(T_i\right)' \left(T_i\right),$$

Soit

$$n^* = Min\left\{ r_i,\ i = 1,\ n\right\}$$

où r_i est le plus petit nombre positif tel que

$$\frac{\lambda_1^i + \lambda_2^i + ... + \lambda_{r_i}^i}{tr\left(\Sigma_i\right)} \geqslant a,$$

λ_l^i est la *lième* plus grande valeur propre de la matrice d'inertie Σ_i. Les n^* nouvelles variables sont centrées et données par

$\forall l = 1,\ n^*$

$$\begin{cases} W_{(i)}^{l} = T_i.u_l \\ \overline{W_i^l} = 0 \end{cases} .$$

Cependant, l'intervalle $\Im_i^l$ qui décrit l'individu ω_i pour les nouvelles variables $W_{(i)}^l$ est estimé par

$$\Im_i^l = \left[-\frac{1}{2}\sigma_i^l; +\frac{1}{2}\sigma_i^l \right],$$

où

$$\sigma_i^l = \left(\frac{1}{L_i - 1} \sum_{t=1}^{L_i} \left[(T_i) \,.\, (u_l)_t \right]^2 \right)^{1/2} .$$

4.1.2 Problème d'optimisation

On désire regrouper les N individus en K classes homogènes. Les valeurs prises par les individus sont soit à l'interieur d'un pavé de $\mathbb{R}^d$ soit des vecteurs de $\mathbb{R}^d$. L'homogéinité des classes se mesure a l'aide d'un critére de la somme des inerties intra -classes. Ce critére s'exprime

$$Cr\,(P, L) = \sum_{k=1}^{K} H_k = \sum_{k=1}^{K} \left[\sum_{\{l\} \in P_k} \delta\,(x_l, l_k) \right],$$

l_k est le noyau de la classe P_k, x_l est l'observation de l'individu ω_l et δ est un indice de distance entre objets et éléments représentatifs des classes.

Soit IP_k, IL_k respectivement l'ensemble des k-partitions, l'ensemble des k-représentatifs éléments des classes

$$P \in IP_k \Rightarrow P = \{P_1, \ldots, P_k\}$$

et

$$L \in IL_k \Rightarrow L = (l_1, \ldots,\ l_k)$$

Le problème de classification se réduit à la recherche d'une partition à k classes et de ses k noyaux qui minimisent le critére C_r. Il s'agit de résoudrele problème d'optimisation suivant

$$\underset{P \in IP_K, L \in IL_k}{Min} C_r\,(P, L)\,.$$

Les algorithmes habituellement utilisées pour résoudre ce type de problème sont du type $k - means$. Ces algorithmes consistent à définir une fonction d'affectation f et une fonction de représentation g .Les fonctions f et g sont utilisées alternativement pour faire décroître le critére. f et g doivent vérifier les conditions suivantes

$$f : IL_k \longmapsto IP_k$$

$$L \in IL_k \longmapsto f(L) = P = \{P_1, \ldots, P_k\}$$

$$\underset{P \in IP_K}{Min} [C_r(P, L)] = C_r(f(L), L)$$

$$g : IP_k \longmapsto IL_k$$

$$P \longmapsto g(P) = L = (l_1, \ldots, l_k)$$

$$\underset{L \in IL_K}{Min} [C_r(P, L)] = C_r(P, g(P))$$

Dans ce chapitre, nous présentons une approche pour classifier des objets dont les descriptions sont entachées d'erreurs de mesure et nous étudions trois cas.

4.1.3 Algorithme de classification

On choisi $L^{(0)}$ et on utilise alternativement les fonctions f et g. L'algorithme se déroule comme suit

$$L^{(0)} \xrightarrow{f} P^{(1)} \xrightarrow{g} L^{(1)} \xrightarrow{f} P^{(2)} \xrightarrow{g} ... \xrightarrow{f} P^{(n)} \xmapsto{g} L^{(n)} \longmapsto ...P^{(*)} \xmapsto{g} L^{(*)}$$

L'algorithme s'arrête dés la partition générée ne change plus. On construit ainsi 2 suites V_n et U_n.

Proposition 21

La suite

$$U_n = Cr\left(P^{(n)},\ L^{(n)}\right)$$

converge en décroissant et la suite

$$V_n = \left(P^{(n)}, L^{(n)}\right)$$

est stationnaire à partir d'un certain rang

Preuve:[32]

4.2 Etude de cas

4.2.1 Alternative 1: Les erreurs sont cummulées au niveau des noyaux des classes

Soit $\Re_{V_d}$ l'ensemble des pavés de $\mathbb{R}^d$

$$R \in \Re_{V_d} \Longleftrightarrow R = \prod_{j=1}^{d} \Im^j \text{ où } \Im^j \subset \mathbb{R}.$$

Les objets sont décrits par des vecteurs de $\mathbb{R}^d$. Les noyaux des classes sont observés à l'interieur de pavés de $\mathbb{R}^d$. Pour chaque pavé $R \in \Re_{V_d}$, on associe l'ensemble

$$S_d = \left\{ som(.,\ t) = s_t \in \mathbb{R}^d.,\ t = 1,\ \ldots,\ 2^d \right\},$$

Proposition 22

$$|S_d| = 2^d$$

Pour $d = 2 \hookrightarrow |S_2| = 2^2 = 4 \Rightarrow S_2 = \{som(.,\ t);\ t = 1,\ 2, 3, 4\}$
Les sommets associés pour ces 2 intervalles $\{\Im_i =]m_i;\ M_i[\,; i = 1,\ 2\}$ sont

$$\begin{cases} som(.,(1)) = s_1 = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \\ som(.,(2)) = s_2 = \begin{pmatrix} m_1 \\ M_2 \end{pmatrix} \\ som(.,(3)) = s_3 = \begin{pmatrix} M_1 \\ m_2 \end{pmatrix} \\ som(.,(4)) = s_4 = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \end{cases}$$

Indice de distance entre objets On défini l'application δ par

$$\delta :\ \Omega \times \Omega \longmapsto \Re^+$$

$$(\omega_i, \omega_l) \ \longmapsto \ \delta\ (\omega_i, \omega_l) = \|x_i - x_l\|,$$

où x_i, $x_l \in \Re^d$et $\|.\|$ est la norme euclidienne classique, δ est une distance entre objets.

Indice d'aggrégation entre objets et groupe d'objets. Définition 9
L'application D définie par

$$D :\ \Omega \times \Re_{V_d} \longmapsto \mathbb{R}$$

$$(\omega_i, R) \longmapsto D\ (\omega_i, R) = \frac{1}{2^d} \sum_{t=1}^{2^d} \|x_i - l_t\|$$

où l_t est le *tième* sommet du pavé R qui représente la classe C, $l_t \in \mathbb{R}^d$ et $\|.\|$ est la norme euclidienne classique. D est un indice d'aggrégation entre objets et groupe d'objets et mesure la distance moyenne entre le pavé R et les vecteurs $\left\{x_i \in \mathbb{R}^d; \omega_i \in C\right\}$.

$$D\left(\omega_i, R\right) = \frac{1}{2^d} \sum_{t=1}^{2^d} \left(\sum_{j=1}^{d} \left(x_{ij} - \left(l_t\right)_j \right)^2 \right)^{\frac{1}{2}},$$

x_{ij} et $\left(l_t\right)_j$ sont respectivement la $jième$ coordonnée de x_i et l_t

Critère de classification L'application C_r

$$C_r : IP_k \times IL_k \longmapsto \mathbb{R}_+$$

$$(P, L) \longmapsto C_r\left(P, L\right) = \sum_{k=1}^{K} \sum_{\{i\} \in P_k} \left[\frac{1}{2^d} \sum_{t=1}^{2^d} \|x_i - l_k\| \right]$$

est le critére d'adéquation C_r s'écrit

$$C_r\left(P, L\right) = \sum_{k=1}^{K} \sum_{\{i\} \in P_k} \left[\frac{1}{2^d} \sum_{t=1}^{2^d} \left[\left(\sum_{j=1}^{d} \left(x_{ij} - l_{tj}^k \right)^2 \right)^{1/2} \right] \right]$$

Ce critére exprime l'adéquation entre les individus et les noyaux des classes. Dans le cas où les erreurs sont cumulées au niveau des noyaux. Les observations des individus sont données sans erreurs.

La fonction d'affectation f et la fonction de représentation g sont données par

$$f : IL_k \longmapsto IP_k$$

$$L \longmapsto f\left(L\right) = P$$

$$P = \{P_1, \ldots, \ P_k\} \in IP_k$$

et

$$P_t = \{\omega \in \Omega / \ D(\omega, \ l_t) \leq D(\omega, \ l_m) \ \text{ et } \ t \prec m \ \text{ s'il y a égalité } \}$$

Proposition 23

f vérifie

$$\underset{P \in IP_k}{Min} \quad C_r(P, \ L) = C_r(f(L), \ L)$$

$$g : IP_k \longmapsto IL_k$$

$$P \longmapsto g(P) = L^*$$

$$L^* = \{l_1, \ldots, \ l_k\} \in IL_k$$

avec

$$l_t = \frac{1}{|P_t|} \left(\frac{1}{2^d}\right) \sum_{\{i\} \in P_k} (x_i),$$

Proposition 24

g vérifie

$$\underset{L \in IL_K}{Min} \left[C_r(P, \ L)\right] = C_r(P, \ g(P) = L^*)$$

En effet, sans restreindre la généralité, étudions le cas où $d = 2$

On a : Pour tout $i = 1, \ldots, n$

$\omega_i \to R_i = \Im_i^1 \times \Im_i^2$ où $\Im_i^1 = \left[\mu_{i1} - \frac{\sigma_{i1}}{2}; \mu_{i1} + \frac{\sigma_{i1}}{2}\right]; \Im_i^2 = \left[\mu_{i2} - \frac{\sigma_{i2}}{2}; \mu_{i2} + \frac{\sigma_{i2}}{2}\right]$

Posant

$$\begin{cases} \mu_{i1} = x_i; \sigma_{i1} = a_i \\ \mu_{i2} = y_i; \sigma_{i2} = b_i \end{cases}$$

Les 4 sommets correspondants sont

$$S(i,1) = S_{i1} = \begin{pmatrix} x_i - a_i \\ y_i - b_i \end{pmatrix}; S(i,2) = S_{i2} = \begin{pmatrix} x_i - a_i \\ y_i + b_i \end{pmatrix}$$

$$S(i,3) = S_{i3} = \begin{pmatrix} x_i + a_i \\ y_i + b_i \end{pmatrix}; S(i,4) = S_{i4} = \begin{pmatrix} x_i + a_i \\ y_i - b_i \end{pmatrix}$$

Pour $k = 1, K; L_k \in L \Rightarrow L_k \in \mathbb{R}^2 \Rightarrow L_k = \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}$

Le critère à optimiser s'écrit

$$Cr\left(P,L\right)=\sum_{k=1}^{K}\left[\sum_{\omega_i\in P_k}\left[\frac{1}{2^d}\sum_{t=1}^{2^d}\left\|S\left(i,t\right)-L_k\right\|\right]\right]$$

On suppose que les objets de la classe P_k sont numérotés par $\{1,\ldots,n_k\}$

$$\text{et}\quad\begin{array}{l}x^{(k)}=x\\y^{(k)}=y\end{array}$$

On a

$$\left\|S\left(i,1\right)-L_k\right\|^2=\left(\left(x_i-a_i\right)-x\right)^2+\left(\left(y_i-b_i\right)-y\right)^2$$

$$\left\|S\left(i,2\right)-L_k\right\|^2=\left(\left(x_i-a_i\right)-x\right)^2+\left(\left(y_i+b_i\right)-y\right)^2$$

$$\left\|S\left(i,3\right)-L_k\right\|^2=\left(\left(x_i+a_i\right)-x\right)^2+\left(\left(y_i+b_i\right)-y\right)^2$$

$$\left\|S\left(i,4\right)-L_k\right\|^2=\left(\left(x_i+a_i\right)-x\right)^2+\left(\left(y_i-b_i\right)-y\right)^2$$

$$\begin{aligned}\varphi\left(x,y\right) &= 2\sum_{i=1}^{n_l}\left[\left(\left(x_i-a_i\right)-x\right)^2+\left(\left(x_i+a_i\right)-x\right)^2\right]+\\ &\quad 2\sum_{i=1}^{n_l}\left[\left(\left(y_i-b\right)-y\right)^2+\left(\left(y_i+b\right)-y\right)^2\right]\end{aligned}$$

$$\Rightarrow\left\{\begin{array}{l}\frac{\partial\varphi(x,y)}{\partial x}=0\\\frac{\partial\varphi(x,y)}{\partial y}=0\end{array}\right.\Rightarrow x=\frac{1}{n_k}\sum_{i=1}^{n_l}\left(x_{i1}\right)\quad;\quad y=\frac{1}{n_k}\sum_{i=1}^{n_l}\left(x_{i2}\right)$$

Remarque 4
Les noyaux sont des vecteurs de $\mathbb{R}^d$.

4.2.2 Alternative 2: Les erreurs sont cumulées au niveau de la description des individus et les noyaux sont donnés sans erreurs

Dans ce cas ω_i est décrit par le pavé $R_i=\prod_{j=1}^{d}\Im_j^i$ avec

$$\Im_i^l\subset\mathbb{R}$$

Critére d'optimisation Le critére C_r s'écrit

$$C_r(P,L)=\sum_{k=1}^{K}\sum_{\{i\}\in P_k} D(R_i,l_k)$$

où

$$L=(l_1,\ldots,\ l_k)\in IL_k,\ l_k\in\ \mathbb{R}^d$$

et D est définie par

$$D:\Omega\times IL\longmapsto\mathbb{R}*$$

$$(\omega_i,L)\ \longmapsto\ D\ (\omega_i,L)=\frac{1}{2^d}\left(\sum_{t=1}^{2^d}\|S_i(t)-L\|\right)$$

$\|.\|$ est la norme euclidienne, $L,\ S_i(t)\in\mathbb{R}^d$. $S_i(t)$ est le sommet t du pavé R_i.
Le critére s'écrit

$$C_r(P,L)=\ \sum_{k=1}^{K}\sum_{\omega_i\in P_k}\left[\frac{1}{2^d}\sum_{l=1}^{2^d}\left(\sum_{j=1}^{d}\left(S_i(t,j)-(l_k)_j\right)^2\right)^{\frac{1}{2}}\right]$$

La fonction représentative g est donnée par

$$g(P)=L=(l_1,\ldots,\ l_k)$$

avec

$$(l_k)_j=\frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ij})$$

x_{ij} est la moyenne prise par l'individu ω_i pour la variable $j,\ n_k=|P_k|$.
Proposition 25
g vérifie

$$\underset{L\in IL_K}{Min}\quad C_r(P,\ L)=C_r(P,\ g(P))$$

La fonction d'affectation f est

$$f : IL_k \longmapsto IP_k$$

$$L = (l_1, \ldots,\ l_k) \longmapsto f(L) = P = \{P_1, \ldots,\ P_k\}$$

$$P_t = \{\omega \in \Omega / D(\omega,\ l_t) \leq D(\omega,\ l_m);\ \text{et } t \prec m,\ \text{si on a égalité}\}$$

Proposition 26

f vérifie

$$\underset{P \in IP_k}{Min} C_r(P, L) = C_r(f(L), L)$$

On montre que les noyaux ou représentants des classes, dans ce cas, sont des vecteurs de $\mathbb{R}^d$. Il n'y a donc pas 2^d sommets représentants chaque classe mais chaque noyau est représenté par un vecteur de $\mathbb{R}^d$ et

$$D(\omega_i, l) = \|X_i - l\|; l \in \mathbb{R}^d$$

4.2.3 Alternative 3: Les erreurs sont cumulées dans la description des individus et dans les noyaux des classes

Proposition 27

L'application Δ est définie par

$$\Delta : \Re_{V_d} \times \Re_{V_d} \longmapsto \mathbb{R}_+$$

$$(R_!, R_2) \longmapsto \Delta(R_1, R_2)$$

$$\Delta(R_1, R_2) = \frac{1}{2^d} \sum_{l=1}^{2^d} D(S(1,\ l), R_2) = \frac{1}{2^d} \sum_{l=1}^{2^d} \sum_{t=1}^{2^d} \delta(S(1,\ l), S(2,\ t))$$

$$= \frac{1}{2^d} \sum_{l=1}^{2^d} \sum_{t=1}^{2^d} \left[\left(\sum_{j=1}^{d} \left(S^j(1,\ l) - S^j(2,\ t) \right)^2 \right)^{1/2} \right]$$

$$S(1,\ l), S(2,\ t) \in \mathbb{R}^d,$$

δ est la distance euclidienne. Δ est un indice de distance entre pavés de $\mathbb{R}^d$.

Critére à optimiser

$$C_r\left(P,\ L\right)=\sum_{k=1}^{K}\sum_{\omega_i\in P_k}\left[\frac{1}{2^d}\sum_{l=1}^{2^d}\left(\sum_{j=1}^{d}\left(S_i\left(t,\ j\right)-N_k\left(t,\ j\right)\right)^2\right)^{\frac{1}{2}}\right]$$

La fonction de représentation g est

$$g:IP_k\longmapsto IL_k$$

$$P=\left(P_1,\ldots,P_k\right)\longmapsto g\left(P\right)=L=\left(l_1,\ldots,\ l_k\right)$$

Si

$$\omega_i\rightarrow\left[\left(x_{ij}\right)_{j=1,\,d}\ ;\ \left(\varepsilon_{ij}\right)_{j=1,\,d}\right],$$

on montre que le noyau l_k de la classe P_k est donné par

$$l_k\rightarrow\left[\left(\mu_{kj}\right)_{j=1,d}\ ;\ \left(\sigma_{kj}\right)_{j=1,d}\right]$$

avec

$$\mu_{kj}=\left(\mu_k\right)_j\ ,\ \sigma_{kj}=\left(\sigma_k\right)_j$$

et

$$\begin{cases}\left(\mu_k\right)_j=\frac{1}{n_k}\sum\limits_{i=1}^{n_l}\left(x_i\right)_j\\ \left(\sigma_k\right)_j=\frac{1}{n_k}\sum\limits_{i=1}^{n_l}\left(\varepsilon_{ij}\right)\end{cases}.$$

En effet, sans restreindre la généralité, étudions le cas où $d=2$
On a
Pour tout $i=1,\ldots,n$

$$\omega_i\rightarrow R_i=\Im_i^1\times\Im_i^2\quad\text{où }\Im_i^1=\left[\mu_{i1}-\tfrac{\sigma_{i1}}{2};\mu_{i1}+\tfrac{\sigma_{i1}}{2}\right];\Im_i^2=\left[\mu_{i2}-\tfrac{\sigma_{i2}}{2};\mu_{i2}+\tfrac{\sigma_{i2}}{2}\right]$$

Pour $k=1,K;L_k\in L\Rightarrow L_k\in\Re_{V_d}\Rightarrow L_k=I_k^1\times I_k^2$ avec

$$\left\{I_k^j=\left[x_j^{(k)}-a_s^{(k)};x_j^{(k)}+a_s^{(k)}\right];j=1,2\right\}$$

Les 4 sommets associés à chacun des pavés dans l'ordre sont:

$_i \downarrow$ $L_k \downarrow$

$$S(i,1) = S_{i1} = \begin{pmatrix} x_i - a_i \\ y_i - b_i \end{pmatrix} \quad ; \quad N(k,1) = \begin{pmatrix} x - a \\ y - b \end{pmatrix}$$

$$S(i,2) = S_{i2} = \begin{pmatrix} x_i - a_i \\ y_i + b_i \end{pmatrix} \quad ; \quad N(k,2) = \begin{pmatrix} x - a \\ y + b \end{pmatrix}$$

$$S(i,3) = S_{i3} = \begin{pmatrix} x_i + a_i \\ y_i + b_i \end{pmatrix} \quad ; \quad N(k,3) = \begin{pmatrix} x + a \\ y + b \end{pmatrix}$$

$$S(i,4) = S_{i4} = \begin{pmatrix} x_i + a_i \\ y_i - b_i \end{pmatrix} \quad ; \quad N(k,4) = \begin{pmatrix} x + a \\ y - b \end{pmatrix}$$

Le critère à optimiser s'écrit

$$Cr(P,L) = \sum_{k=1}^{K} \left[\sum_{\omega_i \in P_k} \left[\frac{1}{2^d} \sum_{t=1}^{2^d} \|S(i,t) - N(k,t)\| \right] \right]$$

On a

$$\|S(i,1) - N(k,1)\|^2 = ((x_i - a_i) - (x - a))^2 + ((y_i - b_i) - (y - b))^2$$

$$\|S(i,2) - N(k,2)\|^2 = ((x_i - a_i) - (x - a))^2 + ((y_i + b_i) - (y + b))^2$$

$$\|S(i,3) - N(k,3)\|^2 = ((x_i + a_i) - (x + a))^2 + ((y_i + b_i) - (y + b))^2$$

$$\|S(i,4) - N(k,4)\|^2 = ((x_i + a_i) - (x + a))^2 + ((y_i - b_i) - (y - b))^2$$

$$\varphi(a,b,x,y) = 2\sum_{i=1}^{n_l} \left[[(x_i - a_i) - (x - a)]^2 + [(x_i + a_i) - (x + a)]^2 \right] +$$
$$2\sum_{i=1}^{n_l} \left[[(y_i - a_i) - (y - a)]^2 + [(y_i + a_i) - (y + a)]^2 \right] \Rightarrow$$

$$\begin{cases} \frac{\partial \varphi(a,b,x,y)}{\partial x} = 0 \\ \frac{\partial \varphi(a,b,x,y)}{\partial y} = 0 \end{cases} \Rightarrow x = \frac{1}{n_k} \sum_{i=1}^{n_l} (x_i) \quad ; \quad y = \frac{1}{n_k} \sum_{i=1}^{n_l} (y_i)$$

$$\begin{cases} \frac{\partial \varphi(a,b,x,y)}{\partial a} = 0 \\ \frac{\partial \varphi(a,b,x,y)}{\partial b} = 0 \end{cases} \Rightarrow a = \frac{1}{n_k} \sum_{i=1}^{n_l} (a_i) \quad ; \quad b = \frac{1}{n_k} \sum_{i=1}^{n_l} (b_i)$$

La fonction d'affectation f est

$$f : IL_k \longmapsto IP_k$$

$$L \longmapsto f(L) = P$$

$$P = \{P_1, \ldots, \ P_k\} \in IP_k$$

est

$$P_t = \{\omega \in \Omega / \ D(\omega, \ l_t) \leq D(\omega, \ l_m) ; \text{ et } t \prec m \ \text{ si on a égalité } \}$$

4.3 Objets décrits par des ellipsoides de $\mathbb{R}^d$

4.3.1 Critère et problème d'optimisation

Nous désirons regrouper les n individus en k classes où chaque individu sera décrite par son élément représentatif. L'homogénéité des classes est mesurée par

$$H_k = \sum_{\{i\} \in P_k} d_{ma}(W_i, L_k),$$

où L_k est l'élément représentatif du noyau de la classe C_k , $W_i \in \mathbb{N} \times \mathbb{R}^d \times M_d(\mathbb{R})$ et d_{ma} est la distance de Mahannalobis ponderée. H_k mesure la dispersion des individus à l'intérieur de la classe. H_k est aussi l'inertie des objets de la classe autour de son noyau.

Soit P_K l'ensemble des partitions à k classes

$$P \in P_K \Rightarrow P = \{P_1, \ldots, P_K\}$$

est une partition à K classes, L_K l'ensemble des K représentations

$$(L \in L_K) \Rightarrow L = \{L_1, \ldots, L_k\}$$

où

$$L_t = \in \mathbb{N} \times \mathbb{R}^d \times M_d(\mathbb{R}).$$

On défini l'application

$$C_r : P_K \times L_K \to \mathbb{R}_+$$

$$(P, L) \quad \to C_r(P, L) = \sum_{k=1}^{K} \sum_{\{i\} \in P_k} d_{ma}(W_i, L_k)$$

$$d_{ma}\left(W_i, L_k\right) = \frac{N_i \left\|\mu_i - \widetilde{\mu}_k\right\|_{\Sigma_i^{-1}} + \widetilde{N}_k \left\|\mu_i - \mu_l\right\|_{\widetilde{\Sigma}_k^{-1}}}{N_i + \widetilde{N}_k}$$

$$W_i; c_k \in \mathbb{N} \times \mathbb{R}^d \times M_d\left(\mathbb{R}\right)$$

Le critère s'écrit

$$C_r\left(P, L\right) = \sum_{k=1}^{K} \sum_{\{i\} \in P_k} d_{ma}\left(W_i, L_k\right)$$

Ce critère exprime l'adéquation entre les éléments représentatifs des individus dans les classes avec les éléments représentatifs des noyaux des classes. Cr s'ecrit

$$C_r\left(P, L\right) = \sum_{k=1}^{K} \sum_{\{i\} \in P_k} d_{ma}\left(W_i, L_k\right)$$

On recherche le couple (P^*, L^*) qui réalise

$$\underset{\substack{P \in P_K \\ L \in L_K}}{Min} C_r\left(P, L\right)$$

Les algorithmes utilisés pour résoudre ce type de problème sont du type $k - means$. Ces algorithmes sont basés sur la définition de la fonction de représentation g et de la fonction d'affectation f qui seront utilisées simultanément pour faire decroître le critère.

4.3.2 Caractérisation d'une classe d'objets(résultat trés important)

Il s'agit de rechercher le noyau représentant chacune des classes générée par cet algorithme

Si l'on note par $i = 1, \ldots, k$ les k objets dans la classe C.

soit φ l'application definie par

$$\mathbb{R}^d \times M_{(d)} \times \mathbb{N} \xrightarrow{\varphi} \mathbb{R}_+$$

$$(\mu, S, N) \longrightarrow \varphi\left(\mu, S, N\right) = \sum_{i=1}^{K} \left[\frac{N_i}{N_i + N} \left\|\mu_i - x\right\|^2_{S_i^{-1}} + \frac{N}{N_i + N} \left\|\mu_i - x\right\|^2_{S^{-1}}\right]$$

Il s'agit de chercher μ^*, S^*, N^* qui realise le minimum de φ . Résoudre le problème d'optimisation

$$\underset{(x,S,L)\ \in\mathbb{R}^d\times M_{(d)}\times\mathbb{N}}{Min}\ \varphi(\mu,S,N)$$

Sans restreindre la généralité, on s'intéresse au cas où $p = N_i\ \ \forall i$. On cherche à résoudre le problème d'optimisation suivant

$$\underset{(x,S,p)\ \in\mathbb{R}^d\times M_{(d)}}{Min}\ \varphi(x,S)$$

on pose $\mu = x : \mu_i = x_i$
Condition nécéssaire

$$\begin{cases} \frac{\partial\varphi}{\partial x} = 0 & (1) \\ \frac{\partial\varphi}{\partial S} = 0 & (2) \end{cases}$$

$$\frac{\partial\varphi}{\partial x} = 0\,(1) \Longleftrightarrow \sum_{i=1}^{K}\left[(x_i - x)'\left(S_i^{-1} + S^{-1}\right)\right] = 0\ (1) \Leftrightarrow \sum_{i=1}^{K}\left[(x_i - x)'\,S_i^{-1}\right] + \sum_{i=1}^{K}\left[(x_i - x)'\,S^{-1}\right] = 0$$

En ce qui concerne la seconde condition, on remarque que la première expression de φ ne dépend pas de S. Soit

$$\Phi(S) = \sum_{i=1}^{K}\|(x_i - x)\|_{S^{-1}}^2$$

On a

$$\Phi(S+H) - \Phi(S) = \sum_{i=1}^{K}\left[(x_i - x)'\,(S+H)^{-1}\,(x_i - x)\right] - \sum_{i=1}^{K}\left[(x_i - x)'\,(S)^{-1}\,(x_i - x)\right]$$

or

$$(S+H)^{-1} = \left[S\left(I + S^{-1}H\right)\right]^{-1} = \left(I + S^{-1}H\right)^{-1}S^{-1}$$

Le développement de $(I + S^{-1}H)^{-1}$donne

$$\left(I + S^{-1}H\right)^{-1} = I - S^{-1}H + \ldots.. \Rightarrow\ (S+H)^{-1} = \left[I - S^{-1}H + \ldots..\right]S^{-1} \Rightarrow$$

$$\Phi(S+H)-\Phi(S)=\sum_{i=1}^{K}\left[(x_i-x)'\left(S^{-1}.H.S^{-1}\right)(x_i-x)\right]+.....$$

$$=\sum_{i=1}^{K}\left[\left[S^{-1}(x_i-x)\right]'(H)\left[S(x_i-x)\right]\right]+...$$

$$=\sum_{i=1}^{K}\left\|S^{-1}(x_i-x)\right\|_H^2+....$$

$$\Rightarrow\frac{\partial\varphi}{\partial S}=0\ \ (2)\Rightarrow\sum_{i=1}^{K}\left\|S^{-1}(x_i-x)\right\|_H^2=0\Rightarrow S^{-1}(x_i-x)=0\forall i\Rightarrow S^{-1}=0$$

car $x_i\neq x$

$$(1)\Longleftrightarrow\sum_{i=1}^{K}\left[(x_i-x)'\left(S_i^{-1}\right)\right]=0\ \Rightarrow(x)'\sum_{i=1}^{K}\left(S_i^{-1}\right)=\left(\sum_{i=1}^{K}(x_i)'\left(S_i^{-1}\right)\right)\Rightarrow$$

Si

$$\Gamma_k=\sum_{i=1}^{K}\left(S_i^{-1}\right)$$

$$\Rightarrow x=(\Gamma_k)^{-1}\sum_{i=1}^{K}\left(S_i^{-1}\right)(x_i)=\widetilde{\mu}_k$$

Ce qui montre qu'une fois les clusters obtenues, la caractérisation n'est pas entachées d'erreurs et est donnée de manière exacte Ce qui parait tout à fait dans l'ordre des choses.

La distance entre un individu ω_i donné décrit par le couple(ν_i,Σ_i) et la classe C_k caracterisée par son noyau L_t est donnée par

$$d_{ma}(\omega_i,L_k)=\|\mu_i-\widetilde{\mu}_k\|_{\Sigma_i^{-1}}$$

4.3.3 La fonction de représentation

$$g:P_k\rightarrow L_k$$

$$P=\{P_1,\ldots,P_k\}\rightarrow g(P)=\{L_1,\ldots,L_k\}$$

g doit vérifier

$$\underset{L\in L_k}{MinCr}\left(P,L\right)=Cr\left(P,g\left(P\right)\right)\quad(*)$$

Proposition 28
La fonction g définie par

$$P=\{P_1,\ldots,P_k\}\rightarrow g\left(P\right)=\{L_1,\ldots,L_k\}$$

$$\text{où } L_k=\left(\Gamma_k\right)^{-1}\sum_{i=1}^{K}\left(S_i^{-1}\right)\left(x_i\right)\quad\text{avec } \Gamma_k=\sum_{i=1}^{K}\left(S_i^{-1}\right)$$

verifie$(*)$

4.3.4 La fonction d'affectation

$$f:L_k\rightarrow P_k$$
$$L=\{L_1,\ldots,L_k\}\rightarrow f\left(L\right)=\{P_1,\ldots,P_k\}$$

f doit vérifier

$$\underset{P\in P_k}{MinCr}\left(P,L\right)=Cr\left(f\left(L\right),L\right)$$

Proposition 29
Ce minimum est obtenu pour

$$P_l=\{\omega_i\in\Omega\quad\text{tels que}\quad d_{ma}\left(\omega_i,L_l\right)\preceq d_{ma}\left(\omega_i,L_t\right);\forall t\neq l\quad\text{et}\quad l\prec t\text{ en cas d'égalité}\}$$

4.3.5 L'algorithme

On choisi au hasard ou par d'autres artifices $L^{(0)}$et on utilise alternativement les fonctions f et g pour faire décroître le critère L'algorithme de déroule comme suit

$$L^{(0)}\xrightarrow{f}P^{(1)}\xrightarrow{g}L^{(1)}\xrightarrow{f}P^{(2)}.....L^{(*)}\xrightarrow{f}P^{(*)}$$

L'algorithme s'arrête dés que la partition obtenue ne change plus. Nous construisons ainsi 2 suites U_n et V_n telles que

$$\begin{cases}U_n=Cr\left(P^{(n)},L^{(n)}\right)\\ V_n=\left(P^{(n)},L^{(n)}\right)\end{cases}$$

Proposition 30
La suite U_n converge en décroissant et la suite V_n est stationnaire à partir d'un certain rang donc convergente.

4.3.6 Exemples

On désire classifier les 6 objets en 2 classes. Chaque objet est décrit par 3 variables quantitatives V_1, V_2 et V_3 . Les données en entrée se présentent comme suit

$$\omega_1 \mapsto T_1 = \begin{bmatrix} 3 & 5 & 6 \\ 7 & 9 & 10 \\ -5 & 2 & 0 \\ 3 & 1 & -1 \\ 5 & 4 & 2 \end{bmatrix} ; T_2 = \begin{bmatrix} 2 & -1 & 0 \\ 3 & 5 & 11 \\ -2 & 3 & 4 \\ 6 & 7 & 8 \end{bmatrix} ; T_3 = \begin{bmatrix} 4.1 & 2.6 & -1 \\ 2.4 & 0.5 & 4 \\ 5 & 6 & 7 \end{bmatrix}$$

$$T_4 = \begin{bmatrix} 3 & 3.6 & 4 \\ -1 & 17 & 0 \\ 10 & 3 & 12 \\ 13 & 0 & 11 \end{bmatrix} ; T_5 = \begin{bmatrix} 4 & -1 & 2 \\ 3 & 2 & 6 \\ 2 & 3 & 7 \\ 3 & 5 & -11 \\ 1 & 0.2 & 3 \\ 0.4 & 7 & 2 \end{bmatrix} ; T_6 = \begin{bmatrix} 2 & 7 & 2.5 \\ -1 & 3 & 1.5 \\ 2 & 5 & 0 \\ 4 & 0.5 & 14 \\ 6 & 0.7 & 3 \end{bmatrix}$$

- On remarque que les tables $T_1, T_2, ..., T_6$ n'ont pas la même dimension
- .Les moyennes, les écart-types et les intervalles de confiance pour chaque variable et pour chaque individu sont donnés dans les tableaux suivants

$$\omega_1 \rightarrow \begin{bmatrix} mean\,(\mu) & std.deviation\,(\sigma) \\ 2.6 & 4.56070 \\ 4.2 & 3.11448 \\ 3.4 & 4.56070 \end{bmatrix} ; \omega_1 \rightarrow \begin{bmatrix} \mu - \frac{\sigma}{2} & \mu + \frac{\sigma}{2} \\ 0.31965 & 4.88034 \\ 2.64275 & 5.75724 \\ 1.11965 & 5.68035 \end{bmatrix}$$

$$\omega_2 \rightarrow \begin{bmatrix} mean\,(\mu) & std.deviation\,(\sigma) \\ 2.25 & 10.90489 \\ 3.50 & 10.08298 \\ 5.75 & 4.78714 \end{bmatrix} ; \omega_2 \rightarrow \begin{bmatrix} \mu - \frac{\sigma}{2} & \mu + \frac{\sigma}{2} \\ -3.20245 & 7.70244 \\ -1.54149 & 8.54149 \\ 3.35643 & 8.14357 \end{bmatrix}$$

$$\omega_3 \rightarrow \begin{bmatrix} mean\,(\mu) & std.deviation\,(\sigma) \\ 3.8333 & 1.32035 \\ 3.0333 & 2.77549 \\ 3.3333 & 4.04145 \end{bmatrix} ; \omega_3 \rightarrow \begin{bmatrix} \mu - \frac{\sigma}{2} & \mu + \frac{\sigma}{2} \\ 0.317283 & 4.49318 \\ 1.64526 & 4.42075 \\ 1.31228 & 5.35373 \end{bmatrix}$$

$$\omega_4 \rightarrow \begin{bmatrix} mean\,(\mu) & std.deviation\,(\sigma) \\ 6.25 & 6.39661 \\ 5.9 & 7.56571 \\ 6.75 & 5.73730 \end{bmatrix} ; \omega_4 \mapsto \begin{bmatrix} \mu - \frac{\sigma}{2} & \mu + \frac{\sigma}{2} \\ 3.05170 & 9.44831 \\ 2.11715 & 9.68286 \\ 3.88135 & 9.61865 \end{bmatrix}$$

$$\omega_5 \longmapsto \begin{bmatrix} mean\,(\mu) & std.deviation\,(\sigma) \\ 2.2333 & 1.35892 \\ 2.7 & 2.97658 \\ 1.5 & 6.47302 \end{bmatrix} ; \omega_5 \longmapsto \begin{bmatrix} \mu - \frac{\sigma}{2} & \mu + \frac{\sigma}{2} \\ 1.55384 & 2.91276 \\ 1.21171 & 4.18829 \\ 1.73651 & 4.73651 \end{bmatrix}$$

$$\omega_6 \longmapsto \begin{bmatrix} mean\,(\mu) & std.deviation\,(\sigma) \\ 2.6 & 2.60768 \\ 3.24 & 2.79517 \\ 4.2 & 5.59687 \end{bmatrix} ; \omega_6 \longmapsto \begin{bmatrix} \mu - \frac{\sigma}{2} & \mu + \frac{\sigma}{2} \\ 1.29616 & 3.90384 \\ 1.84242 & 4.63759 \\ 1.40157 & 6.99844 \end{bmatrix}$$

• Les sommets associés pour chaque individu s'écrivent

Pour $i = 1,\ n;\ S\left(i,\ l\right) \in \mathbb{R}^d$

$$\left[S\left(i,\ l\right)\right]_j = \mu_i^j + \varepsilon.\sigma$$

$$; \text{ avec } \varepsilon = \pm 1$$

$$\begin{array}{cccccccc}
\overbrace{\mu_i^1 - \sigma_i^1}^{S(i,1)} & \overbrace{\mu_i^1 + \sigma_i^1}^{S(i,2)} & \overbrace{\mu_i^1 - \sigma_i^1}^{S(i,3)} & \overbrace{\mu_i^1 + \sigma_i^1}^{S(i,4)} & \overbrace{\mu_i^1 - \sigma_i^1}^{S(i,5)} & \overbrace{\mu_i^1 + \sigma_i^1}^{S(i,6)} & \overbrace{\mu_i^1 - \sigma_i^1}^{S(i,7)} & \overbrace{\mu_i^1 + \sigma_i^1}^{S(i,8)} \\
\mu_i^2 - \sigma_i^2 & \mu_i^2 - \sigma_i^2 & \mu_i^2 + \sigma_i^2 & \mu_i^2 + \sigma_i^2 & \mu_i^2 - \sigma_i^2 & \mu_i^2 - \sigma_i^2 & \mu_i^2 + \sigma_i^2 & \mu_i^2 + \sigma_i^2 \\
\mu_i^3 - \sigma_i^3 & \mu_i^3 - \sigma_i^3 & \mu_i^3 - \sigma_i^3 & \mu_i^3 - \sigma_i^3 & \mu_i^3 + \sigma_i^3 & \mu_i^3 + \sigma_i^3 & \mu_i^1 + \sigma_i^3 & \mu_i^3 + \sigma_i^3
\end{array}$$

• A ce niveau, 2 formes de la description des objets peuvent être considérées

-Chaque objet est décrit par un vecteur de $\mathbb{R}^d$ qui contient les valeurs moyennes prises par les variables qui les décrivent .

$$\omega_i \mapsto (\mu_i)' = \left[\mu_i^1, \ldots, \mu_i^d\right]$$

où

$$\mu_i^j = \frac{1}{L_i} \sum_{l=1}^{L_i} x_{il}^j.$$

Les données en entrée sont regroupées dans le tableau X.

$$X = \begin{bmatrix}
\mu_i^1 & \mu_i^2 & \mu_i^3 \\
2.6 & 4.2 & 3.4 \\
2.25 & 3.5 & 5.75 \\
3.83 & 3.03 & 3.33 \\
6.25 & 5.9 & 6.75 \\
2,23 & 2.7 & 1.5 \\
2.6 & 3.24 & 4.2
\end{bmatrix}$$

L'algorithme des $k - means$ déroulé pour classifier les 6 individus en 2 classes donne

Centres finaux des classes

$$\begin{bmatrix}
 & class1 & class2 \\
v11 & 2.7 & 6.25 \\
v22 & 3.3 & 5.90 \\
v33 & 3,64 & 6.75
\end{bmatrix}$$

$$\begin{bmatrix}
\text{case number} & \text{cluster} & \text{distance} \\
\omega_1 & 1 & ,90305 \\
\omega_2 & 1 & 2,16773 \\
\omega_3 & 1 & 1,20818 \\
\omega_4 & 2 & ,00000 \\
\omega_5 & 1 & 2,27794 \\
\omega_6 & 1 & ,58051
\end{bmatrix}$$

- Chaque objet est décrit par les 2^d $(d = 3)$ sommets associés.

Les coordonnées des sommets associés aux 6 individus sont
Pour $\omega_1 \downarrow$

,31965	2,64276	1,11965	4,88035	2,64276	1,11965	,31965
5,75724	1,11965	4,88035	5,75724	1,11965	4,88035	2,64276
5,68035	,31965	5,75724	5,68035	,31965	2,64276	5,68035

4,88035 5,75724 5,68035 **1** **4,20132**

Pour $\omega_2 \downarrow$

-3,20245	-1,54149	3,35643	7,70245	-1,54149	3,35643	-3,20245
8,54149	3,35643	7,70245	8,54149	3,35643	7,70245	-1,54149
8,14357	-3,20245	7,70245	8,14357	-3,20245	-1,54149	8,14357

7,70245 5,20783 8,14357 **2** **8,03977**

Pour $\omega_3 \downarrow$

3,17283	1,64526	1,31228	4,49318	1,64526	1,31228	3,17283
4,42075	1,31228	4,49318	4,42075	1,31228	4,49318	1,64526
5,35373	3,17283	4,42075	5,35373	3,17283	1,64526	5,35373

4,49318 4,42075 5,35373 **1** **3,59757**

Pour $\omega_4 \downarrow$

3,05170	2,11715	3,88135	9,44831	2,11715	3,88135	3,05170
9,68286	3,88135	9,44831	9,68286	3,88135	9,44831	2,11715
9,61865	3,05170	9,68286	9,61865	3,05170	2,11715	9,61865

9,44831 9, 68286 9,61865 **2** **8,03977**

Pour $\omega_5 \downarrow$

1,55384	1,21171	1,73651	2,91276	1,21171	1,73651	1,55384
4,18829	1,73651	2,91276	4,18829	1,73651	2,91276	1,21171
4,73651	1,55384	4,18829	4,73651	1,55384	1,21171	4,73651

2,91276 4,18829 4,73651 **1** **3,32424**

Pour $\omega_6 \downarrow$

1,29616 1,84242 1,40157 3,90384 1,84242 1,40157 1,29616 4,63759 1,40157 1,29616
4,63759 1,40157 3,90384 1,84242 6,99844 1,29616 4,63759 6,99844 1,29616

1,84242 6,99844 3,90384 4,63759 6,99844 **1** **3,41777**

Chapitre VI
Conclusion générale

- Dans cette thése, on s'est intéressé à l'analyse des éléments constitutifs d'une structuture de juxtaposition de tableaux multiples qui dans certains cas des objets matriciels.

- Nous avons développé diffèrentes approches adaptées aux types de structure que nous avons liées à des formalismes spécifiques qui permettent d'appréhender les éléments de la structure des données.

- Nous avons montré comment et dans quelles conditions les techniques factorielles peuvent y être adaptées quand il s'agit de décrire les éléments constitutifs de la structure. Nous avons étudié le cas d'une stucuture de juxtaposition de tableaux de mesure et le cas où chaque élément de la strucuture est décrit par un tableau de données quantitatives.

La démarche consiste à résumer chaque élément de la structure par sa représentation sur des sous -espaces de faible dimension . Ces sous espaces sont obtenus à partir des déroulements d'ACP.

- Nous avons discuté les avantages et les inconvénients. Cette approche qui a fait l'objet des travaux de Bouroche[10] pose un problème fondamental qui est: # *La recherche d'un sous espace commun ou compromis à tous les sous espaces construits*# . Il propose que ces choix se fassent par des méthodes heuristiques suivants des critères de minimisation d'inertie mais le problème de fixer la dimension des sous espaces n'a pas été abordé. Ce sous espace affine ou vectoriel doit permettre d'avoir une vision globale des éléments constitutifs pour une étude plus fine.

- Ce problème intéresse d'autres disciplines de la mathématique et en particulier la géometrie différentielle, nous citons les travaux de P.Orlick et H Terao[74][1992] . Ce probleme reste à mon sens assez ouvert.

- Nous avons contourné ce problème en proposant les 2 approches qui constituent l'ossature de cette thèse et qui conduisent à un niveau de connaissance en sortie équivanlent à celui introduit en entrée (condition nécéssaire d'une bonne analyse, Diday [1998]).

- Cette remarque se justifie aisément dans l'étude réalisée dans le chapitre 3 comme le montre le tableau suivant

Entrée
objets décrits par des courbes
représentées par des tableaux
↓
Sortie
Noyaux décrits par des courbes de charges
construits à partir de tableaux représentants les classes

- De plus, dans les 2 approches, nous tenons compte de la variabilité des observations de chaque individu pour les différentes variables qui les décrivent.

- La seconde approche nous parait plus générale tant au point de vue méthodologique qu'au point de vue du formalisme utilisé. Nous avons manipulé des variables aléatoires qui sont des concepts sous tendus par des notions de variablilité et d'extension. L'approche peut s'adapter facilement à des données mesurables ou qualitatives car elle s'appui essentiellement sur des calculs de fréquences relatives ou des probabilités d'apparition de telle ou telle modalité . Elle permet de s'étendre à tous les types de strucuture de données

- . Le fait de prendre les données sans manipulation préalable et mise en forme rend les résultats obtenus plus proches de la réalité que l'on désire appréhender donc conduit à des résultats plus crédibles et facilement interprétables.

Cette approche est basée sur la définition de Shannon Weiner du concept de système physique. La notion de système physique a longtemps fait l'objet de discussions entre scientifiques. Depuis que Carnot a ébauché les concepts fondamentaux de la thermodynamique et défini la caractéristique physique qui mesure le degré du désordre et qu'il a baptisé entropie. Ces discussions continuent avec l'apparition de la notion d'objets probabilistes [32](Diday,...) et de sous ensembles flous (Kauffman,Bezdek,...)

- Nous pensons que le choix de la définition de Shannon du système physique a permis de manière naturelle d'utiliser et d'expliquer l'aspect pratique de la distance de Kullback-Leibler comme un indice de distance entre les éléments constitutifs de structures complexes.

Ce travail ouvre la voie à plusieurs domaines d'applications pratiques et particulièrement en médecine.

Enfin, ce travail ouvre le chemin à plusieurs axes de recherche et d'investigations et les extensions sont nombreuses. Parmis les éclairages possibles, on peut citer

- Analyse d'objets décrits par des distributions paramétrées connues (normales, multinomiales,....)

- Etudes algébriques et propiétrés d'objets aléatoires(complétude,...)

Chapitre VII
Bibliographie

Bibliographie

[1] T. Adamson (1996) , *Data strucuture and algorithms, a first cours*e. Springer Verlag.

[2] A. Anderberg (1973) , *Cluster analysis for applications* . Academic Press

L'un des premiers ouvrages spécialisé en langue anglaise qui contient l'historique des développements qu'a connu la classification automatique.

[3] T.W. Anderson (1974) , *An introduction to multivariate analysis* John Wiley and sons, New York.

Le premier livre qui préconise une modélisation des données en entrée avant d'entamer l'analyse. C'est un livre de référence qui fait le lien entre l'approche classification de l'école "française " et l'approche anglo-saxone dite "clustering

[4] Arabie and Hubert (1992) , *Combinatorial data analysis.* Annual review of psychology. p: 168 – 203.

[5] G.H. Ball,D.J Hall (1967) , *ISODATA,a novel method of data analysis and pattern classification*, Stanford research Inst. California,

[6] J.D.Banfield and B.Basford (1989) , *Model based gaussian and non gaussien clustering.* Technical report n° **186** Dept. of Statistical, GN 22, University of Washington.

[7] JP.Benzecri (1973) , *L'analyse des données* tome 1 & 2 .Dunod, Paris.

Ouvrage spécialisé en langue francaise contenant d'importantes considérations historiques du sujet et de rigoureux développements formels sur la notion de classification automatique.

[8] J.C. Bezdek (1975) , *Mathematical models for systematic and taxinomy* .Freidman,San Francisco,USA.

[9] HH. Bock (1975) , *On significiance tests in cluster analysis.* Journal of Classification, p: 77–108, Springer Verlag, N.Y.

[10] JM. Bouroche (1975) , *Analyse de données ternaires; Double analyse en composantes principales(DACP).* Thèse de l'Université de Paris VI

[11] X.Bry (1997) , *Analyses factorielles multiples.* Economica, Paris.

[12] C.Burt (1950) ,*The factorial analysis of qualitative data.* British J of statistical phsychology **3, 3**, p: 166–185.

Burt a été incontestablement l'inventeur du point de vue méthodologique de l'analyse factorielle multiple. Dans cet article ,il préconise le calcul du tableau qui

regroupe les croisements deux à deux de toutes les variables associées qui interviennent. Ce tableau qui porte son nom est une juxtaposition de tables de contingence. Après une normalisation de celui-ci ,il devient un tableau disjonctif complet sur lequel opère l'analyse factorielle multiple (AFM)

[13] T.Calcom (1981) , *Analyse conjointe de matrices de données.*Thèse de l'Université de Grenoble, France

[14] A.Carlier (1985) , A*pplication de l'analyse des évolutions et de l'analyse intra-période.* Revue de Statistique et Analyse des Données **10-1,** p: 27–53.

[15] A. Carlier, C. Lavit, M. Pages, M.O. Pernin, J.C.Turlot, (1988) , *A comparative review of methods which handles a set of indexed data tables*, Multiway data analysis, p 85 -102

[16] J.D.Carroll and J.J.Chang (1970) , *Analysis of individual differrences in multidimensionnal scaling via N way generalisation of "Eckart-Young " decomposition.* Psychometrika, **49- 3**, p: 403–423.

[17] P.Cazes (1980) , *Analyse de certains tableaux rectangulaires décomposés en blocs.* Les Cahiers de l'Analyse des Données , **5**, p: 387– 403.

[18] P.Cazes (1986) , *Une généralisation des correspondances multiples et des correspondances hiérarchiques.* Cahiers du BURO, **46-47**, Université de Pierre et Marie Curie, Paris , France

[19] P.Cazes, A. Chouakria, E.Diday, Y Scheketman (1997) , *Analyse en composantes principales d' objets décrits par des intervalles.* Revue de Statistiques appliquées. Vol **45**, n°3.

[20] G.Celeux (1988) , *Classification et modèles:* RSA XXXVI, **4**, p: 43–58.

[21] J.L. Chandon and S. Pison (1981) , *Analyse typologique: théorie et applications.* Masson, Paris.

[22] L.Csizard (1967) , *Information type distance measures and indirect observation.* Stud. Scien Math., Hungar, Vol. **2,** p :299–318.

[23] R.M. Cormack and C Hubert (2000) , *Hierarchical dependency models for multivariate survival data with censoring* . Life Data, **6**, pp: 299–320.

[24] R.M. Cormack (1971) , *A review of classification.*J.of Royal statist. Society, serie A, **134**, part.3, pp: 321– 367.

L'un des premiers articles faisant état de la synthèse des techniques de classification.

[25] A.D. Cordon (1987) , *A review of hierarchical classification.* J.of Royal Statist. Society, serie A, **150**, part.2 , p: 119–137.

Complément de l'article de Cormack par l'ajout des travaux les plus récents sur la classification hiérarchique

[26] L.Cuttman (1941) , *The quantificationof a class of attributes*, p:319-348, Soc. Sc. Res. Council,NY

[27] N.Day (1969) , *Estimating the components of mixture of normal distribution.* Biometrika, **56**, p: 463–474.

[28] F.Dazy and J.F Le Barzik (1997) , *L'analyse des données évolutives :méthodes et applications.* Edition technip, Paris.

C'est un ouvrage consacré aux techniques factorielles adaptées à l'analyse d'une juxtaposition de tableaux indexés par le temps. Les applications proposées sont très intéressantes et touchent le domaine de la vie sociale française '(étude de l'évolution de la délinquance et de l'évolution des votes en France). Les méthodes statis et dacp sont largement développées

[29] F.A.T. De Carvalho (1994) , *Proximity coefficients between boolean objects. New approachs in classication and data analysis*, Diday (Eds). Springer Verlag, p: 387–394.

[30] E.Diday (1971) , *La méthode des nuées dynamiques.* RSA **19**, N° 19, p: 19–34.

La méthode des nuées dynamiques généralise la technique proposée par Ball & Hall[65] et permet de modifier le choix initial des centres de classes. Le problème de classification se pose en terme d'optimisation d'un critère dont l'algorithme des nuées dynamiques permet de donner une solution locale dépendante du choix initial. Diday a mis au point une procédure de groupements stables (formes fortes) qui permet de remédier au moins partiellement au principal inconvénient des algorithmes proposés à savoir la dépendance de la solution obtenue en fonction du choix initial.

[31] E.Diday and all (1980) ,*Optimisation en classification automatique* tome 1 & 2. INRIA, Paris.

Rassemble plusieurs travaux sur l'application de l'algorithme des nuées dynamiques et des développements de la classification par les techniques des centres mobiles. Les travaux ont fait l'objet de thèses et diverses applications sont exposées.

[32] E Diday (1995) , *Probabilist, possibilist and belief objects for knowledje analysis* Annals of operationnel research, **55**, p: 227–276.

[33] P.Ducemetière (1970) , *Les méthodes numériques de classification automatique.* RSA,Vol **48** n° 4.

[34] R.O.Duda and P.E Hart (1973) , *Pattern classification and science analysis.* J.Wiley, N.Y.

[35] C.Eckart, G Young (1936) : The approximation of one matrix by another of lower rank, *Psychometrika*,**1**, p 211- 218

[36] B Escofier and J. Pages (1988) , *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation.* Dunod, Paris.

[37] B.Escofier (1989) , *Quelques indices pour comparer des tableaux de contingence* Revue de Statistique et d'Analyse des Données, **1**, p: 39–51.

[38] B.Everit and D Hand (1981) , *Finite mixture distributions.* Chapman and Hall, U.S.A.

[39] RA. Fisher (1936) ,*The use of multiple measurements in taxinomic problems.* Ann Of eugenics, **7**, p: 179 –188.

Dans cet article Fisher a montré que dans le cas où les individus ne seraient pas décrits par un seul paramètre, le calcul d'une partition exacte est possible car il existe une relation d'ordre entre les individus ce qui limite considérablement l'éventail des partitions

[40] E.W Forgy,(1965) ,Cluster analysis of multivariate data,Biometrics, **21**, 3, p 728

[41] T.Foucart (1983) , *Une nouvelle approche de la méthode STATIS* . RSA, XX1, **2**, p: 61–75.

[42] T.Foucart (1984) , *Analyse factorielle de tableaux multiples.*Masson,France

[43] C.Gowda & E.Diday (1991) , *Symbolic clustering using a new dissimilarity measure.*Pattern Recognition Vol **24**, N° 6

[44] M.Jambu and M.O.Lebeaux (1978) , *Classification automatique pour l'analyse des données* Dunod,Paris

[45] R. A. Harshman,(1970) ,*Foundation of the parafac procedure*, Phonetics,**16**, UCLA, Los Angeles

[46] J A. Hartigan (1975) .*Clustering algoritms* J.Wiley, New York.

[47] L. Kauffman and P.J.Rousseeuw (1990) , *Finding groups in data.* J.Wiley, New York

[48] C.Hayashi (1956) ,Theory and examples of quantification *II, procedure .of the institue StaT .Math*, **4**(2),P19-30

[49] P. Horst (1965) ,*Factor analysis of data matrices*, Holt. Rinehart, Winston, NY

[50] R. J. Ketteling (1971) , *Canonical analysis of several sets of variables*, Biometrika, **58**(3), p 433-450

[51] WP Krijner (1993) , *The analysis of three-way arrays constrained parafac methods.* DSWO Press.

[52] R. Kodell,and J.Chen (2001) , *Infering effects on tumor frequencies and times to observation in the analysis of tumor nultiplicity data.* Biometrical Journal, 43, **4**, p:447–460.

[53] P. R Kroomenberg (1983) ,*Three mode principal component analysis.* DSWO Press.

[54] S Kullback (1959) , *Information theory and statistics*. J.Wiley, New York

[55] S. Kullback and R.A. Leibler (1967) , *Information and sufficiency*. Ann. Math Statis., **32**, p: 79–86

[56] C. Lavit (1988) , *Analyse conjointe de tableaux quantitatifs*. Masson, France.

[57] L.Lebart et N.Tabard (1973), *Recherches sur la description automatique des données*, Credoc,cr 13

[58] L.Lebart, A.Morineau et M. Piron (1996) , *Statistique exploratoire multidimensionnelle*. Dunod, France

Cet ouvrage est une synthèse des techniques factorielles utilisées en analyse de tableaux multiples(DACP,Statis,analyse procusteenne,regression multiple,analyse canonique,AFM analyse discriminate,...) . Une grande partie de cet ouvrage est consacré aux techniques de classification automatique. Une introduction à la validation de ces techniques constitue le dernier chapitre du livre (Tests ,simulation par le bootstrap ....). La présentation de ces différentes approches est faite avec un aperçu historique clair et sans ambiguïté.

[59] A. Leclerc (1975) , *L'analyse des correspondances sur juxtaposition de tableaux de contingence* RSA, **23**, p: 221–248.

Dans cet article sont étudiées les différentes propriétés concernant les sous tableaux de Burt.

[60] I. C. Lerman (1970) , *Les bases de la classification automatique*. Gauthiers Villard, Paris.

[61] I .C.Lerman (1980) , *Classification et analyse ordinale des données.* Dunod, Paris.

[62] I. C. Lerman (1980), *Analyse globale des éléments constitutifs d'une juxtaposition de tableaux de contingence* RSA XXVIII, **2.**

[63] H. L'Hermier Des Plantes (1976) , *Structuration des Tableaux A Trois Indices de la Statistique (STATIS).* Thèse de l'Université de Montpelier. France.

[64] G. Mac Lachlam (1982), *he classification and mixture maximum likelihood approch to cluster. analysis.*Hand Book of statistical ,Vol **2**, Krishnaiak. P.R and Kanak eds, pp: 199–208, Amesterdam, Holland.

[65] J. B. Macqueen (1967) , Some methods for classification and analysis multivariate observations. Proc. Sym . Math Statist and Probab. (5th), Berkely, **1**, pp: 281–297.

Dans cet article est introduite la méthode de classification par des partitions dite : méthode des K-means.

[66] A. Morinneau, A.E Sammartino, M.Getter-Summa, C.Pardoux (1994) , *Analyse des données et modélisations des séries chronologiques.* RSA, **42,** (4), pp: 61–81.

[67] F. Murtagh (1985) , *Multidimensionnal clustering algorithms* COMPSTAT lectures **4**, Physica Verlag,Vienna.

[68] N. Naab(2001) ,Analyse factorielle pour analyser les éléments constitutifs d'une structure de tableaux multiples,thése de magister,USTHB

[69] J.P. Nakache (1973) , *Influence du codage des données en analyse factorielle des correspondances.* RSA, **21**, N°2.

[70] A. Nishisato (1980) , *Analysis of categorial data,dual scaling and its applications.*Université de Torranto, Press.

C'est un ouvrage où est développé l'analyse des correspondances multiples sous le nom de dual scaling.D'autres auteurs se sont penchés sur les techniques d'analyse de tableaux multiples et ont donné des extensions à l'AFM sous des noms tels que: Homogeneity analysis (J De Leuvv,...).

[71] C. Perruchet (1983) , *Une analyse bibliographique des épreuves de classificabilité en analyse des données.* RSA Vol. **8,** p: 18–41.

[72] B.C. Petters and H.F Waker (1978) , *An iterative procedure for obtaining maximum estimate of the parameters for a mixture of normal distributions.* SIAM, JAM, **35**, p : 362–378.

[73] B.C. Petters and H.F Waker (1978) , *The numerical evaluation of the maximum likelihood ùixture proportions.* SIAM, JAM, **35**, p: 447–462.

[74] P. Orlick et H Terao (1992) :*Arrangements of hyperplanes ,Springer Verlag*

[75] JO. Ramsay (2000) , *Functional compment of variation inhardwriting.* JASA, serie B, **95,** p: 9–15.

[76] A. Rebbouh (1983) , *Classification automatique des données avec erreurs de mesure et application aux données médicales.* Thèse de Magister Université des sciences d'Alger.

[77] A. Rebbouh (1993) , *Classification des données avec erreurs de mesure: modèle gaussien.* IFCS, Paris

[78] A. Rebbouh et S.Djemai (1996) , *Classification d'objets aléatoires.* Prépublication, Institut de mathématiques université des sciences d'Alger, N° 183/36/13.

[79] A. Rebbouh (2005) , *Clustering the constitutive element of a structure of disjonctif table data.* 37èmes journées de statistique de la SFDS, juin, Pau, France.

[80] A. Rebbouh (2006) , *Clustering the constitutive element of a structure of measuring table data.* Communications in statistics simulation and computation, **35**, 2, p: 751–764.

[81] A. Rebbouh (2007) , *Classification d"objets aléatoires, deuxièmes journées de Statistiques appliquées et théoriques de Biskra,avril 2007.*

[82] A. Rebbouh (2008) ,*Clustering the constitutive element of a structure of binary data tables*

à paraitre dans JAMDS # Journal of Applied mathematics and Decision Science#

[83] R.A. Rener and H.F. Walker(1984) , *Mixture density maximum likelihood and the em algorithm*, Dekker, N.Y.

[84] M. Roux (1985) , *Algorithme de classification.* Masson, Paris.

[85] F. H. Ruspini (1969) , *A new approch to clustering.* Infor and control, **15**, p: 22–32.

[86] G. Saporta (1981) , *Méthodes exploratoires d'analyse de données temporelles.* Thèse de l'Université de Paris VI.

[87] G. Saporta (1990) , *Probabilités, analyse des données et statistique.* Edition technip-Paris.

[88] A. Schroeder (1976) , *Analyse d'un mélange de distributions de probabilités de même type.* RSA, Vol **24** ,n° 1.

[89] A.J. Scott and M.J.Symons (1971) , *Clustering methods based on likelihood ratio.* Biometrika, **27**, p: 387–397.

[90] C E Shannon (1948) , *A mathematical theory of communication* Bell System technical journal, **27**(3) p: 349–423 & 623–659.

[91] P H. Sneath & RR Sokal (1973) , *Numerical taxinomy.* Freeman, NY

Livre de base et de référence sur le clustering.

[92] M. Symons (1981) , *Clustering criteria and multivariate normal mixture.* Biometrika, **37**, p: 37–43.

[93] PH Tassi et S Legait (1990) , T*héorie des probabilités,en vue des applications statistiques.* Edition technip, Paris.

[94] N.Tennenhaus et Prieuret (1974) , *Analyse des séries chronologiques multidimensionnelles.* Revue RAIRO, **2**, p: 5–16.

[95] N.Tennenhaus et F.W Young (1985) , *An analysis and synthesis of multiple correspondance analysis,optimal scaling,dual scaling homogeinity analysis and other methods for quantiflying categorical multivariate data* Psychometrica, **50**, p: 91–119.

[96] R.L. Thorndike (1953) , *Who belongs in the family.* Psychometrika, **18**, p: 267–276.

Thorndike a introduit la notion de seuils ou protection destinée à modifier éventuellement le nombre de classes. Ses travaux ont été le prélude à la mise au point de plusieurs algorithmes de classification.

[97] L.R.Tucker (1964) ,The extension of factor analysis to three -dimensional matrices,Harris.(ed), Univ. of Wisconsin,Press p:109-127

[98] L.R. Tucker (1966), *Some mathematical notes on three mode factor analysis* Psychometrika, **31**, p: 231–279.

Cet article est un exposé synthétique des diverses approches de l'AFM.

[99] B.Van Cutsem (eds) (1989) , *Classification and dissimilarity analysis.* Springer - Verlag, N.Y.

[100] Ventsel (1973) , *Théorie des probabilités.* Edition MIR, Moscou.

[101] PGN.Von der Heidjer (1987) , *Correspondance analysis longitudinal categorial data.* DSWO Press

[102] J.H.Wolfe (1970) , *Pattren clustering by multivariate mixture analysis Multivararariate.* Behavioral research, **5**, pp: 329–350.

.

1 Appendix

Program of the algorithm that computes the ascending hierarchical of elements of a categorical data structure written under excel VBA.

```
Sub test()
Dim matin(6, 6) As Variant
Dim matrixM(6, 6) As Variant
Dim matrixT(6, 6) As Variant
Dim vect(6) As Variant
Dim delta(6, 6) As Variant
Dim vectmin(6) As Variant
Dim delta2(6, 6)
size_matin = 6
Sheets("F1").Select
' vectmin contains the minimum value and the corresponding two objects
' input data matrix should be entered into the sheet F1
For i = 1 To size_matin
For j = 1 To size_matin
matin(i, j) = Cells(i, j)
Next
Next
' computation of matrix M
nl = 0
For l2 = 1 To size_matin
For l1 = 1 To size_matin
matrixM(l2, l1) = -Log(matin(l2, l1)) / Log(2)
Next
Next
' all the intermediate outputs will be put in the sheet F2
Sheets("F2").Select
' transposition of matrix M
For i = 1 To size_matin
For j = i + 1 To size_matin
X = matrixM(i, j)
matrixM(i, j) = matrixM(j, i)
matrixM(j, i) = X
Next
Next
' product of the transpose of matrix M by the input data matrix
For i = 1 To size_matin
For j = 1 To size_matin
s = 0
For k = 1 To size_matin
s = s + matin(i, k) * matrixM(k, j)
Next
```

```
matrixT(i, j) = s
Next
Next
nligne = size_matin
Do While nligne >= 3
mindelta = 100
For l1 = 1 To nligne
For l2 = l1 + 1 To nligne
x1 = matrixT(l1, l2)
x2 = matrixT(l2, l1)
x3 = matrixT(l1, l1)
x4 = matrixT(l2, l2)
delta(l2, l1) = x1 + x2 - (x3 + x4)
If mindelta > delta(l2, l1) Then
mindelta = delta(l2, l1)
vectmin(2) = l1
vectmin(3) = l2
End If
delta(l1, l2) = delta(l2, l1)
Next
delta(l1, l1) = 0
Next
vectmin(1) = mindelta
nl = nl + 1
' reconstruction of the matrix delta
lig = 0
For i = 1 To nligne
If i <> vectmin(2) And i <> vectmin(3) Then
lig = lig + 1
col = 0
For j = 1 To nligne
If j <> vectmin(2) And j <> vectmin(3) Then
col = col + 1
delta2(lig, col) = delta(i, j)
End If
Next
End If
Next
lig = nligne - 1
col = 0
For i = 1 To nligne
If i <> vectmin(2) And i <> vectmin(3) Then
If delta(vectmin(2), i) < delta(vectmin(3), i) Then
min1 = delta(vectmin(2), i)
Else
```

```
min1 = delta(vectmin(3), i)
End If
col = col + 1
delta2(lig, col) = min1
delta2(col, lig) = min1
End If
Next
delta2(lig, lig) = 0
nligne = nligne - 1
For i = 1 To nligne
For j = 1 To nligne
matrixT(i, j) = 0
delta(i, j) = 0
If i <= nligne And j <= nligne Then
matrixT(i, j) = delta2(i, j)
End If
Next
Next
Loop
Sheets("F1").Select
End Sub
```