République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur
et de la recherche scientifique
Université de Saida - Dr Moulay Tahar
Département Mathématiques
Master : Analyse Sthocastique, Statistique des processus
et Applications (ASSPA)

Polycopié Rédigé par

Fatima Benziadi

Méthodes Avancées d'Analyse des données

Méthodes Avancées d'Analyse des données


بسم الله الرحمن الرحيم

Table des matières

Intoduction

Pour toute étude statistique, il est nécessaire de décrire et explorer les données avant d'en tirer de quelconques lois ou modèles prédictifs.

Dans beaucoup de situations, les données sont trop nombreuses pour pouvoir être visualisables. Le traitement d'un tableau de grand dimension est un problème fondamental en statistique multidimensionnelle. Aujourd'huit, des vastes données d'enquêtes sont dépouillées et fournisent des grands tableaux qui se prettent à interpréter graphiquement.

La statistique descriptive nous font habitués à étudier des variables les unes après les autres, de construire autant d'histogrammes que de variables. Comment faire pour que, à ces nombreux graphiques se substituent un seul graphe ? C'est-à- dire comment devant, la profusion des discriptions fournies par l'analyse de variable par variable, donner une vision globale de l'ensemble des résultats ? L' Analyse des données, permet de répondre à ces questions.

D'aprés J.P.Fénellon 1982 [21], l'analyse des données (ADD) est un ensemble de téchniques pour découvrir la structure compliquée d'un tableau de données à plusieurs dimensions et de le traduire par une structure plus simple et qui la résume au mieux. Cette structure peut le plus souvent, être représentée graphiquement.

L'analyse des données recouvre principalement deux ensembles de téchniques : les premiers sont appellées "Analyses factorielles" ou "Méthodes factorielles", qui sont :

1. Analyse en composantes principales (ACP)[19] ;
2. Analyse canonique (AC)[29] ;
3. Analyse factorielle de correspondance (AFC)[3][14] ;
4. Analyse factorielle de correspondance multiple (AFCM)[24].
5. Analyse discriminante (AD)[35]

Les secondes sont appelées " Les méthodes de la classification automatiques", qui sont

1. Les méthodes alogarihtmiques [32] ;[36].

2. Les méthodes probabilistes [2].

Parmi ces deux groupes, les premiers occupent une place de choix, car elles sont utilisées soit seules soit conjointement avec les secondes.

Plus précisement, ces méthodes descriptives ne supposent, a priori, aucun modèle sous-jacent, de type probabiliste. Ainsi, lorsqu'on considère un ensemble de variables quantitatives sur lesquelles on souhaite réaliser une analyse factorielle, il n'est pas nécessaire de supposer que ces variables sont distribuées selon des lois normales. Néanmoins, l'absence de données atypiques, la symétrie des distributions sont des propriétés importantes des séries observées pour s'assurer de la qualité et de la validité des résultats.

Bien que l'étude de la structure de vastes ensembles de données soit récente, les principes dont les méthodes d'analyse de données s'inspirent sont anciennes.

Historiquement, K.Pearson(1901) [34] est le premier qui a donné le principe de l'analyse en composantes principales : " les variables colonnes du tableau à analyser étant considérées comme des vecteurs d'un espace à dimensions, on proposait de réduire la dimension de l'espace en projetant le nuage des points variables sur le sous-espace de dimension pk (k petit fixé) permettant d'ajuster au mieux le nuage". Ch. Spearman (1904)[37] est le premier qui a introduit la définition d'un facteur de l'analyse factorielle. C.Brutt et L.L.Thurstone (1930) ont résumé à l'aide d'un petit nombre de facteurs une information multidimemsionnelle. Puis, H.Hotelling (1933) a dévloppée l'analyse en composantes principales (ACP). D'un point de vue plus récent L. Lebart (1979) [30] écrit, "l'analyse en composantes principales est une technique de représentation des données, ayant un caractère optimal selon certains critères algébriques et géométriques spécifiés et que l'on utilise en général sans référence à des hypothèses de nature statistique ou à un modèle particulier".

Enfin, l'analyse factorielle des correspondances introduite par J.P Benzécri (1980)[3], est actuellement en vogue. Elle fournit, sans hypothèses à priori des représentations simplifiées dans un certain sens à l'interprétation. Laissons sur ce point la parole de Professeur J.P Benzécri : "l'analyse des correspondances telle qu'on la pratique en 1977 ne se borne pas à extraire des facteurs de tout tableau de nombres positifs. Elle donne pour la préparation des données des règles telles que le codage sous-forme disjonctive complète, aide à critiquer la validité des résultats, principalement par des calculs de contribution, fournit des procédés efficaces de discrimination et de régression, se conjugue harmonieusement avec la classification automatique". Sa logique est claire : le modèle doit suivre les données non l'inverse, le modèle probabiliste est jugé trop

contraignant : "statistique n'est pas probabilité".

Les deux méthodes précédentes et celles qui en ont été dérivées, comme l'analyse factorielle discriminante présentée par Fisher en 1936 [26], qui permet de décrire la liaison entre une variable qualitative et un ensemble de variables quantitatives et l'analyse canonique introduite par Hotelling en 1936 [29] est dont l'objectif initial était d'exprimer au mieux à l'aide d'un petit nombre de couples de variables la liaison entre deux ensembles de caractères quantitatifs.

S'agissant de la classification automatique, compte tenu de la multiplicité des techniques existantes et l'effervescence qui règnent autour de ce domaine, car selon R.M. Cormack [17] plus de 1000 articles sont publiés par an sur ce thème, il est vraiment difficile de faire l'historique de ces méthodes ; en effet nombreux sont les chercheurs qui ont contribué à leur mise en oeuvre et dont les précurseurs sont : Buffon (1749), Adanson (1757) et Linné (1758). Pour terminer cette page d'histoire, mentionnons l'analyse des données non métriques introduite par une nouvelle école de statisticiens américains sous le nom de " multidimensional scaling " (J.D. Carrol, J.B. Kruskal, R.N. Shepard,...) et dont les principales méthodes sont :

- l'analyse des proximités ;
- l'analyse des préférences ;
- l'analyse de mesure conjointe (qui permet d'expliquer une variable qualitative ordinale à l'aide des variables nominales).

Ces méthodes ont trouvé leurs applications surtout dans le domaine du marketing [6]

Types de tableaux analysables

les données se représentent généralement sous la forme d'un tebleau rectangulaire, dont les lignes correspondent à des individus ou unités statistiques et les colonnes à des variables ou caractaires, ses valeurs peuvent être quantitatives ou qualitatives, lorsque dans un tableaux toutes les variables sont quantitatives, on peut établir un tableau quantitatif ou numérique. Par exemple, on observe sur un ensemble de n individus, $I = \{x_1, \cdots, x_n\}$, un certain nombre de mesures $J = \{Poid, Taille, ge, ...\}$. Ce tableau est encor appllé tableau de mesures. Dans le cas où toutes les variables sont qualitatives, on peut définire plusieurs types de tableaux par exemple, le tableau d'éffectifs, tableau de fréquences, tableau de profiles, tableau de contingence et tableau de Brute...

Analyse générale

On part d'un tableau réctangulaire X reliant deux ensembles finis, ce tableau peut admettre deux présentations :

- L'une dans un sous-espace vectoriel de $\mathbb{R}^p$ avec un nuage de n points correspondant chacun à une ligne (un individu) qui s'appele nuage d'individus.
- L'autre dans un sous-espace vectoriel de $\mathbb{R}^n$ avec un nuage de p points, correspondant chacun à une colonne (une variable), s'appele nuage de variables.

  L'analyse factorielle a été introduite par T.Foucart [27] ; elle cherche à ajuster le nuage de n ou p points par un sous-espace vectoriel de $\mathbb{R}^p$ respectivement $\mathbb{R}^n$ de dimension $k < p, n$. Les méthodes de la classifications automatique ont été introduite par M.O.LEBEAUX [31] ; elles consistent à trouver une subdivision de l'ensemble d'individus ou l'ensemble de variables en k classes homogènes.

Chapitre 1

L'analyse en composantes principales (ACP)

Dans la plupart des situations, on dispose de p observations sur chaque individu constituant la population d'étude, on a donc à prendre p variables par individu (p$>$ 3). L'étude séparée de chacune de ces variables, donne quelques informations mais inssufisantes car elle laisse de côté les liaisons ente elles, ce qui est pourtant souvent ce qui l'on veut étudier.

L'analyse en composantes principales est alors la bonne méthode pour étudier les données multidimentionnelles, lorsque les variables sont de type numérique.

1.1 Présentation d'ensembles

On part d'un tableau de données représentant toutes les données en palçant en lignes les individus et en colonnes les variables.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \in \mathcal{M}_{n\times p}(\mathbb{R})$$

x_{ij} : est la mesure de la variable x^j sur l'individu x_i, $i = 1, \cdots, n$ *et* $j = 1, \cdots, p$.

Dans la suite, on considère que X est un tableau de données centrées(pour que le centre de gravité du nuage d'individus $g = (\overline{x^1}, \overline{x^2}, \cdots, \overline{x^p})$ coïncide avec lorigine de $\mathbb{R}^p$.)

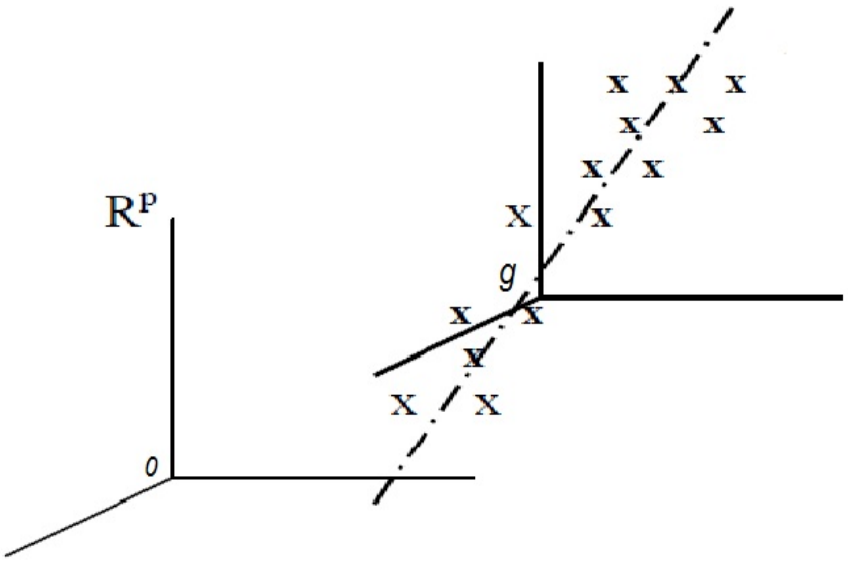

FIGURE 1.1 – Nuage des données dans l'espace $\mathbb{R}^p$.

$$X = \begin{pmatrix} x_{11} - \overline{x^1} & x_{12} - \overline{x^2} & \cdots & x_{1p} - \overline{x^p} \\ x_{21} - \overline{x^1} & x_{22} - \overline{x^2} & \cdots & x_{2p} - \overline{x^p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x^1} & x_{n2} - \overline{x^2} & \cdots & x_{np} - \overline{x^p} \end{pmatrix} \in \mathcal{M}_{n\times p}(\mathbb{R})$$

En pratique, On donne même importance à chaque variable, on travaille avec des données centrées et réduites.

$$X = \begin{pmatrix} \frac{x_{11}-\overline{x^1}}{\sigma_{x^1}} & \frac{x_{12}-\overline{x^2}}{\sigma_{x^2}} & \cdots & \frac{x_{1p}-\overline{x^p}}{\sigma_{x^p}} \\ \frac{x_{21}-\overline{x^1}}{\sigma_{x^1}} & \frac{x_{22}-\overline{x^2}}{\sigma_{x^2}} & \cdots & \frac{x_{2p}-\overline{x^p}}{\sigma_{x^p}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\overline{x^1}}{\sigma_{x^1}} & \frac{x_{n2}-\overline{x^2}}{\sigma_{x^2}} & \cdots & \frac{x_{np}-\overline{x^p}}{\sigma_{x^p}} \end{pmatrix} \in \mathcal{M}_{n\times p}(\mathbb{R})$$

Soit on prend X comme un ensemble de n points de dimension p, soit on le prend comme un ensemble de p points de dimension n.

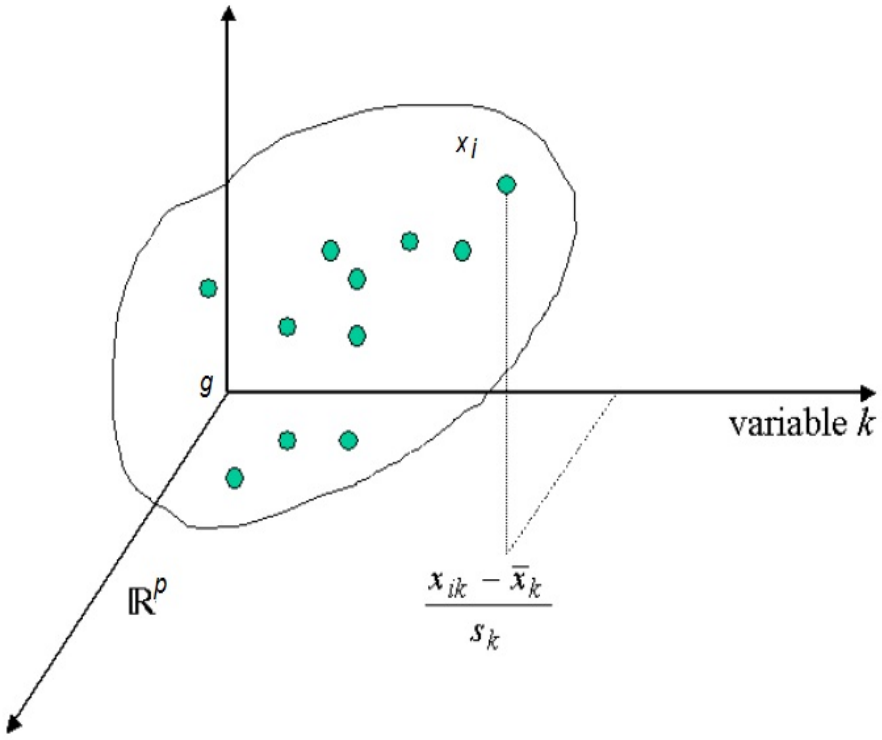

FIGURE 1.2 – Nuage des individus dans l'espace $\mathbb{R}^p$.

1.1.1 Les nuages de points initiaux

Nuage d'individus

Nous représentons graphiquement les individus par un nuage de points dans un sous-espace E de $\mathbb{R}^p$ et l'information intéressante pour l'individu est la distance entre les points.

Nuage de variables

Dans l'espace de variables $F \subseteq \mathbb{R}^n$, on définit la métrique de poids $N = D_p = \frac{1}{n} Id_n$ tel que :

$$\begin{aligned} \forall x^j, x^k \in F : D_p(x^j, x^k) &= < x^j, x^k >_{D_p} \\ &= (x^j)^t D_p x^k \\ &= \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} = cov(x^j, x^k) \end{aligned}$$

On déduit que $D_p(x^j, x^j) = \| x^j \|_{D_p} = var(x^j)$

On travaille avec des variables normées donc $D_p(x^j, x^k) = r(x^j, x^k)$,(où r est la rorrélation entre x^j et x^k) d'autre part $-1 \leq r \leq 1$ c'est-à-dire la distance entre toutes les variables est entre -1 et 1, donc le nuage de variables est situé dans une sphère de rayon 1 et de centre 0.

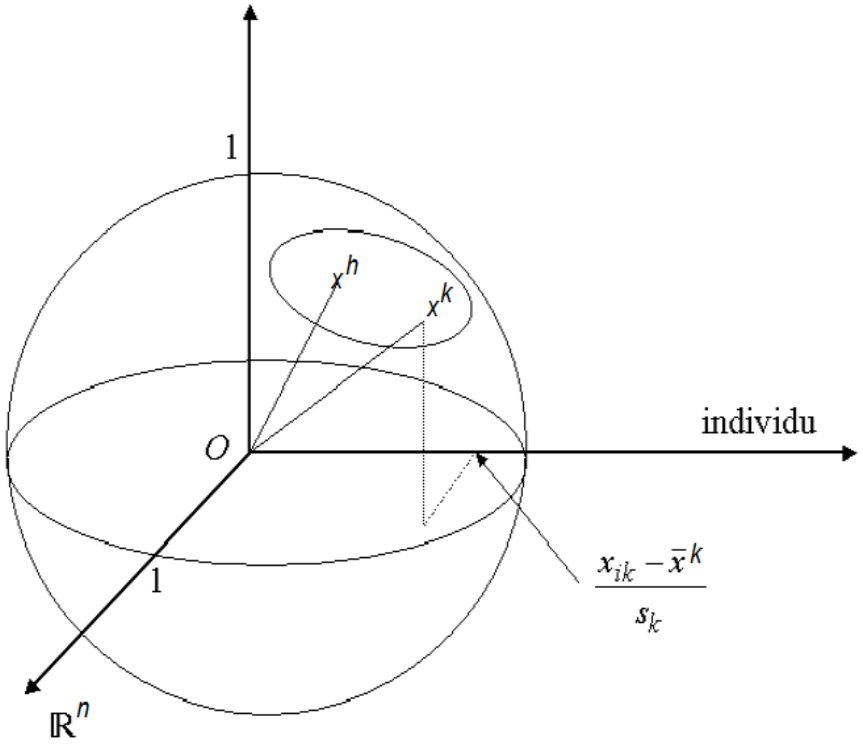

FIGURE 1.3 – Nuage des variables dans l'espace $\mathbb{R}^n$.

1.1.2 Le schéma de dualité

Soit $\{e_1, e_2, \cdots, e_p\}$ la base canonique de $E \subseteq \mathbb{R}^p$, $\{f_1, f_2, \cdots, f_n\}$ la base canonique de $F \subseteq \mathbb{R}^n$ alors ,

$$x_i \in E \quad \Leftrightarrow \quad \exists x_{ij} \in \mathbb{R};\ j = 1.\cdots.p \ \ tq \ \ x_i = \sum_{j=1}^{p} x_{ij} e_j,$$

$$x^j \in F \quad \Leftrightarrow \quad \exists x_{ij} \in \mathbb{R};\ i = 1.\cdots.n \ \ tq \ \ x^j = \sum_{i=1}^{n} x_{ij} f_i$$

Soit $h \in \{1, \cdots, p\}$ et soit $e_h^* \in E^*$, on a : $e_h^*(x_i) = e_h^*\left(\sum_{j=1}^{p} x_{ij} e_j\right) = \sum_{j=1}^{p} x_{ij} e_h^*(e_j) = x_{ih}$

On déduit que $e_h^*(x_i)$ c'est exactement la valeur de la variable x^h obtenue sur x_i , e_h^* peut être assimulé à la variable x^h, on le concidère comme représentant de la variable x^h dans E^*. Nous remarquons que $\forall j \in \{1, \cdots, p\}$, on a :$Xe_j^* = x^j \in F \subseteq \mathbb{R}^n$. Il résulte que X est une application linéaire de E^* dans F, par conséquence, X^t est aussi une application linéaire de F^* dans E.

Soient M, N deux métriques définies respectivement sur E, F , on peut shématiser toutes les informations précédentes dans le schéma suivant qui s'appelle le schéma de dualité :

$$\begin{array}{rcl} \mathbb{R}^p \supseteq E & \longleftarrow^{X^t} & F^* \\ V \uparrow M \downarrow & & \uparrow N \downarrow W \\ E^* & \longrightarrow^{X} & F \subseteq \mathbb{R}^n \end{array}$$

Apartir de ce schéma, on peut définir d'autre applications $V = (X)^t NX, \quad W = XM(X)^t$.

1.2 Le problème de l'analyse en composantes principales

Le problème de l'analyse en composantes principales est de trouver un sous-espace E_k de dimension $k < p$ de tel sorte que l'inertie[1] par rapport à E_k noté I_{E_k} soit minimale. Pour construire cet sous-espace E_k il faut trouver les axes qui l'engendrent le problème devient à chercher le premier axe Δ_1 passant par le centre de gravité g et d'inertie I_{Δ_1} minimale , puis le deuxième axe Δ_2 passant par g et d'inertie I_{Δ_2} minimale et $\Delta_1 \perp \Delta_2$, jusqu'à le $k-$ième axe Δ_k.

1.2.1 La recherche de Δ_1

On cherche un axe Δ_1 passant par g d'inertie I_{Δ_1} minimale, car c'est l'axe le plus proche de nuage d'individus, si l'on doit projeter ce nuage sur cet axe, c'est lui qui donnera l'image la moins déformée du l'image initiale.

1. L'inertie est la moyenne de distances entre un nuage de points et un point ou une droite ou un sous-espace

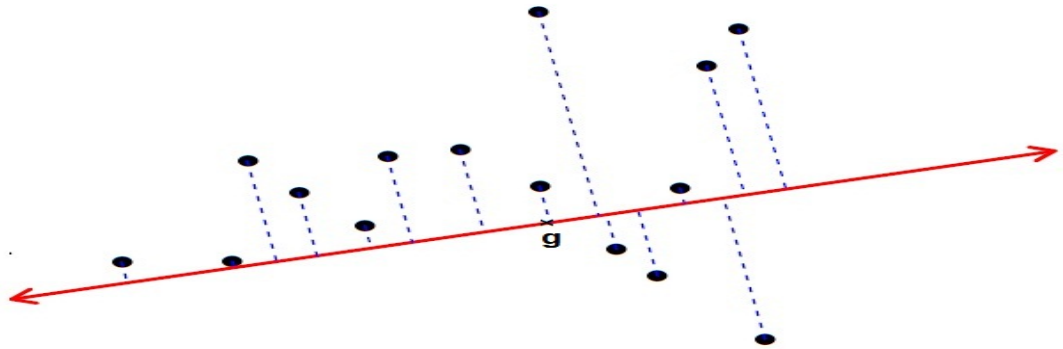

FIGURE 1.4 – Le premier axe passe par g.

1.2.2 L'inertie d'une droite

Soit u_1 un vecteur dans E normé , et soit Δ_1 la droite engendrée par u_1, alors ,

$$E = \Delta_1 \oplus \Delta_1^{\perp} \Rightarrow I_E = I_{\Delta_1} + I_{\Delta_1^{\perp}}$$

Soit $x_i \in E \Rightarrow \exists \alpha_i \in \Delta_1, \exists \beta_i \in \Delta_i^{\perp}$ *tel que* $x_i = \alpha_i + \beta_i$

donc

$$\begin{aligned} I_{\Delta_1} &= \frac{1}{n}\sum_{i=1}^{n} \| x_i - \alpha_i \|_M^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \| \beta_i \|_M^2 \\ I_{\Delta_1^{\perp}} &= \| \alpha_i \|_M^2 \end{aligned}$$

Proposition 1.2.1 *[21]* $I_{\Delta_1^{\perp}} = MVM(u_1, u_1)$

Preuve 1.2.1 $\alpha_i \in \Delta_1 \Rightarrow \exists c_i \in \mathbb{R},\ \alpha_i = c_i u_1$.

c_i est la coordonnée de la projection $M-$orthogonale de x_i sur Δ_1, donc, on peut écrire

$$c_i = M(x_i, u_1).$$

Alors, $\| \alpha_i \|^2_M = M(c_i u_1, c_i u_1) = c_i^2$, *donc,*

$$I_{\Delta_1^\perp} = \frac{1}{n} \sum_{i=1}^{n} c_i^2,$$

Pour cela , soit

$$C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} M(x_1, u_1) \\ M(x_2, u_1) \\ \vdots \\ M(x_n, u_1) \end{pmatrix} = XMu_1 \in \mathbb{F}$$

D'une part on a : $\| C \|^2_N = \frac{1}{n} \sum_{i=1}^{n} c_i^2$ *, d'autre part on a :*$I_{\Delta_1^\perp} = \frac{1}{n} \sum_{i=1}^{n} c_i^2$, *donc ,*

$$\begin{array}{rcl} I_{\Delta_1^\perp} & = & \| C \|^2_N \\ & = & N(XMu, XMu) \\ & = & u^t M X^t N X M u \\ & = & u^t M V M u \end{array} \tag{1}$$

■

Théorème 1.2.1 *(théorème de Huygense)[21]*

L'inertie du nuage d'individus par rapport au centre de gravité $g = (0, 0, \cdots, 0)$ est donnée par

$I_g = \frac{1}{n} \sum_{i=1}^{n} \| x_i \|^2_M$ *et vérifié*

$$I_g = I_\Delta + I_{\Delta^\perp} \tag{2}$$

Preuve 1.2.2 *on pose : h_Δ la projection orthogonale de x_i sur Δ et $h_{\Delta^\perp}$ la projevtion orthogonale de x_i sur $\Delta^\perp$.*

D'après le théorème de Pythagorthe on a :

$$\begin{aligned} d^2(g, x_i) &= d^2(x_i, h_{\Delta^\perp}) + d^2(x_i, h_\Delta) \\ &= d^2(g, h_\Delta) + d^2(g, h_{\Delta^\perp}) \\ &= I_\Delta + I_{\Delta^\perp} \end{aligned}$$

■

Remarque 1.2.1 *D'après le thèorème de Huygence (1.2.1) on déduit que la recherche de* Δ *qui minimise* I_Δ *équivalent à la recherche de* $\Delta^\perp$ *qui maximise* $I_{\Delta^\perp}$

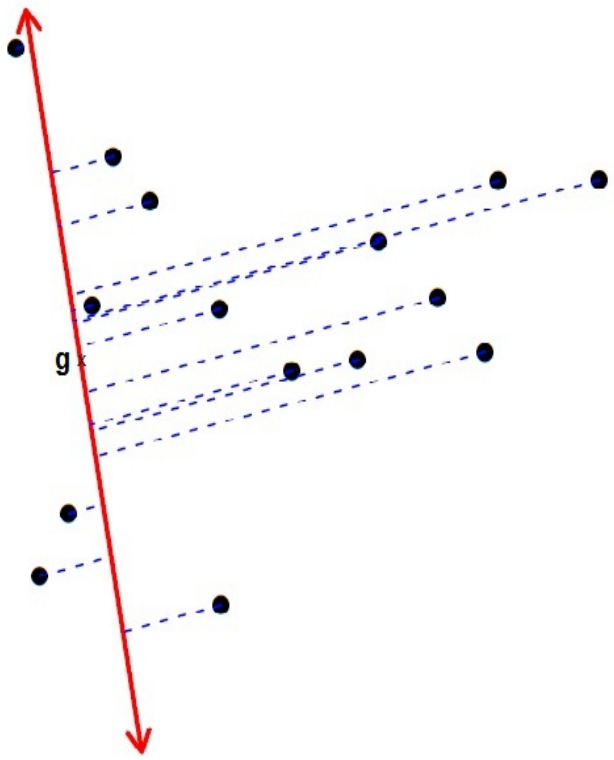

FIGURE 1.5 – Les projections sur l' axe orthogonale .

1.2.3 Le maximum de MVM

Le problème à résoudre est de trouver le vecteur u qui maximise $MVM(u, u)$ sous la contrainte $M(u, u) = 1$. Le problème de la recherche d'un optimum d'une fonction à plusieurs variables liées par une contrainte (les inconnus sont les coordonnées de u), la méthode

de multiplicateur de Lagrange peut alors être utilisée. Dans le cas de la recherche de u_1, il faut calculer les dérivées partielles de la fonction suivante :

$$\begin{aligned} g(u_1) &= g(u_{11}, u_{12}, \cdots, u_{1p}) = MVM(u_1, u_1) \\ &= u_1^t MVMu_1 - \lambda_1((u_1^t Mu_1) - 1) \end{aligned}$$

Proposition 1.2.2 *On utilise les dérivées partielles de la fonction g on trouve :*

$$\frac{dg(u_1)}{du_1} = 2MVMu_1 - 2\lambda_1 u_1 = 0$$

Preuve 1.2.3 *On a :*

$$\frac{dg(u_1)}{du_1} = \begin{pmatrix} \frac{\partial g(u_1)}{\partial u_{11}} \\ \frac{\partial g(u_1)}{\partial u_{12}} \\ \vdots \\ \frac{\partial g(u_1)}{\partial u_{1p}} \end{pmatrix}$$

et

$$\begin{aligned} \frac{d(u_1^t MVMu_1)}{du_1} &= \begin{pmatrix} \frac{\partial u_1^t}{\partial u_{11}}(MVMu_1) + (u_1^t MVM)\frac{\partial u_1}{\partial u_{11}} \\ \frac{\partial u_1^t}{\partial u_{12}}(MVMu_1) + (u_1^t MVM)\frac{\partial u_1}{\partial u_{12}} \\ \vdots \\ \frac{\partial u_1^t}{\partial u_{1p}}(MVMu_1) + (u_1^t MVM)\frac{\partial u_1}{\partial u_{1p}} \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} \frac{\partial u_1^t}{\partial u_{11}}(MVMu_1) \\ \frac{\partial u_1^t}{\partial u_{12}}(MVMu_1) \\ \vdots \\ \frac{\partial u_1^t}{\partial u_{1p}}(MVMu_1) \end{pmatrix}}_{A_1} + \underbrace{\begin{pmatrix} (u_1^t MVM)\frac{\partial u_1}{\partial u_{11}} \\ (u_1^t MVM)\frac{\partial u_1}{\partial u_{12}} \\ \vdots \\ (u_1^t MVM)\frac{\partial u_1}{\partial u_{1p}} \end{pmatrix}}_{A_2} \end{aligned}$$

On peut remarquer que les lignes de A_1 et de A_2 sont égales puisque chacune est la transposée de l'autre, donc :

$$\frac{d(u_1^t MVMu_1)}{du_1} = 2\frac{d(u_1^t)}{du_1}(MVMu_1) = 2MVMu_1$$

car

$$\frac{du_1^t}{du_1} = \begin{pmatrix} \frac{\partial u_1^t}{\partial u_{11}} \\ \frac{\partial u_1^t}{\partial u_{12}} \\ \vdots \\ \frac{\partial u_1^t}{\partial u_1} \end{pmatrix} = \mathbb{I}_p$$

■

Alors le système à résoudre est le suivant :

$$\begin{cases} MVMu_1 - \lambda_1 Mu_1 &= 0 \\ u_1^t Mu_1 - 1 &= 0 \end{cases} \tag{3}$$

La première équation de (3) nous donne

$$VMu_1 = \lambda_1 u_1, \tag{4}$$

donc u_1 est le vecteur propre de VM associé à la valeur propre λ_1. Si on multiplie à gauche les deus mombres de (4) par u_1^t , on trouve :

$$u_1^t MVMu_1 = \lambda_1 u_1^t Mu_1$$

D'après la deuxième équation de (3), on trouve :

$$u_1^t MVMu_1 = \lambda_1 \tag{5}$$

D'après (1) on déduit

$$I_{\Delta_1^\perp} = \lambda_1 \tag{6}$$

qui doit être maximale, cela signifie que la valeur propre λ_1 est la plus grande valeur propre de VM , alors l'axe Δ_1 pour lequel le nuage d'individus a une inertie I_g minimale, a comme vecteur de base u_1, le premier vecteur propre associé à la plus grande valeur propre λ_1 de la matrice VM qui est la matrice de corrélation Γ de X.

On peut rechercher les autres axes, en suivant la même procédure, les nouveaux axes sont les vecteurs propres de Γ correspondant aux valeurs propres ordonnées au sens décroissante

$\lambda_1 > \lambda_2 > \ldots > \lambda_p$ .La matrice Γ est symétriqe réelle, elle possède p vecteurs propres réels, formant une base orthonormée de E

$$\begin{cases} u_1 \perp u_2 \perp \ldots \perp u_p \\ \Delta_1 \perp \Delta_2 \perp \ldots \perp \Delta_p \\ I_{\Delta_1^\perp} > I_{\Delta_2^\perp} > \ldots > I_{\Delta_p^\perp} \end{cases} \tag{7}$$

1.2.4 Éléments principaux de l'ACP

Les axes principaux

Les axes principaux sont les droites engendrées par les k premiers vecteurs propres associés au k premières valeurs propres de la matrice $\Gamma = VM$, ces axes forment une base orthonormée de plan principal E_k.

Les facteurs principaux

Les facteurs principaux de l'ACP d'un tableau X sont les vecteurs de base de l'espace E^*. notés par $\{u_1^*, \ldots, u_p^*\}$.

On a $\forall i \in \{1, \ldots, p\} \quad VMu_i = \lambda_i u_i \Rightarrow MV \underbrace{Mu_i}_{u_i^*} = \lambda_i \underbrace{Mu_i}_{u_i^*} \Rightarrow MVu_i^* = \lambda_i u_i^*$.

Les facteurs principaux sont les vecteurs propres associés aux valeurs propres de la matrice MV.

Les composantes parincipales

les composantes principales sont les vecteurs de base de l'espace F_k, notée par $\{C_1, \ldots, C_k\}$. Soit $F_k \subset F$, l'espace le plus proche au nuage des variables projetés. D'après le schéma de dualité on a,

$$\forall i \in \{1, \ldots, n\} \;\; C_i = XMu_i$$

D'autre part on a :

$$\begin{aligned} \forall i \in \{1, \ldots, p\} \;\; VMu_i = \lambda_i u_i \;\; &\Rightarrow \;\; XMVMu_i = \lambda_i XMVMu_i \\ &\Rightarrow \;\; \underbrace{XMX^t}_{W} N \underbrace{XMu_i}_{C_i} = \lambda_i \underbrace{XMu_i}_{C_i} \\ &\Rightarrow \;\; WNC_i = \lambda_i C_i \end{aligned}$$

Les composantes principales sont les vecteurs propres associés aux valeurs propres de la matrice WN

Proposition 1.2.3 *La composante C_i est le vecteur dont les coordonnées sont les projections $M-$orthogonales de n individus sur l'axe engendré par u_i tel que :*

1. $\overline{C_i} = 0$
2. $Var(C_i) = \lambda_i$

En effet,

1. $\overline{C_i} = \sum_{i=1}^{n} p_i M(x_i, u_i) = \sum_{i=1}^{n} M(p_i x_i, u_i) = M(\sum_{i=1}^{n} p_i x_i, u_i) = M(\overline{x^i}, u_i) = 0$
2. $Var(C_i) = \overline{C_i}^2 = D_p(C_i, C_i) = u_i^t M \underbrace{X^t D_p X}_{V} M u_i = MVM(u_i, u_i) = \lambda_i$

1.3 Contribution des axes à l'inertie totale

En utilisant le thèorème de Huygens (1.2.1), on peut décomposer l'inertie totale du nuage d'individus

$$I_g = I_{\Delta_1^\perp} + \cdots + I_{\Delta_p^\perp} = \lambda_1 + \cdots + \lambda_p$$

La contribution absolue de l'axe Δ_k à l'inertie I_g notée *cta* est la valeur propre qui lui associée

$$cta(\Delta_k \setminus I_g) = \lambda_k$$

Sa contribution relative est égale à

$$ctr(\Delta_k \setminus I_g) = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

On peut étendre ces définitions à tous les sous espaces engendrés par les axes principaux .Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes Δ_1, Δ_2 est égal à :

$$ctr(\Delta_1 + \Delta_2 \setminus I_g) = \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i}$$

On se contente souvent de faire des représentations du nuage d'individus dans un sous espace engendré par les k premiers axes principaux si cet sous-espace explique un pourcentage d'inertie proche de 1.

1.4 Interprétation des données

1.4.1 Représentation des individus dans le plan principal

Pour faire la représentation graphique des individus dans le plan principal, il suffit de calculer les coordonnées des individus dans les axes principaux. Pour obtenir $\widehat{x}_{ik}$, coordonnée de l'individu x_i sur l'axe Δ_k, on projette orthogonalement le vecteur $\overrightarrow{gx_i}$ sur cet axe et on obtient :

$$\widehat{x}_{ik} = C_k u_k = M(x_i, u_k) \Rightarrow C_i = XMu_i$$

Remarque 1.4.1 *Les coordonnées de la i-ième composante principale C_i sont les projections $M-$orthogonales de n individus $x_1, \cdots, x_n$ sur le i-ième axe principal Δ_i.*

1.4.2 Représentation des variables

Soit $F_k \subset F$ le sous-espace engendré par les k premières composantes principales, la projection D_p-orthogonale de la variable X^j sur F_k revient à projeter x^j sur les composantes qui constituent une base orthogonale de F_k, pour un couple de composantes principales (Cj, C_k) on représente ces corrélations linéaires avec chaque variable x^j sur une figure qui s'appelle, cercle de corrélation, c'est à dire chaque x^j est représenté par un point $\widehat{x^j} = (r(x^j, C_j), r(x^j, C_k))$

1.5 Les critères de la qualité de l'ACP

1.5.1 Le nombre d'axes à retenir

Plusieurs critères ont été proposer pour choisir le nombre d'axes à retenir parmi ces critères, on a le critère de la part de l'inertie, ce critère propose de retenir les axes qu' ont une grande part d'inertie.

1.5.2 Critère de $\cos^2$

Cette qualité est mesurée par le $\cos^2(\theta)$ où θ est l'angle entre X_i et les axes du plan principal, si $\cos^2(\theta) \longrightarrow 1$, x_i est plus proche de plan principal donc est bien présenté.

1.5.3 Aide à l'interprétation

Si pour les variables numériques, la vusialisation des vecteurs à l'intérieur du cercle des corrélations donne toute l'information nécéssaire à l'analyse, il peut être utile de définir, pour chaque individu, les aides suivantes :

– La contribution à l'inertie du nuage :

$$ctr(x_i) = \frac{\| x_i \|_M^2}{I_g}$$

– La contribution à l'inertie portée par un axe Δ_k

$$ctr_k = \frac{\| x_i \|_M^2}{\lambda_k}$$

Par construcution $\sum_{i=1}^{n} ctr(x_i) = 1$

- Un élément peut être contributif quasi indépendant de l'axe si cos^2 faible et ctr faible,
- Un élément est très contributif mais peu illustratif de l'axe si cos^2 faible et ctr forte ,
- Un élément peut être contrubitif mais bien ullstratif de l'axe si cos^2 forte et ctr faible,
- Un élément est particulièrement caractéristique de l'axe si cos^2 forte et ctr forte.

1.6 Exemple

Réaliser une ACP du tableau X suivant :

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix}$$

•Le centre de gravité, $g = (1, 1, 1)$

•$\sigma_{x^1} = \frac{\sqrt{2}}{2}, \quad \sigma_{x^2} = 1, \quad \sigma_{x^3} = 1$

• La matrice centrée et réduite de X

$$X_{cr} = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & -1 \\ \sqrt{2} & 1 & -1 \\ -\sqrt{2} & -1 & 1 \end{pmatrix}$$

• La matrice de corrélation

$\Sigma = V = X_{cr}^t N X_{cr} = \frac{1}{4} X_{cr}^t X_{cr}$

$$V = \begin{pmatrix} 1 & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

• Les valeurs propres de VM

$$\lambda_1 = 1 + \frac{\sqrt{2}}{2} \quad , \lambda_2 = 1, \quad \lambda_3 = 1 - \frac{\sqrt{2}}{2}$$

• Les 2 premiers vecteurs propres de VM

$u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad ||u_1||_M^2 = 2 \Longrightarrow u_1^{'} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}$ ($u_1^{'}$ est le vecteur normé de u_1)

$u_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad ||u_1||_M^2 = 1.$

$u_1^{'}, \ u_2$ sont les deux premiers axes principaux.

• Les composantes principales

$$C_1 = X_{cr} M u_1^{'t} = \begin{pmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 1 + \frac{1}{\sqrt{2}} \\ -1 - \frac{1}{\sqrt{2}} \end{pmatrix}, \ \overline{C}_1 = 0 \ , \ ||C_1||_N^2 = \lambda_1 = 1 + \frac{1}{\sqrt{2}}$$

$$C_2 = X_{cr} M u_2^t = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}, \ \overline{C}_2 = 0 \ , \ ||C_2||_N^2 = \lambda_2 = 1$$

•La présentation graphique :

1. Présentation d'individus :

$$\begin{aligned}\widehat{x}_1 &= (\tfrac{-1}{\sqrt{2}}, -1)\\ \widehat{x}_2 &= (\tfrac{1}{\sqrt{2}}, -1)\\ \widehat{x}_3 &= (1 + \tfrac{1}{\sqrt{2}}, 1)\\ \widehat{x}_4 &= (-1 - \tfrac{1}{\sqrt{2}}, 1)\end{aligned}$$

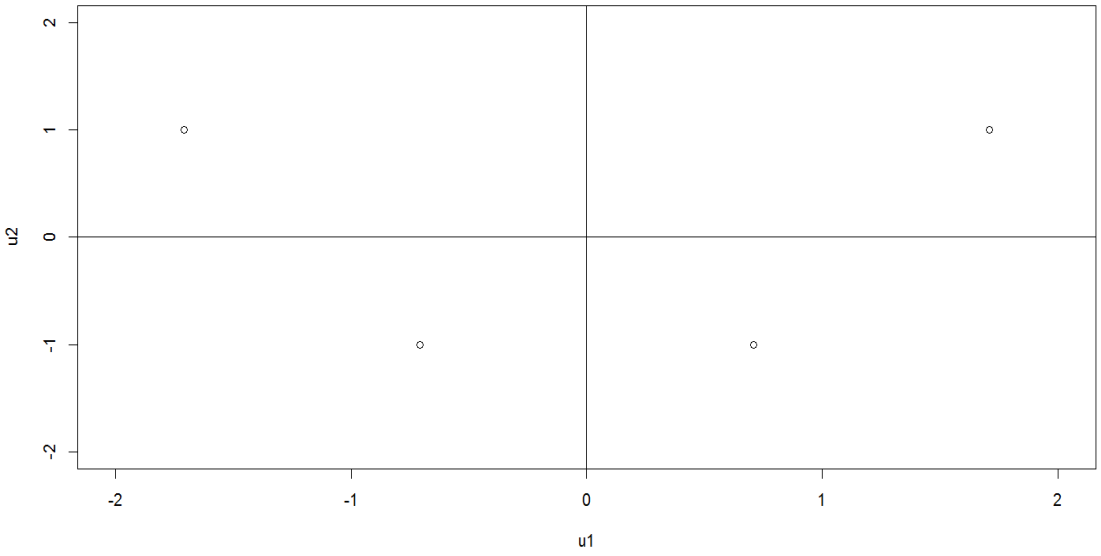

FIGURE 1.6 – La présentation des individus .

2. Présentation des variables
   Calcul des corrélations entre les variables x^1, x^2, x^3 et C_1et les variables x^1, x^2, x^3 et C_2

$$\begin{aligned}\widehat{x}^1 &= (0.9330512, 0)\\ \widehat{x}^2 &= (0.9114594, 0)\\ \widehat{x}^3 &= (-0.7317374, 0.5773503)\end{aligned}$$

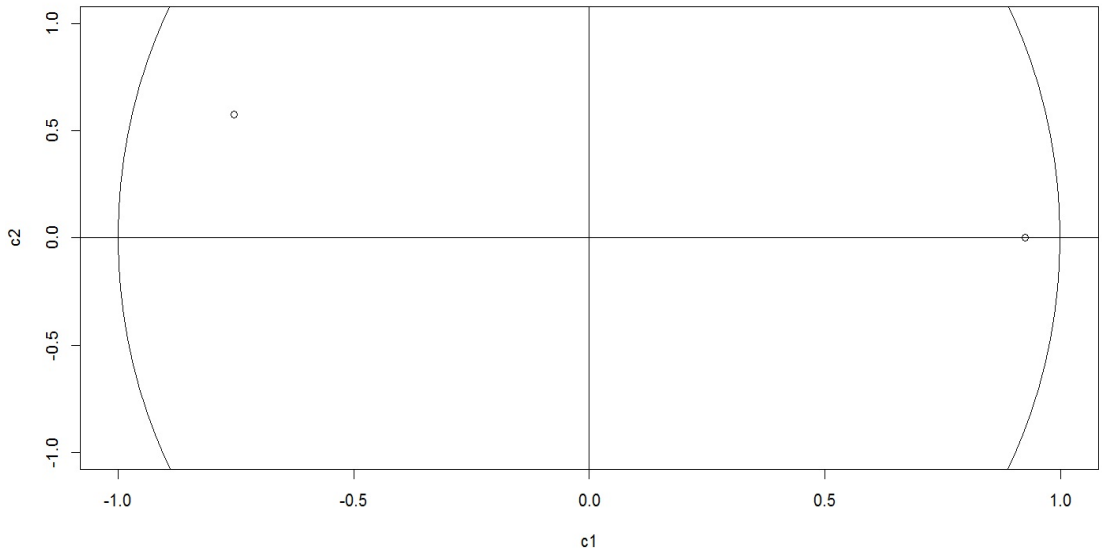

FIGURE 1.7 – La présentation des variables .

1.7 Exercices

Exercice 1

On concidère le tableau des données X suivant

$$X = \begin{pmatrix} 2 & 3 \\ 2 & 1 \\ 3 & 1 \end{pmatrix}$$

1. Donner le tableau des données centrées et réduites de X,
2. Déterminer la matrice des corrélation Σ,
3. Détermimer les valeurs et les vecteurs propres de Σ,
4. Quelles sont les axes principaux,
5. Déduire les composantes principales,
6. Présenter les individus et les variables de X dans les sous-espaces principaux.

Exercice 2

On dispose du classement de 11 individus sur 3 matières : math, musique et français. Le classement en math revient à numéroter les individus. Le tableau des classements selon les trois matières est le suivant :

Math	1	2	3	4	5	6	7	8	9	10	11
Musique	6	1	3	4	5	2	9	7	10	11	8
Français	2	5	6	3	7	4	11	10	9	1	8

1. Calculer le centre de gravité g du nuage des individus,
2. Calculer le tableau centré et réduit,
3. Calculer la matrice de var-cov de X,
4. Quelle est l'inertie du nuage d'individus ?
5. Chercher les axes principaux u_i pour X,
6. Chercher les composantes principales c_i,
7. Quelle est la contribution absolue de l'axe u_1 à l'inertie du nuage ?
8. Quel est le taux d'inertie extrait par l'axe u_1 ?
9. Effectuer la représentation graphique du plan engendré par u_1, u_2,
10. Effectuer la repr ?sentation graphique du plan engendré par c_1, c_2,
11. Étudier la qualité de la représentation graphique.

Chapitre 2

L'analyse canonique

2.1 Introduction

L'analyse canonique (AC) est une méthode d'analyse multidimensionnelle qui consiste à traiter deux tableaux des données quantitatifs. l'analyse canonique est une généralisation de l'ACP à deux tableaux numériques d'une part et d'autre part est une généralisation de la régréssion linéaire multiple dans le sens où la multiplication se porte au niveau de la variable réponse.

L'objectif général de l'A.C est d'écrire les liaisons pouvant exister entre deux groupes de variables quantitatives observés sur le même ensemble d'individus, afin d'expliquer un groupe avec l'autre.

2.2 Les données

Les données se représentent sous forme de deux tableaux X et Y

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \qquad Y = \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix}$$

- $x_{ij}, i = 1, \cdots, n \; et \; j = 1, \cdots, p$ représente la mesure la $j - ime$ variable x^j sur le i-ième individu x_i
- $y_{ik}, i = 1, \cdots, n \; et \; k = 1, \cdots, q$ représente la mesure la $k^{-}ime$ variable y^k sur le i-ième individu x_i

- $E_X \subseteq \mathbb{R}^p$ le sous-espace de $\mathbb{R}^p$,c'est l'espace d'individus mesurés par le premier paquet de variables $\{x^1, \cdots, x^p\}$
- $E_Y \subseteq \mathbb{R}^q$ le sous-espace de $\mathbb{R}^q$, c'est l'espace d'individus mesurés par le deuxième paquet de variables $\{y^1, \cdots, y^q\}$
- La matrice de poids $N = \frac{1}{n} Id_n$
- $V_{XX} = X^t NX,\ V_{YY} = Y^t NY\ \ V_{XY} = X^t NY$ et $V_{YX} = Y^t NX$
- $\mathbb{P}_1 = X(X^t NX)^{-1} X^t N = XV_{XX}^{-1} X^t N$ La projection N-orthogonale sur le sous-espace engendré par les colonnes de X
- $\mathbb{P}_2 = Y(Y^t NY)^{-1} Y^t N = YV_{YY}^{-1} Y^t N$ la projection $N-$orthogonale sur le sous-espace engendré par les colonnes de Y.

2.3 Le principe de la méthode

On cherche à déterminer les ressemblances entre les variables du tableau Y et les variables du tableau X, ou juste à décrire les relations entre ces deux groupes de variables. Pour cela, on cherche à synthétiser ces ressemblances. On recherche alors les combinaisons linéaires des variables de X et de Y qui soient les plus corrélées possible. L'idée est de chercher le couple (ξ_1, φ_1) de vecteurs normés, tels que :

$$\xi_1 = Xa \qquad \varphi_1 = Yb \tag{1}$$

forment l'angle le plus faible.

ξ_1, φ_1 sont appelés les premières variables canoniques, Les vecteurs a_1 et b_1 les premiers facteurs canonoques ne sont pas uniques. Pour assurer leur unicité, on supose que les vecteurs ξ_1 et φ_1 d'être des vecteurs mormés. r_1 La corrélation entre ξ_1 et φ_1 est appelée première correlation canonique.

En géneral, ξ_1 et φ_1 n'expliquent pas l'ensemble des liaisons entre les X et les Y. On cherche alors deux nouvelles variables normées non correléees avec ξ_1 et φ_1 de corrélation maximale (après ξ_1 et φ_1). On continue le procédure et on définit ainsi les s couples de

variables canoniques et une suite de corrélations canoniques décroissantes :

$$\begin{array}{ll} \xi_1 = a_{11}x^1 + a_{21}x^2 + \cdots + a_{P1}x^P = Xa_1 & \varphi_1 = b_{11}y^1 + b_{21}y^2 + \cdots + b_{q1}y^q = Yb_1 \\ \xi_2 = a_{12}x^1 + a_{22}x^2 + \cdots + a_{P2}x^P = Xa_2 & \varphi_2 = b_{12}y^1 + b_{22}y^2 + \cdots + b_{q2}y^q = Yb_1 \\ r_1 \geq r_2 \geq r_3 \cdots r_s & s = \min(p, q) \end{array}$$

2.3.1 La recherche des variables canoniques

Le problème consiste à chercher deux vecteurs ξ_1 et φ_1 tel que

$$r(\xi_1, \varphi_1) = \frac{cov(\xi_1, \varphi_1)}{\sigma_{\xi_1}, \sigma_{\varphi_1}} = \cos < \xi_1, \varphi_1 > \tag{2}$$

soit maximale. Une interprétation géométrique consiste à rechercher des directions de E_X et E_Y les plus proches possibles (d'angle minimal) tel que $Xa_1 \approx Yb_1$, pour ce faire, on définit deux sous-espaces F_1 de F et F_2 de F, engendrés par les combinaisons linéaires des variables $(x^1, \cdots, x^p)$ et $(y^1, \cdots, y^q)$ tels que :

$$\begin{array}{rcl} F_1 & = & \left\{\xi_i \ tq \ \xi_i = a_{1i}x^1 + a_{2i}x^2 + \cdots + a_{Pi}x^P = Xa_i\right\} \\ F_2 & = & \left\{\varphi_j \ tq \ \varphi_j = b_{1j}y^1 + b_{2j}y^2 + \cdots + b_{qj}y^q = Yb_j\right\} \end{array}$$

On cherche ξ tel que sa distance avec φ soit minimale.

On peut remarquer que φ_1 l'élément de F_2 le plus proche de ξ_1 l'élément de F_1 qui doit être $\mathbb{P}_1\varphi_1$, l'élément de F_1 le plus proche de φ_1 est sa projection D_p-orthogonale $\mathbb{P}_1\varphi_1$ sur F_1, et de manière réciproque , l'élément de F_2 le plus proche de ξ est sa projection D_p-orthogonale $\mathbb{P}_2\xi_1$ sur F_2 , donc φ_1 doit être colinéaire avec $P_1\varphi_1$,c'est-à-dire

$$\xi_1 = r(\xi_1, \varphi_1)\mathbb{P}_1\varphi_1 \tag{3}$$

$$\varphi_1 = r(\xi_1, \varphi_1)\mathbb{P}_2\xi_1 \tag{4}$$

De l'équation (3) et l'équation (4) on trouve :

$$\left\{ \begin{array}{rcl} \mathbb{P}_1\mathbb{P}_2\xi_1 & = & \lambda_1\xi_1 \\ \mathbb{P}_2\mathbb{P}_1\varphi_1 & = & \lambda_1\varphi_1 \end{array} \right. \tag{5}$$

On montre que $\lambda_1 = r^2(\xi_1, \varphi_1)$.

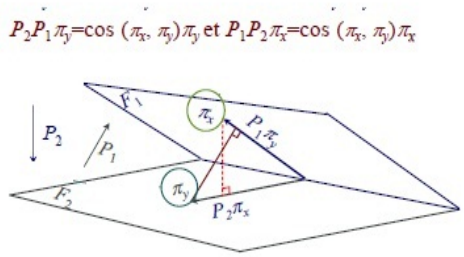

FIGURE 2.1 – Présentation des projections D_p-orthogonales sur les sous-espace F_1 et F_2 dans $\mathbb{R}^n$.

La solution de ce problème est la recherche des valeurs et vecteurs propres de P_1P_2 et P_2P_1

$$\begin{cases} XV_{XX}^{-1}V_{XY}V_{YY}^{-1}Y^tD_p\xi_1 &= \lambda_1\xi_1 \\ YV_{YY}^{-1}V_{YX}V_{XX}^{-1}X^tD_p\varphi_1 &= \lambda_1\varphi_1 \end{cases} \tag{6}$$

ξ_1 et φ_1 sont les premiers vecteurs propres des opérateurs $XV_{XX}^{-1}V_{XY}V_{YY}^{-1}Y^tD_p$ et $YV_{YY}^{-1}V_{YX}V_{XX}^{-1}X^tD_p$ respectivement associés à la même plus grande valeur propre λ_1.

Du système (6) et l'équations (1) on déduit que les premiers facteurs canoniques sont les premiers vecteurs propres des opérateurs $V_{XX}^{-1}V_{XY}V_{YY}^{-1}V_{YX}$ et $V_{YY}^{-1}V_{YX}V_{XX}^{-1}V_{XY}$ respectivement associés à la même plus grande valeur propre $\lambda = 1$

$$\begin{cases} V_{XX}^{-1}V_{XY}V_{YY}^{-1}V_{YX}a_1 &= \lambda_1 a_1 \\ V_{YY}^{-1}V_{YX}V_{XX}^{-1}V_{XY}b_1 &= \lambda_1 b_1 \end{cases} \tag{7}$$

Remarque 2.3.1 • *On peut définir le schéma de dualité pour deux tableaux quantitatifs dé-*

finis sur le même ensemble d'individus comme suit :

$$\begin{array}{rcccl} \mathbb{R}^p \supseteq E_X & \longleftarrow^{X^t} & F^* & \longrightarrow^{Y^t} & E_Y \subseteq \mathbb{R}^q \\ V_{XX}^{-1} \downarrow & & \uparrow N & & \downarrow V_{YY}^{-1} \\ E_X^* & \longrightarrow^{X} & F & \longleftarrow^{Y} & E_Y^* \end{array}$$

- *D'après le schéma de dualité on a :*

$$\forall e_i \in \mathbb{E}_X^*, \quad \forall f_i \in \mathbb{E}_Y^* \quad V_{XX}^{-1} e_i = a_i \quad V_{YY}^{-1} f_i = b_i \tag{8}$$

On déduit de l'équation (8) et système (7) que les vecteurs e_1 et f_1 sont les premiers vecteurs propres des opérateurs $V_{XY}V_{YY}^{-1}V_{YX}V_{XX}^{-1}$ et $V_{YX}V_{XX}^{-1}V_{XY}V_{YY}^{-1}$

$$\begin{cases} V_{XY}V_{YY}^{-1}V_{YX}V_{XX}^{-1}e_1 &= \lambda_1 e_1 \\ V_{YX}V_{XX}^{-1}V_{XY}V_{YY}^{-1}f_1 &= \lambda_1 f_1 \end{cases} \tag{9}$$

les vecteurs e_1 et f_1 sont appelés les premiers axes canoniques.

2.4 Interprétation graphique

2.4.1 Interprétation des variables

Pour donner les nouvelles coordonnées des variables sur les nouveaux axes, on calcule la corréletion des nouvelles variables ξ_i et φ_i avec les anciens de X et Y, on fait présenter sur un même plan l'ensemble des variables de X et Y, on trace un cercle des corrélations, l'axe correspondant à la $k - ime$ étape est la moyenne entre ξ_k et φ_k notée $C_k = \frac{1}{2}(\xi_k + \varphi_k)$ et la variable x^j sur l'axe k, a pour coordonée $r(x^j, C_k)$

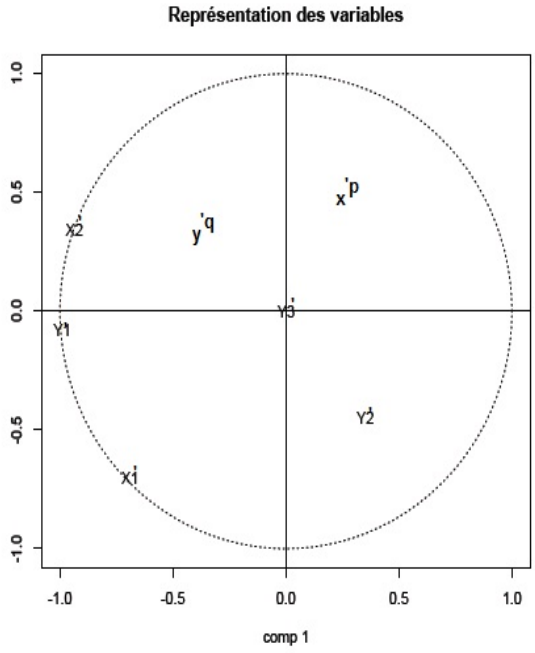

FIGURE 2.2 – Représentation des variables dans le cercle unité.

2.4.2 Interprétation des individus

On a deux choix pour présenter les individus de chaque tableau, soit on choisit la base $\{\xi_1, \cdots, \xi_k\}$ ou la base $\{\varphi_1, \cdots, \varphi_k\}$. La présentation se fait d'une manière analogue à celle de l'ACP. les deux nuages d'individus de chaque tableau sont représentés deux foix. Il s'agit de comparer la description des individus donnée par la variable canonique ξ_k avec la description des individus donnée par la deuxième variable canonique φ_k.

Le graphe fait apparaitre les individus pour lesquels les variables canoniques sont proches et ceux pour lesquels sont éloignés. L'écart résuduel quantifie cet éloignement

$$|x_{1i}^k - x_{2i}^k|$$

x_{1i}^k est la coordonnée de x_i sur l'axe e_k et x_{2i}^k et la coordonnée de x_i sur l'axe f_k

-

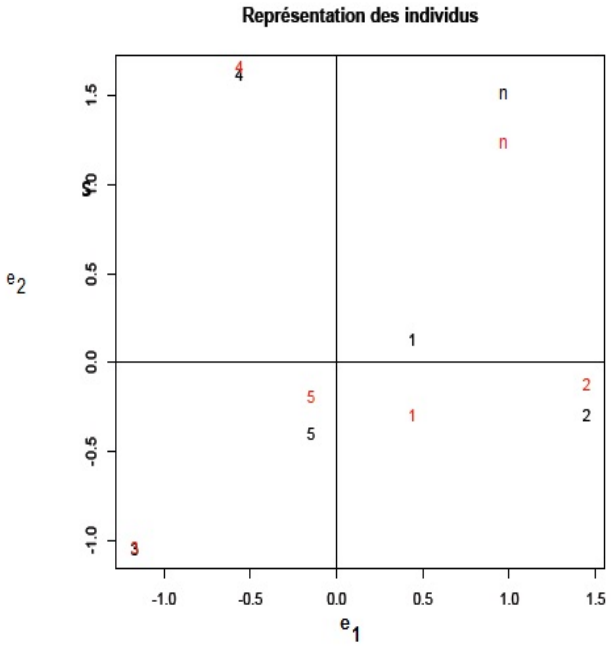

FIGURE 2.3 – Représentation des individus .

Chapitre 3

Analyse factorielle de correspondance

3.1 Introduction

L'analyse factorielle de correspondance a été introduite par J.P.Benzécri (1980 [3] sous le nom d'analyse des correspondances binaires. Elle aété dévloppée en suite par P.CIBOIS[14]. Elle peut aussi être vue comme une ACP d'un tableau de contingence avec une métrique spéciale celle du χ_2. Les domaines d'application de l'AFC sont donc différents de ceux de l'ACP qui est adaptée aux tableaux de mesures quantitatives.
Pour cette analyse aussi nous pouvons donner une longue liste des disciplines ayant trouvé réponse à leur problème par l'AFC. Ainsi, l'écologie, l'économie, et d'autres encore dans lesquelles il peut être intéressant d'étudier les liaisons entre deux variables nominales, ont fourni un grand nombre de données.

3.2 les données

L'AFC étudie le lien (la correspondance) entre deux variables qualitatives X, Y mesurées sur le même ensemble d'individus $I = \{x_1, x_2, \cdots, x_n\}$, la variable X a p modalités $\{x^1, x^2, \cdots, x^p\}$ et la variable Y a q modalités $\{y^1, y^2, \cdots, y^q\}$

Définition 3.2.1 *[10] Un tableau de contingence est un tableau d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus.*

Ce tableau est représenté sous la forme suivante :

	y^1	$\cdots$	y^q	$n_{i.}$
x^1	n_{11}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$n_{k.}$
x^p	n_{p1}	$\cdots$	n_{pq}	$n_{p.}$
$n_{.j}$	$n_{.1}$	$\cdots$	$n_{.q}$	n

(1)

où

- n_{ij} *est le nombre d'individus possédant à la fois la modalité* x^i *de la première variable* X *et la modalité* y^j *de la seconde variable* Y*.*
- $n_{i.} = \sum_{j=1}^{q} n_{ij}$ *est l'effectif de la modalité* x^i *de* X*.*
- $n_{.j} = \sum_{i=1}^{p} n_{ij}$ *est l'effectif de la modalité* y^j *de* Y*.*
- $\sum_{i=1}^{p}\sum_{j=1}^{p} n_{ij} = n.$

Définition 3.2.2 *[24] Le tableau suivant, c'est le tableau des fréquences conjoints notée par* f_{ij} *qui sont données par :*

$$f_{ij} = \frac{n_{ij}}{n}$$

et les fréquences marginales sont données par :

$$f_{.j} = \sum_{i=1}^{p} f_{ij} \qquad f_{i.} = \sum_{j=1}^{q} f_{ij} \text{ et } \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1$$

	y^1	$\cdots$	y^q	$f_{i.}$
x^1	f_{11}	$\cdots$	f_{1q}	$f_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$f_{k.}$
x^p	f_{p1}	$\cdots$	f_{pq}	$f_{p.}$
$f_{.j}$	$f_{.1}$	$\cdots$	$f_{.q}$	1

(2)

3.2.1 L'objectif de l'AFC

l'AFC étudie un tableau de contingence ou de fréquence pour déterminer les liaisons entre les deux variables qualitatives X et Y. Pour l'ACP ces liaisons sont définis par les coefficients de corrélation par contre pour l'AFC, on dit qu'il y a indépendance entre les deux variables considérées si,

$$f_{ij} = f_{i.}f_{.j}, \forall i \in I, \forall j \in J$$

Nous disons qu'il y a une relation entre ces deux variables, ou que ces deux variables sont liées si elles ne sont pas indépendantes.

Les objectifs donc sont les mêmes que ceux de l'ACP dans le sens où l'AFC cherche donc à obtenir une structure simple des lignes et une structure simple des colonnes, puis de relier ces deux typologies.

3.3 Le principe de l'AFC

On a deux présentations de tableau de fréquence, On peut le considérer comme un nuage des lignes et comme un nuage des colonnes.

Lorsque le tableau est considéré en ligne les données sont normalisées en divisant par $f_{i.}$, la nouvelle ligne ainsi crée est appelée profil-ligne. Cette normalisation a pour but de considérer les liaisons entre les deux variables à travers de l' ?cart entre les pourcentages en lignes.

$$x^i = \begin{pmatrix} \frac{f_{i1}}{f_{i.}} \\ \vdots \\ \frac{f_{iq}}{f_{i.}} \end{pmatrix} \quad \forall i = 1, \cdots, p$$

Si le tableau de fréquences est considéré comme un nuage colonnes, les données sont normalisées, nous divisons par $f_{.j}$ la nouvelle colonne est dite profile-colonne

$$y^j = \begin{pmatrix} \frac{f_{1j}}{f_{.j}} \\ \vdots \\ \frac{f_{pj}}{f_{.j}} \end{pmatrix}$$

Ainsi il y a indépendance lorsque les lignes du tableau de fréquences sont proportionnelles. Par symétrie il en est de même pour les colonnes.

3.3.1 La ressemblance entre les profils

La ressemblance entre deux lignes ou entre deux colonnes est définie par une distance entre profils. La distance employée est celle du χ_2 et elle est définie de façon symétrique. pour les lignes et les colonnes. Ainsi entre deux lignes x^i et x^k elle est donnée par :

$$d_{\chi_2}(x^i, x^k) = \sum_{j \in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{kj}}{f_{k.}} \right)^2 \tag{3}$$

et entre deux colonnes y^j et y^h par :

$$d_{\chi_2}(y^j, y^h) = \sum_{i \in I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ih}}{f_{.h}} \right)^2 \tag{4}$$

La matrice diagonale

$$D_I = \begin{pmatrix} \frac{1}{f_{1.}} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_{2.}} & 0 \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{f_{p.}} \end{pmatrix}$$

est la matrice de poids, définie la métrique dans $\mathbb{R}^I$, tandis que

$$D_J = \begin{pmatrix} \frac{1}{f_{.1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_{.2}} & 0 \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{f_{.q}} \end{pmatrix}$$

définie celle dans $\mathbb{R}^J$.

3.3.2 Les nuages des profils

Le nuage des profils-lignes

Lorsque nous nous intéressons aux modalités de la première variable X, il faut considérer les données comme une juxtaposition de profils-lignes. Ainsi, chaque profil-ligne i peut être représenté comme un point de l'espace $\mathbb{R}^q$ dont chacune des q coordonnées représente une modalité de la seconde variable. L'utilisation de la distance entre deux profils est celle de χ_2, elle revient à affecter le poids $\frac{1}{f_{.j}}$ à la $j-me$ coordonnée de $\mathbb{R}^q$.

Pour l'AFC, les poids affectés à chaque point du nuage sont imposés et ne sont pas identiques. Le point x^i a pour poids la fréquence marginale $f_{i.}$. Ce poids est naturel puisqu'il est proportionnel à l'effectif de la classe d'individus qu'il représente. La coordonnée du point x^i sur l'axe j est donné par $\frac{f_{ij}}{f_{i.}}$.

Le centre de gravité de nuage des profils-lignes munis de ces poids, noté g_X , est la moyenne pondérée de tous les points sur tous les axes j. La coordonnée de g_X sur l'axe j est donc donnée par :

$$\sum_{i=1}^{p} f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j}$$

Donc le centre de gravité est le vecteur dans $\mathbb{R}^q$ suivant :

$$g_X = (f_{.1}, f_{.2}, \cdots, f_{.q})$$

Le nuage des profils-colonnes

Le nuage des profils-colonnes est constitué d'une facçon identique à celle du nuage des profils-lignes, lorsque nous nous intéressons aux modalités de la seconde variable Y, on considère les données comme une juxtaposition de profils-colonnes. Chaque profil-colonne y^j peut être représenté comme un vecteur dans l'espace $\mathbb{R}^p$ dont chacune des p coordonnées représente une modalité de la première variable X. Le profile y^j a pour coordonnée sur l'axe k la proportion $\frac{f_{jk}}{f_{.k}}$, et le poids qui lui est associé est $f_{.j}$.

Le centre de gravité g_Y de nuage des profils-colonnes munis de leur poids a pour *ime*

coordonnée :

$$\sum_{j=1}^{q} f_{.j} \frac{f_{ij}}{f_{.j}} = f_{i.}$$

Donc le centre de gravité de nuage de profiles-colonnes est le vecteur dans $\mathbb{R}^p$ défini par :

$$g_Y = (f_{1.}, f_{2.}, \cdots, f_{p.})$$

3.4 L'analyse factorielle de correspondances

3.4.1 L'ACP de nuage de profils-lignes

Soit le tableau Z_1 qui a pour lignes les profiles-lignes défini comme suit

$$Z_1 = \begin{pmatrix} \frac{f_{11}}{f_{1.}} & \frac{f_{12}}{f_{1.}} & \cdots & \frac{f_{1q}}{f_{1.}} \\ \frac{f_{21}}{f_{2.}} & \frac{f_{22}}{f_{2.}} & \cdots & \frac{f_{2q}}{f_{2.}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p.}} & \frac{f_{p2}}{f_{p.}} & \cdots & \frac{f_{pq}}{f_{p.}} \end{pmatrix}$$

Le schéma de dualité associé à Z_1 est le suivant :

$$\begin{array}{rcl} \mathbb{R}^q \supseteq E_1 & \longleftarrow^{Z_1^t} & F_1 \subseteq \mathbb{R}^p \\ M_X \downarrow & & \downarrow D_X \\ E_1^* & \longrightarrow^{Z_1} & F_1^* \end{array}$$

où

- E_1 est le sous-espace de $\mathbb{R}^q$ engendré par les lignes de Z_1, cet espace joue le rôle de l'espace d'individus pour l'ACP.
- F_1 est le sous-espace de $\mathbb{R}^p$ engendré par les colonnes de Z_1, cet espace joue le rôle de l'espace de variables pour l'ACP.
- M_X est la métrique définie dans E_1, c'est une matrice symétrique de dimension $p \times p$ dont les coordonnées sont les distances de χ_2 entre les lignes de Z_1 deux à deux, elle est définie

comme suit :

$$M_X = \begin{pmatrix} \chi_2(x^1, x^1) & \chi_2(x^1, x^2) & \cdots & \chi_2(x^1, x^p) \\ \chi_2(x^2, x^1) & \chi_2(x^2, x^2) & \cdots & \chi_2(x^2, x^p) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_2(x^p, x^1) & \chi_2(x^p, x^2) & \cdots & \chi_2(x^p, x^p) \end{pmatrix}$$

L'AFC du tableau Z_1 est équivalent à l'ACP de (Z_1, M_X, D_X), donc notre problème revient la recherche d'une suite d'axes orthonormés $\{u_i, i = 1, \cdots, k\}$ qui constituent un plan de dimension $k < q$ sur lequel le nuage de profils-lignes est projeté. Chaque axe u_i doit rendre maximum l'inertie du nuage profils-lignes projeté . En pratique, nous devons centrer ce nuage (ce nuage est normé), ainsi le centre de gravité G_X devient l'origine des axes. Une fois le nuage centré, le profile ligne x^i devient un vecteur dans $\mathbb{R}^q$ défini par :

$$x^i = (\frac{f_{i1}}{f_{i.}} - f_{.1}, \frac{f_{i2}}{f_{i.}} - f_{.2}, \cdots, \frac{f_{iq}}{f_{i.}} - f_{.q})$$

et le tableau de profile-lignes centré et réduit s'écrit sous la forme :

$$Z_1 = \begin{pmatrix} \frac{f_{11}}{f_{.}} - f_{.1} & \frac{f_{12}}{f_{1.}} - f_{.2} & \cdots & \frac{f_{1q}}{f_{.}} - f_{.q} \\ \frac{f_{21}}{f_{2.}} - f_{.1} & \frac{f_{22}}{f_{2.}} - f_{.2} & \cdots & \frac{f_{2q}}{f_{2.}} - f_{.q} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p.}} - f_{.1} & \frac{f_{p2}}{f_{p.}} - f_{.2} & \cdots & \frac{f_{pq}}{f_{p.}} - f_{.q} \end{pmatrix}$$

La recherche des axes qui rendent maximum l'inertie du nuage centré et réduit et chaque profil-ligne étant muni d'un poids $f_{i.}$, l'inertie est donnée par :

$$\sum_{i=1}^{p} f_{i.} \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$$

Par une démarche semblable à celle de l'ACP, on cherche le premier vecteur unitaire u_1 qui rende cette inertie maximale, le deuxième vecteur unitaire u_2 orthogonal à u_1 qui vérifie le même critère, etc. à l'exception du fait que les lignes interviennent à travers de leur profil, que la distance entre les profils est celle du χ_2 et que chaque élément i est affecté d'un poids $f_{i.}$, donc la solution est la réalisation d'une analyse spectrale de l'opérateur $Z_1^t D_X Z_1 M_X$.

3.4.2 L'ACP de nuage de profils-colonnes

Puisqu'en AFC les lignes et les colonnes jouent un rôle symétrique, l'ajustement du nuage des profils-colonnes est semblable à celui de nuage de profiles-lignes. Pour cela, on définit le tableau des profiles-colonnes centré et réduit par :

$$Z_2 = \begin{pmatrix} \frac{f_{11}}{f_{.1}} - f_{1.} & \frac{f_{21}}{f_{.1}} - f_{2.} & \cdots & \frac{f_{p1}}{f_{.1}} - f_{p.} \\ \frac{f_{12}}{f_{.2}} - f_{1.} & \frac{f_{22}}{f_{.2}} - f_{2.} & \cdots & \frac{f_{p2}}{f_{.2}} - f_{p.} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f_{1q}}{f_{.q}} - f_{1.} & \frac{f_{2q}}{f_{.q}} - f_{.2} & \cdots & \frac{f_{pq}}{f_{.q}} - f_{p.} \end{pmatrix}$$

L'inertie de nuage de profiles-colonnes est donnée par :

$$\sum_{j=1}^{q} f_{.j} \sum_{i=1}^{p} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2$$

Le schéma de dualité associé à Z_2 est le suivant :

$$\begin{array}{rcl} \mathbb{R}^p \supseteq E_2 & \longleftarrow^{Z_2^t} & F_2 \subseteq \mathbb{R}^q \\ M_Y \downarrow & & \downarrow D_Y \\ E_2^* & \longrightarrow^{Z_2} & F_2^* \end{array}$$

où

- E_2 est un sous-espace de $\mathbb{R}^p$ engendré par les lignes de Z_2, cet espace joue le rôle de l'espace d'individus pour l'ACP.
- F_2 est un sous-espace de $\mathbb{R}^q$ engendré par les colonnes de Z_2, cet espace joue le rôle de l'espace de variables pour l'ACP.
- M_Y est la métrique définie sur E_1, c'est une matrice symétrique de dimension $q \times q$ dont les cordonnées sont la distance de χ_2 entre les lignes de Z_2 deux à deux, elle est définie comme suit :

$$M_Y = \begin{pmatrix} \chi_2(y^1, y^1) & \chi_2(y^1, y^2) & \cdots & \chi_2(y^1, y^q) \\ \chi_2(y^2, y^1) & \chi_2(y^2, y^2) & \cdots & \chi_2(y^2, y^q) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_2(y^q, y^1) & \chi_2(y^q, y^2) & \cdots & \chi_2(y^q, y^q) \end{pmatrix}$$

L'AFC de tableau Z_2 est équivalent à l'ACP de (Z_2, M_Y, D_Y) c'est à dire que l'analyse spectrale de l'opérateur $Z_2^t D_Y Z_2 M_Y$

3.4.3 Interprétation

L'ACP de nuage de profils-lignes et celle de profils-colonnes sont équivalentes et nous donnent les mêmes résultats. La représentation graphique se fait sur celle de l'ACP avec les mêmes règles d'interprétation.

3.5 Conclusion

L'ACP et l'AFC sont diférentes en plusieurs points, elles fournissent des éclairages complémentaires. L'AFC est une méthode puissante pour synthétiser et résumer de vastes tableaux de contingence. En pratique elle est appliquée à beaucoup d'autres tableaux, notamment les tableaux individus-variables. Les individus sont alors considérés comme des variables.

Dans le cas de tableaux de contingence, le principal objectif de cette analyse est de dégager les liaisons entre deux variables. L'analyse des correspondances multiples que nous exposons dans le chapitre suivant permet l'étude des liaisons entre plus de deux variables.

Chapitre 4

Analyse de correspondances multiples

4.1 Introduction

l'AFCM est une généralisation de l'AFC à plusieurs variables qualitatives. Cette analyse a particulièrement été étudiée par B. Escofer et J.PAGES(1988)[24].

Cette analyse très simple est non plus adaptée aux tableaux de contingence de l'AFC, mais aux tableaux disjonctifs complets que nous décrivons ci-dessous. Ces tableaux sont des tableaux logiques pour des variables codées. Les propriétés de tels tableaux font de l'AFCM une méthode spécifique aux règles d'interprétation des représentations simples. Elle permet donc l'étude des liaisons entre plus de deux variables qualitatives, ce qui étend le spectre d'étude de l'AFC. L'A FCM est donc trés bien adaptée au traitement d'enquêtes lorsque les variables sont qualitatives. Il est également possible de n'appliquer cette méthode plusieurs fois en ne prenant en compte que quelques variables.

4.2 les données

L'ACM permet l'étude de tableaux décrivant une population de I individus et J variables qualitatives. Une variable qualitative peut être décrite par une application de l'ensembles des I individus dans un ensemble fini non structuré, par exemple non ordonné. Les données peuvent donc être représentées de façon classique à l'aide d'un tableau d'effectif, cependant deux autres représentations sont également utilisées : le tableau disjonctif complet (T.D.C) et le tableau de Brut (T.B)

Définition 4.2.1 *[27] Le tableau disjonctif complet comporte une colonne pour chaque modalité des variables étudiées et une ligne pour chaque individu, les céllules du tableau contiennent 1*

ou 0 selon que l'individu considéré présente la modalité correspondante ou non.

Exemple 1 Sur 7 individus on observe 3 variables qualitative, x^1 *a 3 modalités* $\{x_1^1, x_2^1, x_3^1\}$*,* x_2 *a 3 modalités et* x^3 *a 4 modalités* $\{x_1^3, x_2^3, x_3^3, x_4^3\}$*, tels que :*

ind_1	x_1^1	x_1^2	x_4^3
ind_2	x_2^1	x_1^2	x_1^3
ind_3	x_3^1	x_3^2	x_3^3
ind_4	x_2^1	x_1^2	x_3^3
ind_5	x_3^1	x_3^2	x_4^3
ind_6	x_1^1	x_3^2	x_1^3
ind_7	x_3^1	x_2^2	x_2^3

Le tableau disjonctif complet de cet tableau est le suivant :

	x_1^1	x_2^1	x_3^1	x_1^2	x_2^2	x_3^2	x_1^3	x_2^3	x_3^3	x_4^3	$x_{i.}$
ind_1	*1*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*1*	$x_{1.} = 3$
ind_2	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*0*	$x_{2.} = 3$
ind_3	*0*	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*1*	*0*	$x_{3.} = 3$
in_4	*0*	*1*	*0*	*1*	*0*	*0*	*0*	*0*	*1*	*0*	$x_{4.} = 3$
ind_5	*0*	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*0*	*1*	$x_{5.} = 3$
ind_6	*1*	*0*	*0*	*0*	*0*	*1*	*1*	*0*	*0*	*0*	$x_{6.} = 3$
ind_7	*0*	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*0*	*0*	$x_{7.} = 3$

Définition 4.2.2 *[27] Le tableau de Brut est un tableau qui contient une ligne et une colonne pour chaque modalité des variables étudiées, chaque céllule du tableau indique le nombre d'individus qui possèdent à la fois la modalité ligne et la modalité colonne correspondante, c'est un tableau symétrique et un tableau en blocs.*

Exemple 2 Le tableau de Brut de l'exemple1 est donné par :

	x_1^1	x_2^1	x_3^1	x_1^2	x_2^2	x_3^2	x_1^3	x_2^3	x_3^3	x_4^3
x_1^1	*2*	*0*	*0*	*1*	*0*	*1*	*1*	*0*	*0*	*1*
x_2^1	*0*	*2*	*0*	*2*	*0*	*0*	*1*	*0*	*1*	*0*
x_3^1	*0*	*0*	*3*	*0*	*1*	*2*	*0*	*1*	*1*	*1*
x_1^2	*1*	*2*	*0*	*3*	*0*	*0*	*1*	*0*	*1*	*1*
x_2^2	*0*	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*0*	*0*
x_3^2	*1*	*0*	*2*	*0*	*0*	*3*	*1*	*0*	*1*	*1*
x_1^3	*1*	*1*	*0*	*1*	*0*	*1*	*2*	*0*	*0*	*0*
x_2^3	*0*	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*0*	*0*
x_3^3	*0*	*1*	*1*	*1*	*0*	*1*	*0*	*0*	*2*	*0*
x_4^3	*1*	*0*	*1*	*1*	*0*	*1*	*0*	*0*	*0*	*2*

4.2.1 l'AFCM d'un tableau disjonctif complet

Comme pour l'AFC, nous allons considérer le tableau disjonctif complet en profils-lignes et en profils-colonnes. Pour se faire nous modifions ce tableau pour considérer les fréquences. Les fréquences f_{ik} sont données par

$$\frac{x_{ik}}{np}.$$

Les marges sont données par :

$$f_{i.} = \sum_{k=1}^{K} \frac{x_{ik}}{np} = \frac{1}{n}$$

avec $x_{ik} \in \{1, 0\}$, et K est le nombre de toutes les modalités de toutes les variables étudiées.

$$f_{.k} = \sum_{i=1}^{n} \frac{x_{ik}}{np} = \frac{I_k}{np}$$

Le tableau de fréquences est alors défini comme suit :

$inds/modalit$	x_1^1	$\cdots$	$x_1^{n_1}$	$\cdots$	x_p^1	$\cdots$	$x_p^{n_p}$	$f_{i.}$
ind_1	$\frac{x_{11}}{np}$	$\cdots$	$\frac{x_{1n_1}}{np}$	$\cdots$	$\frac{x_{11_p}}{np}$	$\cdots$	$\frac{x_{1n_p}}{np}$	$\frac{1}{n}$
ind_2	$\frac{x_{21}}{np}$	$\cdots$	$\frac{x_{2n_1}}{np}$	$\cdots$	$\frac{x_{21_p}}{np}$	$\cdots$	$\frac{x_{2n_p}}{np}$	$\frac{1}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ind_n	$\frac{x_{n1}}{np}$	$\cdots$	$\frac{x_{nn_1}}{np}$	$\cdots$	$\frac{x_{n1_p}}{np}$	$\cdots$	$\frac{x_{nn_p}}{np}$	$\frac{1}{n}$
$f_{.j}$	$\frac{I_1}{np}$	$\cdots$	$\frac{I_{n_1}}{np}$	$\cdots$	$\frac{I_{1_p}}{np}$	$\cdots$	$\frac{I_{n_p}}{np}$	

Nuage de profiles-lignes

Chaque individu du nuage des individus est représenté par les modalités qu'il possède. Le profils-lignes x_i est un vecteur dans $\mathbb{R}^K$ défini par

$$x_i = (\frac{x_{i1}}{np}n, \cdots, \frac{x_{iK}}{np}n)^t = (\frac{x_{i1}}{p}, \cdots, \frac{x_{iK}}{p})^t$$

avec un poids identique pour chaque individu (car la marge est constante) de $\frac{1}{n}$.

Le centre de gravité G du nuage d'individus (profiles-lignes) a pour k-ième coordonnée

$$\sum_{i=1}^{n} \frac{x_{ik}}{p}.\frac{1}{n} = \frac{I_k}{n.p} = f_{.k}$$

alors

$$g = (f_{.1}, \cdots, f_{.K})$$

La ressemblance entre deux individus (deux profiles-lignes) est caractérisée par La distance de Khi-deux qui est quantifiée par :

$$d^2(x_i, x_l) = \sum_{k=1}^{K} \frac{np}{I_k} \left(\frac{x_{ik}}{p} - \frac{x_{lk}}{p} \right)^2 = \frac{n}{p} \sum_{k=1}^{K} \frac{1}{I_k} (x_{ik} - x_{lk})^2 \tag{1}$$

Cette expression est remarquable car $(x_{ik} - x_{lk})^2 = 1$ si un seul individu possède la modalité k et 0 sinon. Cette distance croît logiquement avec le nombre de modalités qui diffèrent pour les individus i et l, ce qui est recherché. Le poids de la modalité k dans la distance est l'inverse de sa fréquence : $\frac{n}{I_k}$. Ainsi si un individu possède une modalité rare, il sera éloigné de tous les autres individus et du centre de gravité. LAFCM du tableau de profiles-lignes est équivalente à l'ACP de (X, D_X, M_X) c'est-à- dire l'analyse spectrale de l'opérateur $X^t D_X X M_X$ où

– X est le tableau dont les lignes sont les profiles-lignes défini par

$$X = \begin{pmatrix} \frac{x_{11}}{p} & \cdots & \frac{x_{1K}}{p} \\ \frac{x_{21}}{p} & \cdots & \frac{x_{2K}}{p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}}{p} & \cdots & \frac{x_{nK}}{p} \end{pmatrix}$$

– D_X est la métrique de poids donnée par :

$$D_X = \begin{pmatrix} \frac{1}{f_{1.}} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_{2.}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{f_{.n}} \end{pmatrix}$$

– M_X est la matrice définie par la métrique de χ_2 définie comme suit :

$$M_X = \begin{pmatrix} \chi_2(x_1, x_1) & \chi_2(x_1, x_2) & \cdots & \chi_2(x_1, x_n) \\ \chi_2(x_2, x_1) & \chi_2(x_2, x_2) & \cdots & \chi_2(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_2(x_n, x_1) & \chi_2(x_n, x_2) & \cdots & \chi_2(x_n, x_n) \end{pmatrix}$$

L'AFCM de nuage de profiles-colonnes

Chaque modalité peut être représentée par un profil-colonne, c'est-à-dire par les valeurs prises par tous les individus pour la modalité considérée. Ainsi une modalité k est un vecteur dans $\mathbb{R}^n$ a pour k-ième coordonnée

$$\frac{f_{ik}}{f_{.k}} = \frac{x_{ik}}{I_k} \quad \forall k = 1, \cdots, K$$

alors chaque profile-colonne noté $P(k)$ est donné par

$$P(k) = (\frac{x_{1k}}{I_k}, \cdots, \frac{x_{nk}}{I_k})^t$$

La k-ième coordonnée de centre de gravité G_K du nuage des modalités (profiles-colonnes) est donnée par $\sum_{i=1}^{n} \frac{x_{ik}}{I_k}\frac{I_k}{np} = \frac{\sum_{i=1}^{n} x_{ik}}{np}$ alors

$$G_K = \left(\frac{\sum_{i=1}^{n} x_{i1}}{np}, \cdots, \frac{\sum_{i=1}^{n} x_{iK}}{np}\right)^t$$

La ressemblance entre deux modalitś k et h est donnée par la distance :

$$d^2(P(k), P(h)) = \sum_{i=1}^{n} n \left(\frac{x_{ik}}{I_k} - \frac{x_{ih}}{I_h}\right)^2 \tag{2}$$

En notant que $(x_{ik})^2 = x_{ik}$ qui ne prennent que les valeurs 1 ou 0, cette distance peut s'écrire :

$$d^2(k, h) = \frac{n}{I_k I_h}\left(I_k + I_h - 2\sum_{i \in I} x_{ik}x_{ih}\right) \tag{3}$$

ce qui est le nombre d'individus possédant une et une seule des deux modalités h ou k multiplié par$\frac{n}{I_k I_h}$. Cette distance croît donc avec le nombre d'individus possèdant une et une seule des deux modalités k et h et décroît avec l'effectif de chacune de ces modalités. Ainsi, par construction, deux modalités d'une même variable sont éloignées l'une de l'autre (puisqu'elles ne peuvent pas être possédées par le même individu). Deux modalités possédées par exactement

les mêmes individus sont confondues, tandis que les modalités rares sont éloignées de toutes les autres et du centre de gravité G_K.

D'une manière similaire que le nuage de profiles-lignes, on définit le tableau Y qui a pour lignes les profiles-colonnes et la matrice diagonale de poids D_Y dont le diagonale $\frac{1}{f_{.k}}$ pour $k = 1, \cdots, K$ et la matrice M_Y définie par les distances entre les profiles-colonnes deux-à-deux.L'AFCM de tableau de profiles-colonnes est équivalente à l'ACP de (Y, D_Y, M_Y).

Remarque 4.2.1 *L'AFCM d'un tableau de Brut est équivalent à l'AFC de tableau de Brut c'est-à-dire l'ACP de chaque bloc de cet tableau avec les métriques D_i de poids et M_i la métrique de χ_2*

4.3 Conclusion

L'AFCM est donc une analyse factorielle qui permet l'étude de plusieurs variables qualitatives, de ce fait elle est une généralisation de l'AFC. Elle est donc applicable aux tableaux de variables qualitatives. Le fait de pouvoir interpréter l'AFCM de plusieurs façons rend cette méthode trés riche et d'emploi facile. Elle peut être très complémentaire de l'ACP .

Chapitre 5

Analyse discriminante

5.1 Introduction

L'analyse factorielle discriminante a été introduite par J.M.ROMEDER [35], c'est une méthode comune entre les méthodes factorielles et les méthodes de classification supervisée qui nécessitent une connaissance des classes à prioré.Cette méthode traite les tableaux des données mixes qui contiennent des variables quantitatives et une variable qualitative. L'analyse factorielle discriminante est une méthode descriptive et prédictive fondée sur un modèle paramétrique. Elle est également appelée analyse linéaire discriminante (Linear Analysis Discriminant (LDA) en anglais). En effet, cette méthode peut être vue comme une analyse factorielle, car son aspect descriptif fait appel à des calculs d'axes principaux. C'est une méthode avant tout prédictive qui discrimine les individus selon des classes connues. Son aspect prédictif de classement de nouveaux individus peut en fait faire appel à d'autres méthodes de classification géométriques ou probabilistes.

5.2 les données

Soit $I = \{x_1, x_2, \cdots, x_n\}$ un ensemble de n individus ou observations sur lequel on observe un ensemble de p variables quantitatives $\{x^1, x^2, \cdots, x^p\}$ et une variable qualitative Y possédant k modalités, on va construire k classes selon les modalités de la variable qualitative y . Les k classes sont à priori connues.On note par x_{ij}^l la valeur de la variable x^j pour l'individu x_i dans le groupe c_l, et n_l le cardinal du groupe c_l, $l \in \{1, \cdots, k\}$ $\sum_{l=1}^{k} n_l = n$.

Le tableau X est la juxtraposition de de k groupes des données :

$$X = \begin{pmatrix} x_{11}^1 & \cdots & x_{1p}^1 \\ \vdots & \ddots & \vdots \\ x_{n_1 1}^1 & \cdots & x_{n_1 p}^1 \\ x_{11}^2 & \cdots & x_{1p}^2 \\ \vdots & \ddots & \vdots \\ x_{n_2 1}^2 & \cdots & x_{n_2 p}^2 \\ \vdots & \ddots & \vdots \\ x_{11}^k & \cdots & x_{1p}^k \\ \vdots & \ddots & \vdots \\ x_{n_k 1}^k & \cdots & x_{n_k p}^k \end{pmatrix} \tag{1}$$

5.3 L'objectif de l'AFD

L'analyse discriminante a deux objectifs différents :

- Le premier objectif est descriptif (factoriel) repose sur le choix d'un plan sur lequel on représente les k groupes de tel sorte que les projections de k centres de gravité sont les plus éloignés possible d'une part, d'autre part, les projections des éléments de chaque classe sont les plus proches possibles de la projection de son centre de gravité (des classes homogènes). Autrement dit, la recherche d'un plan qui maximise l'inertie inter-classes (variance inter-classes) et minimise l'inertie intra-classes (variance intra-classes).
- Le second objectif est un objectif de classement, c'est la réponse de la question suivante : Peut-on déterminer le groupe d'appartenance d'une nouvelle observation (nouvel individu) à partir des p mesures quantitatives ? Cet objectif est un objectif d'affectation d'un nouvel individu dans une classe. Il s'agit d'un problème de classement par opposition au problème de classification qui est la construction de classes les plus homogènes possibles dans un échantillon.

5.4 Principe de l'AFD

5.4.1 Aspect descriptif

L'idée de la discrimination repose sur le choix d'un plan sur lequel on représente les k groupes simultanément de telle sorte que :

- •Les projections de k centres de gravité sont éloignées,
- chaque sous-nuage appartenant à une seule classe sont les plus homogènes possibles autour de ces centres de gravité.

Pour ce faire il faut maximiser les variances interclasses (entre les classes) et minimiser les variances intraclasses (à l'intérieur des classes). Nous parlons également de variances externes et internes. La figure (5.1) représente un nuage de n individus partagé en trois classes dans l'espace $\mathbb{R}^p$. Notons n_q le nombre d'individus dans la classe q et l'ensemble des individus de la classe q, $c_q = \{x_1, \cdots, x_{n_q}\}$. G représente le centre de gravité global du nuage des individus dans $\mathbb{R}^p$, et g_q le centre de gravité de la classe c_q qui est donné par le vecteur :

$$g_q = (\frac{1}{n_q}\sum_{i=1}^{n_q} x_{i1}, \cdots, \frac{1}{n_q}\sum_{i=1}^{n_q} x_{ip}) \tag{2}$$

Définition 5.4.1 *[23] Soient $c_1, \cdots c_k$ k groupes qui forment une partition de l'ensemble d'individus I dont les centres de gravités sont $g_1, \cdots, g_k$. La matrice de variance-covariance intraclasse est définie par :*

$$D = \frac{1}{n}\sum_{q=1}^{k}\sum_{i=1}^{n_q}(x_i - g_q)(x_i - g_q)^t \tag{3}$$

Définition 5.4.2 *[23] La matrice de variance-covariance interclasse est donnée par :*

$$E = \frac{1}{n}\sum_{l=1}^{k} n_q (g_q - G)(g_q - G)^t \tag{4}$$

Proposition 5.4.1 *[23] La décomposition de Huygens nous donne*

L'inertie totale du nuage d'individus est égale à la somme de l'inertie interclasse et l'inertie intraclasse.

Cette proposition montre également que la matrice de variance-covariance totale T du nuage est la somme des deux matrices de la variance-covariance interclasse et de la variance-covariance intraclasse :

$$T = D + E \tag{5}$$

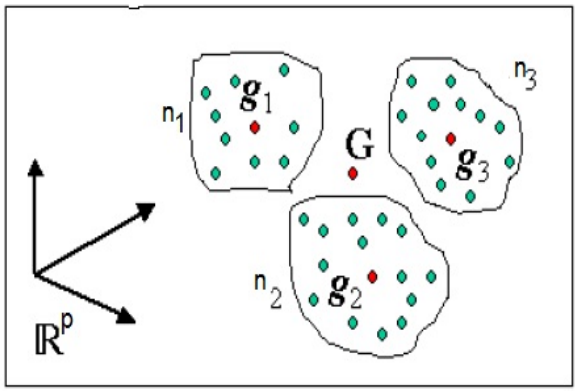

FIGURE 5.1 – Représentation du nuage des individus partitionnés dans l'espace $\mathbb{R}^p$.

La figure (5.2) illustre cette proposition. Le même nuage est représenté deux fois en reliant les points pour le calcul de la variance-covariance totale à gauche et de la somme des variance-covariances interclasse et intraclasse à droite.

Preuve 5.4.2 *La matrice de covariance totale est donnée par :*

$$v_{kk'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - G_k)(x_{ik'} - G_{k'}) = \frac{1}{n}\sum_{q=1}^{k}\sum_{i=1}^{n}(x_{ik} - G_k)(x_{ik'} - G_{k'}) \tag{6}$$

où

$$G_k = \frac{1}{n}x_{ik} \tag{7}$$

Or

$$(x_{ik} - G_k) = (x_{ik} - g_{qk}) + (g_{qk} - G_k) \tag{8}$$

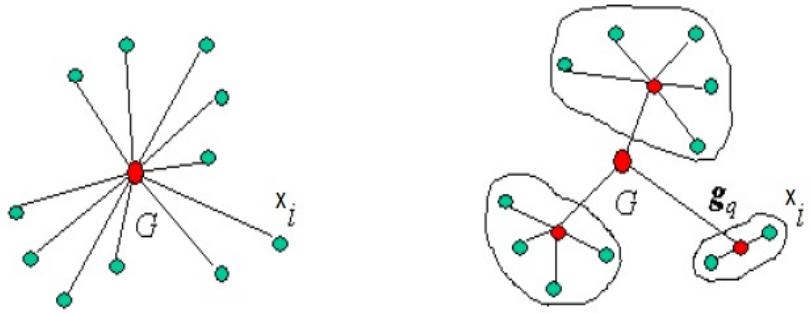

FIGURE 5.2 – Illustration de la formule de Huygens.

nous remarquons ainsi que

$$\sum_{i=1}^{n_q} (x_{ik} - g_{qk}) \left(g_{qk'} - G_{k'}\right) = \sum_{1}^{n_q} (g_{qk} - G_k) \left(x_{ik'} - g_{qk'}\right) = 0 \tag{9}$$

Donc uniquement deux des quatre termes de la partie droite de l'?quation (6) sont non nuls et nous pouvons écrire :

$$v_{kk'} = b_{kk'} + w_{kk'} \tag{10}$$

avec

$$b_{kk'} = \frac{1}{n} \sum_{q=1}^{k} n_q \left(g_{qk} - G_k\right) \left(g_{qk'} - G_{k'}\right) \tag{11}$$

et

$$w_{kk'} = \frac{1}{n} \sum_{q=1}^{k} \sum_{i=1}^{n_q} \left(x_{ik} - g_{qk}\right) \left(x_{ik'} - g_{qk'}\right) \tag{12}$$

ce qui démontre la proposition. ■

L'analyse du problème

L'AFD consiste à trouver des nouveaux axes tels que les projections des k centres de gravité sur ces axes doivent être les plus éloignées possible, tandis que les projections de chaque sous-nuage sur ces axes doivent être les plus concentrées autour des projections des leurs centres de gravité. La démarche à suivre est :

1. La recherche du premier axe est donc celui qui maximise la variance-covariance interclasse et minimise la variance-covariance intraclasse.
2. la recherce du deuxième axe est celui qui est non corrélée à la première et qui discrimine au mieux les classes au sens du même critère (maximisation de la variance-covariance interclasse et minimisation de la variance-covariance intraclasse). Les autres axes sont déterminés de la même façon.

La recherche du premier vecteur séparant le mieux possible les k groupes

Soit u_1 un vecteur sur lequel seront éffectués les projections des individus de X.

Si on note par X_c la matrice centrée de la matrice X , N la matrice de poids sur $\mathbb{R}^n$ et $M = \mathbb{1}_p$,alors

$$T = X_c^t N X_c = \frac{1}{n} X_c^t X_c \tag{13}$$

Les coordonnées des projections des individus de X sur l'axe qui engendré par le vecteur u_1 sont données par la composante C

$$C = X_c M u_1$$

La variance de ces projections est donnée par

$$N(C,C) = C^t N C = \frac{1}{n}(X_c M u_1)^t X_c M u_1 = u_1^t T u_1 = T(u_1, u_1) \tag{14}$$

D'après l'équation (5), on a :

$$var(C) = (u_1^t E u_1) + (u_1^t D u_1) \tag{15}$$

Le problème de l'analyse discriminante est alors la recherche d'un vecteur u_1 qui maximise la quantité $u_1^t E u_1$ et minimise la quantité $(u_1^t D u_1)$ sous la contrainte de normalisation $u_1^t T u_1 = 1$. Ce problème est équivalent à la recherche de u_1 le maximum de

$$\frac{(u_1^t E u_1)}{(u_1^t T u_1)} \tag{16}$$

et le minimum de

$$\frac{(u_1^t D u_1)}{(u_1^t T u_1)} \tag{17}$$

D'après Lagrange, on a :

$$\begin{aligned} \frac{\partial}{\partial u_1}\left(\frac{u_1^t E u_1}{u_1^t T u_1}\right) = 0 \quad &\Leftrightarrow \quad \frac{\partial}{\partial u_1}\left(\frac{u_1^t E u_1}{u_1^t T u_1}\right) - \frac{\partial}{\partial u_1}(\lambda_1(u_1 T u_1) - 1) = 0 \\ &\Leftrightarrow \quad 2Eu_1 - 2\lambda_1 T u = 0 \\ &\Leftrightarrow \quad Eu_1 = \lambda_1 T u_1 \\ &\Leftrightarrow \quad T^{-1}Eu_1 = \lambda_1 u_1 \end{aligned}$$

On déduit que u_1 est le premier vecteur propre de la matrice $T^{-1}E$ associé à la première valeur propre λ_1. Les autres axes sont engendrés par les autres vecteurs propres de la matrice $T^{-1}E$ associés aux valeurs propres ordonées au sens décroissant.

interprétation des résultats

Comme pour les autres méthodes factorielles, il est possible de représenter les individus dans les plans factorielles discriminants. Il est aussi possible comme pour l'ACP de représenter les variables en traçant le cercle de corrélation des p variables. Afin de mesurer la qualité de la représentation, les mêmes indicateurs que l'ACP peuvent être employés. Par exemple la qualité de représentation d'un nuage par un axe engendré par u_s est donnée par le rapport :

$$\frac{\lambda_s}{\sum_{i=1}^{n} \lambda_i} \tag{18}$$

La contribution absolue du centre de gravité g_q à l'axe engendré par u_s est définie par :

$$\frac{n_q}{n}\left(u_s t T^{-1} g_q\right)^2 \tag{19}$$

et la contribution relative du centre de gravité g_q à l'axe engendré pa u_s est définie par :

$$\frac{n_q}{n\lambda_s}\left(u_s t T^{-1} g_q\right)^2 \tag{20}$$

5.4.2 L'aspect classement

Nous souhaitons trouver la classe d'affectation d'un nouvel individu $x_{i'}$. Il existe plusieurs règles d'affectation (ou de classement) de cet individu dans une classe c_q. Nous en présentons ici quelques unes géométriques et probabilistes.

L'approche géométrique

L'idée est très simple, il s'agit de calculer les distances (défini par T^{-1}) entre le nouvel individu est les k centres de gravité de k groupes. On classera la nouvelle observation dans le groupe qui a une distance minimale.

Nous devons donc définir la distance entre l' individu $x_{i'}$ et le centre de gravité g_q du groupe c_q, qui doit être minimale, comme suit :

$$\begin{aligned} d^2(x_{i'}, g_q) &= (x_{i'} - g_q)^t T^{-1}(x_{i'} - g_q) \\ &= x_{i'}^t T^{-1} x_{i'} - 2x_{i'}^t T^{-1} g_q + g_q^t T^{-1} g_q \end{aligned} \tag{21}$$

Le problème devient la maximisation de la quantité suivante :

$$f_q = x_{i'}^t T_{-1} g_q - \frac{1}{2} g_q^t T^{-1} g_q \tag{22}$$

Les fonctions $f_i, \quad i = \{1, \cdots, k\}$, sont appelées les fonctions de classification ou les fonctions discriminantes et on affecte le nouvel individu à la classe qui maximise cette fonction.

L'approche probabiliste

L'idée est de classer la nouvelle observation $x_{i'}$, dans la classe pour laquelle la probabilité conditionnelle d'appartenir à cette classe, étant donné la valeur observée , est maximale (proche de 1).

Cette probabilité conditionnelle est calculée par deux façons :

1. Par la fonction discriminante :

$$\mathbb{P}(c_i/x_{i'}) = \left(\sum_{j=1}^{k} exp(f_j - f_i)\right)^{-1} \tag{23}$$

où f_i est fonction discriminante .

2. **La formule de Bayse** :
D'après la règle de Bayse, on a :

$$\mathbb{P}(c_i/x_{i'}) = \frac{\mathbb{P}(x_{i'}/c_i)\mathbb{P}(c_i)}{\sum_{i=1}^{k} \mathbb{P}(x_{i'}/c_i)\mathbb{P}(c_i)} \tag{24}$$

Il sufit alors de maximiser $\mathbb{P}(x_{i'}/c_i)\mathbb{P}(c_i)$. Cependant pour estimer cette probabilité il faut connaître les probabilités *à priori* $\mathbb{P}(c_i)$, ce qui n'est pas toujours le cas. Elles peuvent être estimées, il faut de plus estimer la probabilité $\mathbb{P}(x_{i'}/c_i)$ qui nécessite de faire l'hypothèse de la distribution. La distribution gaussienne qui peut être justifiée par la loi forte des grands nombres est souvent employée. De plus elle ne nécessite que l'estimation de deux paramètres (la moyenne et la variance).

5.5 Conclusion

L'AFD est une méthode très utilisée de nos jours. Sa simplicité de mise en œuvre fait que nous la retrouvons dans de nombreux logiciels. Elle est adéquate pour la représentation des données dans des espaces qui discriminent au mieux les individus selon des classes connues. Cette représentation permet de dégager des informations à partir d'un grand nombre de données souvent difficile à interpréter. Elle permet également l'affectation de nouveaux individus dans les classes existantes. Il est alors possible de rendre la méthode adaptative pour tenir compte de ces nouvelles observations.

Chapitre 6

La classification automatique

6.1 Rôle et importance de la classification automatique

Le terme classification recouvre plusieurs significations selon le contexe dans lequel il est utilisé. Le sens qui lui est donné en analyse des données est celui de la distibution en classes (décision par la distance). L'importance de la classification dans les sciences se reflète dans la grande variété des domaines où tant leur nature que leur construction ont fait l'objet de recherche.

Dans le cadre d'un problème de classification, on dispose d'un ensemble des données qui reprend une collection d'individus (objets) non étiquetés. Les classes sont encore inconues. l'objectif est alors d'obtenir des classes d'objets homogènes(d'inertie inta-classe minimale), en favorisant l'hétérogénété entre ces différentes classes (l'inertie inter-classes maximale).

On peut donc dire que toutes les téchniques de classification automatique suivent un même principe général qui consiste à minimiser la distance entre deux objets d'une même classe et maximiser la distance entre deux objets de deux classes distinctes. Le mot classe en Anglais fait référence à l'affectation d'un individu à une classe, il se traduit en Anglais par Cluster analysis. une définition formelle de la classification automatique qui prisse de base à un processus automatisé, amène a se poser les quetions suivantes :

- Comment les objets à classer sont-il définis ?
- Comment définir la notion de de ressemblance entre objets ?
- Qu'est ce qu'une classe ?
- Comment sont structurées les classes ?

- Comment comparer une classification par rapport à une autre ? On peut dire que les méthodes de classification automatique, sont des téchniques qui consistent à distribuer en classes un ensemble d'objets (individus ou variables).

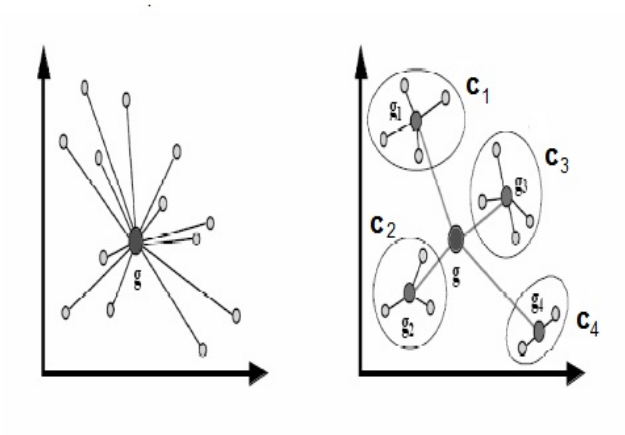

FIGURE 6.1 – La classification en classes un ensemble de points.

Remarque 6.1.1 *A la différence de la classification supervisée, qui étant donné un ensemble de classes déja identifiées, il s'agit de trouver la meilleure classe à laquelle un individu appartient. La classification non supervisée consiste à structurer des classes non encore identifiées qui regroupent ces objets.*

6.1.1 Les étapes d'une classification automatique

Les différentes étapes d'une classification automatique sont :

1. Le choix des données,
2. Le calcul des ressemblances entres les objets,

3. Le choix d'un algorithme de classification,
4. L'interprétation des résutats :
   - évaluation de la qualité de la classification.
   - description des classes obtenues.

Définition 6.1.1 *[20] L'inertie totale d'un nuage de points, notée I_T est la moyenne des distances de tous les points et le centre de gravité de nuage .*

$$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g) \tag{1}$$

Définition 6.1.2 *[20](l'inertie intra-classe) Si le nuage de points est partagé en k classes $c_1, c_2, \cdots, c_k$ celles-ci seront plus homogènes que les inerties de chaque classe,$I_1, I_2, \cdots, I_K$, calculées par rapport à leurs centres de gravité $g_1, g_2, \cdots, g_k$ respectifs, sont faibles. La moyenne de ces inerties est appelée inertie intraclasse :*

$$I_{itra} = \frac{1}{n}\sum_{j=1}^{k}\sum_{x_i \in c_j} d^2(x_i, g_j) \tag{2}$$

Définition 6.1.3 *[20](l'inertie inter-classe) On appelle l'inertie inter-classe la quantité suivante :*

$$I_{inter} = \frac{1}{n}\sum_{i=1}^{k} n_i d^2(g_i, g) \tag{3}$$

où n_i est le cardinal de la classe c_i

Théorème 6.1.1 *[]théorème de Huygens La some de l'inertie intra-classe et l'inertie inter-classe est l'inertie totale d'un nuage de points.*

$$I_T = I_{intra} + I_{inter} \tag{4}$$

6.1.2 La ressemblance entre deux objets

Pour mesurer la ressemblance entre deux objets à classer, deux démarches sont envisagées :

1. On peut dire que deux objets sont ressemblables, s'ils partagent certaines caractéristiques. Ce genre de démarches aboutit à une classification monothétique (les données qualitatives).

2. On peut aussi mesurer la ressemblance en utilisant une mesure de proximité (distance, coéfficient de corrélation...). Cette démarche est dite Polytétique.

Définition 6.1.4 *[31]*

1. *On appelle désimilarité, toute application d à valeurs dans $\mathbb{R}^+$ définie comme suit :*

$$d : I \longrightarrow \mathbb{R}^+$$

telsque :

(a) I est l'ensembe d'objets à classer,

(b) $\forall x_i, x_j \in I$ tels que :
- $d(x_i, x_j) = d(x_j, x_i)$,
- $d(x_i, x_j) = 0 \Longrightarrow x_i = x_j$.

2. *On appelle similarité, toute application S à valeurs dans $\mathbb{R}^+$ définie comme suit :*

$$S : I \longrightarrow \mathbb{R}^+$$

telle que :

(a) $\forall x_i, x_j \in I$ tels que :
- $S(x_i, x_j) = d(x_j, x_i)$,
- $S(x_i, x_i) \geq \Longrightarrow S(x_i, x_j)$.

6.1.3 Présentation des méthodes de classification

On conpte deux familles de techniques de la classification automatique :

– Les premiers sont de type algorithmique qui sont basées sur le calcul de simularité ou désimilarité entre les objets à classer.Ces méthodes sont divisées en deux groupes :

- Les méthodes hiérarchiques qui sont des algorithmes itératifs ,il y a deux algorithmes hiérarchiques, classification ascendante hiérarchique et classification discendente hiérarchique qui seront notées dans la suite par CAH et CDH.
- Les méthodes de partitionnement qui consiste a trouver une bonne partition de l'ensemble d'objets à classer.

– La seconde famille des méthodes de classification est constituée par des approches probabilistes, l'une est basée sur la densité et l'autre est basée sur la convexité de densité.

Définition 6.1.5 *Soit $(A_j)_{j\in J}$ une famille d'ensembles, on dit que les $A_j, i \in J$ forment une partition de l'ensemble I des objets à classer si et eulement si :*

1. $\forall j \neq k, \quad A_j \cap A_k = \emptyset$
2. $\sqcup_{j\in J} A_j = I$

Définition 6.1.6 *Soit $(A_j)_{j\in J}$ une famille d'ensembles, on dit que les $A_j, i \in J$ forment une hiérarchie des classes de l'ensemble I si et seulement si :*

1. $\forall j \neq k, \quad A_j \cap A_k = \emptyset$ *ou* $A_j \subseteq A_k$ *ou* $A_k \subseteq A_j$
2. $\sqcup_{j\in J} A_j = I$

C'est- à- dire, si l'ensemble I est divisé en un nombre fini de classes dont chacune est divisée en un ensemble fini de classes , ect..., on parle alors d'une hiérarchie de classes.

6.2 Classification ascendante hiérarchique (CAH)

Cette aproche consiste à trouver une classification de I en n classes , $n-1$ classes,..., une classe. Avant de démmarer l'algorithme, il faut choisir :

1. Une distance d pour mesurer la ressemblance entre les objets à classer, parmi ces distances, La distance euclidienne définie par :

$$d^2(x_i, x_k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2 \tag{5}$$

2. Un critère d'agrégation qui est la distance entre deux classes, noté par $\delta(c_k, c_h)$. Les critères les plus utilisables sont :
   - Le critère "min" : $\delta(c_k, c_h) = \min\{d^2(x_i, x_j)/x_i \in c_k \ et \ x_j \in c_h\}$
   - Le critère "max" :$\delta(c_k, c_h) = \max\{d^2(x_i, x_j)/x_i \in c_k \ et \ x_j \in c_h\}$
   - Critère de centre de gravité : $\delta(c_k, c_h) = d^2(g_k, g_h)$ avec $g_.$ est le centre de gravité de la classe $c_.$

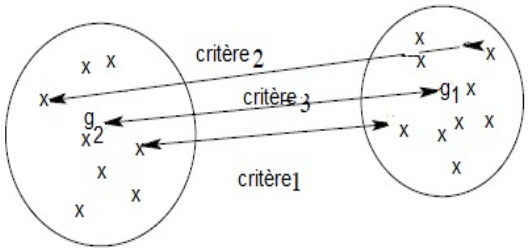

FIGURE 6.2 – Les critères d'agrégations .

6.3 L'arbre hiérarchique ou dendrogramme

On obtient une hiérarchi des classes qui se représentée sous forme d'un arbre de classification)(figure (6.3)).

Sur l'axe horizontal, on trouve les objets à classer ce sont les racines de l'arbre.

Les branches de l'arbre illustrent le différentes étapes de l'algorithme.

Les agrégations sont matérialisés par des points noirs, qui sont appelés "les nœuds" de l'arbre qui portant l'indication de son niveau d'agrégation.

L'axe vertical marque les valeurs du critère d'agrégation à chaque étape de l'algorithme.

FIGURE 6.3 – L'arbre hiérarchique ou dendrogramme.

6.3.1 L'algorithme de la méthode

Soit $I = \{x_1, x_2, \cdots, x_n\}$ l'ensemble d'objets à classer, l'algorithme s'effectue en $(n-1)$ étapes :

La première étape

•On choisit une distance et on calcule le tableau des distances $\triangle$ entre n classes composées par un seul objet défini par :

$$\triangle = \begin{array}{|c|c|c|c|c|} \hline & x_1 & x_2 & \cdots & x_n \\ \hline x_1 & 0 & d^2(x_1,x_2) & \cdots & d^2(x_1,x_n) \\ \hline x_2 & d^2(x_2,x_1) & 0 & \cdots & d^2(x_2,x_n) \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline x_n & d^2(x_n,x_1) & d^2(x_n,x_1) & \cdots & 0 \\ \hline \end{array} \quad (6)$$

C'est un tableau symétrique dont le diagonale égale à 0.

•On cherche parmi les couples $(x_i, x_j), i \neq j$ lequel a une distance minimale, qui forme la première classe c_1

• On choisit un critère d'agrégation et on passe à la deuxième étape.

La deuxième étape

• On recalcule la matrice de distance $\triangle$ entre les $n-1$ classes.

• on construit la deuxième classe c_2 on agrège le couple des classes qui a une distance minimale.

La k-ième étape

• On recalcule la matrice des distance entre les $n-(k-1)$ classes .

• On cherche parmi ces classes, le couple de classes qui a une distance minimale et on définit la classe c_k.

A la dernière étape, on agrège les deux dernières classes.

6.3.2 Exemple 1

soit $I = \{x_1, x_2, x_3, x_4, x_5\}$ un ensemble de 5 objets qui sont les lignes du tableau suivant

$$\begin{pmatrix} 1 & 2 & 3 & 0 & 1 & 4 \\ 1 & 3 & 0 & 2 & 3 & 0 \\ 1 & 1 & 2 & 2 & 1 & 1 \\ 2 & 0 & 3 & 1 & 4 & 0 \\ 1 & 1 & 0 & 3 & 4 & 1 \end{pmatrix}$$

On choisit la distance $d^2(x_i, x_j) = \sum_{k=1}^{6} |x_{ik} - x_{jk}|$ et le critère d'agrégation "min"

La première étape

	x_1	x_2	x_3	x_4	x_5
x_1	0	12	7	11	13
x_2	12	0	7	8	5
x_3	7	7	0	8	6
x_4	11	8	8	0	8
x_5	13	5	6	8	0

$\Longrightarrow c_1 = \{x_2, x_5\}$

La deuxième étape

	x_1	x_3	x_4	c_1
x_1	0	7	11	12
x_3	7	0	8	6
x_4	11	8	0	8
c_1	12	6	8	0

$\Longrightarrow c_2 = \{c_1, x_3\}$

La troisième étape

	x_1	x_4	c_2
x_1	0	11	7
x_4	11	0	8
c_2	7	8	0

$\Longrightarrow c_3 = \{c_2, x_1\}$

La quatrième étape

$c_4 = \{c_3, x_4\}$

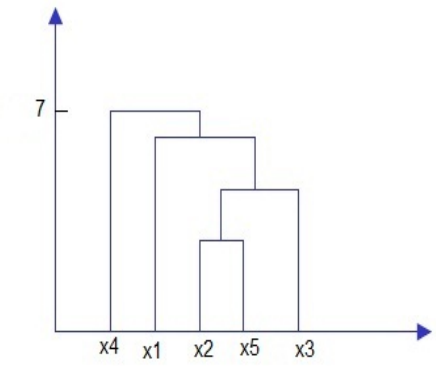

FIGURE 6.4 – L'arbre hiérarchique .

6.4 La classificatioin déscendante hiérarchique (CDH)

Cette méthode de classification construit sa hiérarchie dans le sens inverse que CHA, en commençant par une grande classe contenant tous les objets, à chaque étape k, elle se divise en deux classes l'une contenant un objet et l'autre contenant $n - k$ objets jusqu'à ce que toutes les classes ne contiennent qu'un seul individu.

6.4.1 L'algorithme de la méthode

Pour n objets, l'hiérarchie est construite en $n - 1$ étapes.

Dans la première étape, les données sont partagées en deux classes à l'aide d'un calcul d'une matrice de distance $\triangle$ définie par (6), la première classe contient l'objet le plus

éloigné aux autres objets, cet éloignement se quantifie par :

$$d_{x_i} = \frac{1}{n_k - 1} \sum_{\substack{i \neq j \\ x_j \in c_k}} d^2(x_i, x_j) \tag{7}$$

Dans chacune des étapes suivantes, la classe avec le diamètre le plus grand se divise de la même façon.Aprés $n-1$ divisions, toutes les classes ne contiennent qu'un seul objet.

6.4.2 Exemple 2

On prend les mêmes données de l'exemple1, on trouve

La première étape

	x_1	x_2	x_3	x_4	x_5	$d_{x_i} = \frac{1}{n_k - 1} \sum d^2(x_i, x_j)$
x_1	0	12	7	11	13	10.75
x_2	12	0	7	8	5	8
x_3	7	7	0	8	6	7
x_4	11	8	8	0	8	8.75
x_5	13	5	6	8	0	8

$\Longrightarrow$ $c_1 = \{x_1\}$
$c = \{x_2, x_3, x_4, x_5\}$

la deuxième étape

	x_2	x_3	x_4	x_5	$d_{x_i} = \frac{1}{n_k - 1} \sum d^2(x_i, x_j)$
x_2	0	7	8	5	6.67
x_3	7	0	8	6	7
x_4	8	8	0	8	8
x_5	5	6	8	0	6.33

$\Longrightarrow$ $c_2 = \{x_4\}$
$c = \{x_2, x_3, x_5\}$

la troisième étape

	x_2	x_3	x_5	$d_{x_i} = \frac{1}{n_k - 1} \sum d^2(x_i, x_j)$
x_2	0	7	5	6
x_3	7	0	6	6.5
x_5	5	6	0	5.5

$\Longrightarrow$ $c_3 = \{x_3\}$, $c = \{x_2, x_5\}$

La quatrième étape

$c_4 = \{x_2\}, \quad c_5 = \{x_5\}$

Exercice

1. Donner une CAH de l'ensemble I constitué par les lignes de X, on choisit la métique euclidienne et le critère d'agrégation "max"

$$X = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 2 & 3 & 1 & 2 \\ 3 & 4 & 1 & 2 \\ 4 & 4 & 1 & 1 \\ 2 & 3 & 4 & 4 \end{pmatrix}$$

2. Donner une CDH de I
3. Comparer entre les deux classifications.

6.5 Les méthodes de partitionnement

Ces méthodes cherchent une partition de l'ensemble d'objets à classer , en un certain nombre fixé à priori de classe k, deux conditions soient réalisées :

- Chaque classe doit contenir au moins un objet.
- Chaque objet doit appartenir à une seule classe.

6.6 Les algorithmes $k-$means

6.6.1 L'algorithme de Centre mobile

soit $I = \{x_1, x_2, \cdots, x_n\}$ l'ensemble de n objets à classer et supposons qu' on a k classes.

La première étape

Dans I, on choisit aléatoirement k centres initiaux $\{c_1, c_2, \cdots, c_k\}$ des k classes.

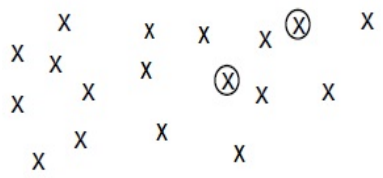

FIGURE 6.5 – Le choix des centres initiaux .

•On calcule un tableau des distances entre $\{c^1, c^2, \cdots, c^k\}$ et tous les objets,

	x_1	x_2	$\cdots$	x_n
c^1	d^1_{11}	d^1_{12}	$\cdots$	d^1_{1n}
c^2	d^2_{21}	d^2_{22}	$\cdots$	d^2_{2n}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
c^k	d^k_{k1}	d^k_{k2}	$\cdots$	d^k_{kn}

- On construit les classes autour des centres $\{c^1, c^2, \cdots, c^k\}$ tels que c_i contient les objets dont la distance avec c^i est minimale par arpport aux autres centres.

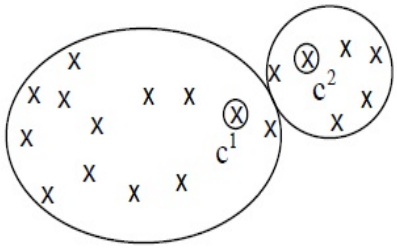

FIGURE 6.6 – La construction des classes autour des centres initiaux .

La deuxième étape

- On calcule les vrais centres de gravités des classes constituées dans la première étape $\{g_1^2, g_2^2, \cdots, g_k^2\}$.
- On calcule le tableau des distances entre $\{g_1^2, g_2^2, \cdots, g_k^2\}$ et tous les objets,

	x_1	x_2	$\cdots$	x_n
g_1^2	d'^1_{11}	d'^1_{12}	$\cdots$	d'^1_{1n}
g_2^2	d'^2_{21}	d'^2_{22}	$\cdots$	d'^2_{2n}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
g_k^2	d'^k_{k1}	d'^k_{k2}	$\cdots$	d'^k_{kn}

- On reconstruit des nouvelles classes autour les centres $\{g_1^2, g_2^2, \cdots, g_k^2\}$

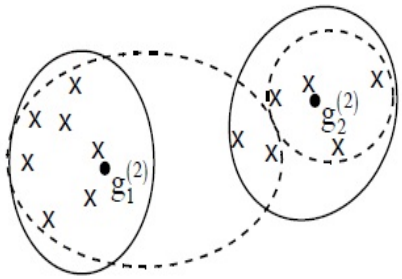

FIGURE 6.7 – La construction des classes autour des centres de gravités .

La l-ième étape

- On calcule les vrais centres de gravités des classes constituées dans l'étape $l-1$ $\{g_1^l, g_2^l, \cdots, g_k^l\}$.
- On calcule le tableau des distances entre $\{g_1^2, g_2^2, \cdots, g_k^2\}$ et tous les objets,

	x_1	x_2	$\cdots$	x_n
g_1^l	d_{11}^{l1}	d_{12}^{l1}	$\cdots$	d_{1n}^{l1}
g_2^l	d_{21}^{l2}	d_{22}^{l2}	$\cdots$	d_{2n}^{l2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
g_k^l	d_{k1}^{lk}	d_{k2}^{lk}	$\cdots$	d_{kn}^{lk}

- On reconstruit des nouvelles classes autour les centres $\{g_1^l, g_2^l, \cdots, g_k^l\}$

L'algorithme s'arrête lorsque les classes formées dans l'étape l sont les mêmes classes formées dans l'étape $l-1$.

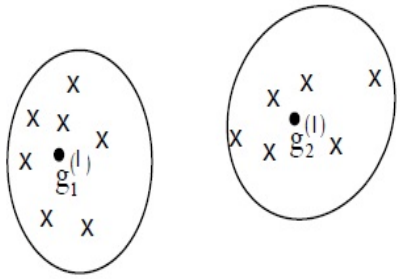

FIGURE 6.8 – La construction des classes autour des centres de gravités à l'étape k .

Exemple 3 :

On applique l'algorithme de centre mobile sur les lignes du tableau suivant :

$$X = \begin{pmatrix} 1 & 2 & 4 & 2 \\ 2 & 1 & 2 & 2 \\ 3 & 3 & 1 & 1 \\ 4 & 1 & 2 & 1 \\ 3 & 4 & 2 & 2 \end{pmatrix}$$

On prend x_1, x_2 comme des centres initiaux et on utilise la distance suivante

$$d^2(x, y) = \sum_{i=1}^{4} |x_i - y_i|$$

La première étape

	x_1	x_2	x_3	x_4	x_5
x_1	0	4	7	7	6
x_2	4	0	5	3	4

$\Longrightarrow \begin{array}{lcl} c_1 & = & \{x_1\} \\ c_2 & = & \{x_2, x_3, x_4, x_5\} \end{array}$

La deuxième étape

$$\begin{array}{lcl} g_1 & = & (1, 2, 4, 2) \\ g_2 & = & (3, 2.5, 1.75, 1.5) \end{array}$$

	x_1	x_2	x_3	x_4	x_5
g_1	0	4	7	7	6
g_2	4.25	3.25	1.75	3.25	2.25

$\Longrightarrow \begin{array}{lcl} c_1 & = & \{x_1\} \\ c_2 & = & \{x_2, x_3, x_4, x_5\} \end{array}$

Stabilité de l'algorithme

6.7 La méthode de PAM (Partition Around Medoids

Dans cet algorithme, chaque classe est représentée par l'un de ses objets ce représentant est appelé "médoïde".

Le principe de cet algorithme est de choisir un ensemble des médoïdes et affecter chaque objet au médoïde le plus proche et itérativement, remplacer chaque médoïde par un autre objet dans la même classe si cela permet de réduire l'inertie intra-classe.

6.7.1 L'algorithme de PAM

- La première étape :Dans l'ensemble d'objets I, on choisit arbitrairement k médoïdes $\{M_1, \cdots, M_k\}$,
- La deuxième étape : On calcule un tableau de distance entre tous les objets et les k médoïdes choisis ,et on affecte chaque objets restant au médoïde le plus proche.
- La troisième étape : Dans la plus grande classe représentée par le médoïde M_j, on choisit aléatoirement un objet non médoïde x_h , et on remplace le médoïde M_j par x_h.

  On recalcule le tableau des distance entre tous les objets et x_h et les autres médoïdes choisis.

  On forme des nouvelles classes autours de médoïdes.

On calcule le coût de remplacement de M_j par x_h, si $T_{jh} < 0$, on accepte le changement.

•La k-ième étape : Dans la même classe, on fait un autre changement, on choisit un autre objet x_l et on remplace x_h par x_l, on clacule le coût T_{hl}, s'il est négatif,on accepte ce changement et on construit les nouvelles classes jusqu'à ce qu'il n'y ait plus de changement.

Le T_{ij} est quantifié par :

$$T_{jh} = \sum_{x_i \in c_h} d^2(x_i, x_h) - \sum_{x_r \in c_j} d^2(x_r, M_j)$$

où,

c_h (respectivement c_j) la classe représentée par x_h (respectivement représentée par M_j)

6.7.2 Exemple 4 :

Appliquer un algorithme de PAM sur les lignes du tableau X suivat :

$$X = \begin{pmatrix} 1 & 2 & 4 & 2 \\ 2 & 1 & 2 & 2 \\ 3 & 3 & 1 & 1 \\ 3 & 1 & 2 & 1 \\ 3 & 4 & 2 & 3 \\ 4 & 4 & 3 & 3 \end{pmatrix}$$

On choisit $\{x_3, x_4\}$ comme des médoïdes et on utilise la même distance que l'exemple 3.

La première étape

	x_1	x_2	x_3	x_4	x_5	x_6
x_3	7	5	0	3	4	6
x_4	6	2	3	0	5	7

$$\Longrightarrow \begin{array}{rcl} c_1 & = & \{x_3, x_5, x_6\} \\ c_2 & = & \{x_1, x_2, x_4\} \end{array} \Longrightarrow T_1 = 10 + 8 = 18$$

La deuxième étape :

Dans la classe c_1, on remplace x_3 par x_5

x_5	7	5	4	6	0	2
x_4	6	2	3	0	5	7

$$\Longrightarrow \begin{array}{rcl} c_1 & = & \{x_5, x_6\} \\ c_2 & = & \{x_1, x_2, x_3, x_4\} \end{array} \Longrightarrow \begin{array}{l} T_2 = 2 + 11 = 13 \\ T_{5,3} = -5 < 0 \end{array}$$

On accepte le changement.

La troisième étape :

On remplace $x_5$par x_6

x_6	7	7	6	7	2	0
x_4	6	2	3	0	5	7

$\Longrightarrow$ $c_1 = \{x_5, x_6\}$, $c_2 = \{x_1, x_2, x_3, x_4\}$ $\Longrightarrow$ $T_3 = 2 + 11 = 13$, $T_{5,6} = 0$

On accepte les classes c_1 dont le représentant x_6 et c_2 dont le représentant x_4, car $d^2(x_5, x_4) < d^2(x_6, x_4)$.

6.7.3 L'avantage et l'inconvénient de PAM

L'avantage de cette méthode est qu'elle est plus robuste que la méthode de centre mobile en présence de bruit. L'inconvénient vient de la complexité des calculs pour chaque itération ce qui est plus coûteux en cas de k et n assez grand, ce qui nous conduit à dire que cet algorithme est efficace en cas des données de petite taille. et le deuxième inconvénient est que les résultats dépendent du choix de k et $M_1, \cdots, M_k$.

Exercice

Comparer entre les deux classifications illustrent de la méthode de centre mobile et la méthode de $k-$médoïdes , de l'ensemble des lignes de X suivant :

$$X = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 2 & 3 & 1 & 2 \\ 3 & 4 & 1 & 2 \\ 4 & 4 & 1 & 1 \\ 2 & 3 & 4 & 4 \end{pmatrix}$$

On prend x_1, x_2 comme des centres et médoïdes initiaux.

6.8 Modélisation probabiliste en classification

L'approche statistique se base sur des modèles probabilistes qui formalisent l'idée de classe.Cette approche permet d'interpréter de façon statistique la classification obtenue.

La modélisation la plus classique est celle du modèle de mélange fini qui peut être paramétrique ou non.

6.8.1 Le problème

Un objet de l'ensemble d'objets à classer $I = \{x_1, \cdots, x_n\}$ sera modélisé par le vecteur X, décrit lui même par p variables quantitatives. Le but de la classification étant d'associer le vecteur X à une des k classes. Nous introduisons la variable auxiliaire Z à valeurs dans $\{1, 2, \cdots, k\}$ tel que $Z = i$ si X appartient à la classe c_i.

Ainsi, le problème de la classification revient à établir une règle de décision δ qui associe au vecteur $X \in \mathbb{R}^p$ un vecteur $Z \in \{1, 2, \cdots, k\}$.

$$\begin{array}{rcl} \delta : \mathbb{R}^p & \longrightarrow & \{1, 2, \cdots, k\} \\ X & \longmapsto & Z \end{array}$$

Cette règle de décision est généralement construite à partir d'un jeu des données (d'apprentissage) et c'est la nature de ce jeu de données qui différencie les deux types d'apprentissage.

1. **L'apprentissage supervisé :**Les données utilisées pour ce type noté $y_1, \cdots, y_n$, sont dites "complètes" car elles contiennent à la fois les vecteurs $x_1, \cdots, x_n$ pris par p variables quantitatives et leurs apprentissages aux k classes $Z_1, \cdots, Z_n$. Les données complètes sont donc l'ensemble des couplets d'observations "labels" c'est-à- dire $\{y_1, \cdots, y_n\} = \{(x_1, z_1), \cdots, (x_n, z_n)\}$
2. **L'apprentissage non-supervisé :**Les données utilisées pour ce type ne sont pas complètes car elles ne contiennent que les vecteurs $x_1, \cdots, x_n$ pris par les p variables quantitatives.

6.8.2 Le principe

La classification probabiliste suppose que les observations $x_1, \cdots, x_n$ de l'ensemble I des objets à classer sont des réalisations d'un vecteur aléatoire X à valeurs dans $\mathbb{R}^p$. Elle suppose que les valeurs $Z_1, \cdots, Z_n$ sont des réalisations de la variable aléatoire Z à valeurs dans $\{1, 2, \cdots, k\}$, le fait de dire que x est une réalisation de X conditionnellement au fait que $Z = i$ revient à dire que l'observation $x \in c_i$. nous introduisons également le vecteur aléatoire $S \in \{0, 1\}^k$ tel que $Z = i \Longrightarrow S = \underbrace{(0, 0, \cdots, \underbrace{1}_{ordrei}, 0, 0 \cdots, 0)}_{k}$

6.8.3 La règle de Bayes

Le cadre probabiliste permet de construire la règle optimale δ^* , dite "règle de Bayes", qui minimise le risque conditionnel $R(\delta/x)$ pour chaque observation x

$$\delta^* = \min_{\delta} R(\delta/x) \tag{8}$$

où

$$R(\delta/x) = 1 - \mathbb{P}(Z = \delta(x)/X = x) \tag{9}$$

La règle δ^* consiste donc à affctter l'observation x à la classe la plus probable à postériori.

$$\delta^* = \max_{i=1\cdots,k} \mathbb{P}(Z = i/X = x) \tag{10}$$

6.8.4 Approche générative et approche discriminative

En classification probabiliste, la règle de décision repose donc sur la probabilité à postériori et c'est la manière de calculer ces probabilités qui déffirencie les deux approches de la classification probabiliste : l'approche discriminante et l'approche générative.

la première modélise directement la probabilité à postériori $\mathbb{P}(Z/X)$.
la deuxième cherche tout d'abord à modéliser la distribution conjointe $\mathbb{P}(X, Z)$ et en déduit ensuite la règle de classification, en utilisant la formule de Bayes :

$$\mathbb{P}(Z/X) = \frac{\mathbb{P}(Z).\mathbb{P}(X/Z)}{\mathbb{P}(X)} \simeq \mathbb{P}(Z).\mathbb{P}(X/Z) \tag{11}$$

Modélisation par mélange de lois

Le modèle de mélange suppose que chaque classe est caractérisé par une distribution de probabilité. Dans un modèle de mélange, on considère que les données $x_1, \cdots, x_n$ constituent un échantillon de n réalisations indépendantes du vecteur aléatoire X à valeurs dans $\mathbb{R}^p$ dont la fonction de densité peut s'écrire de la façon suivante :

$$f(x) = \sum_{i=1}^{k} \pi_i f_i(x) \tag{12}$$

où

- k est le nombre des classes,
- f_i est la densité de la distibution de X conditionnellement à $Z = i$ (la i-ième composante de mélange),
- π_i est la proportion de mélange $\pi_i \in [0, 1], \sum_i \pi_i = 1$

On suppose généralement que la densité f_i des classes appartiennent à une famille paramétrique , $f_i(.) = f_i(., \theta_i)$ et le modèle de mélange s'écrit alors,

$$f(x) = \sum_i \pi_i f(x, \theta_i), \quad \theta_i \in \Theta \tag{13}$$

Θ est l'ensemble des paramètres du modèle.

Dans ce cas, la formule (11) de Bayes s'écrit sous la forme suivante :

$$\mathbb{P}(Z = i/X = x, \theta) = \frac{\pi_i f(x, \theta_i)}{f(x)} \tag{14}$$

où , $f(x) = \sum_{i=1}^{k} \pi_i f(x, \theta_i)$ (formule totale de probabilité)

Le modèle de mélange paramétrique gausien

Parmi les modèles de mélange paramétriques, le modèle Gaussien est certainement le plus utilisé en classification probabiliste. Dans ce cas, les densités de probabilités des variables explicatives conditionnellement aux classes $f(x, \theta_i) \quad \forall i = 1, \cdots, k$, sont supposées être celles de loi normale $N(\mu_i, \Sigma_i)$:

$$f(x, \theta_i) = \frac{1}{(2\pi)^{p/2}(det\Sigma)^{1/2}} exp\left(\frac{-1}{2}(x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i)\right) \tag{15}$$

où $\theta_i = (\mu_i, \Sigma_i)$. Dans le cadre du modèle de mélange paramétrique, le problème de la classification se résume à l'estimation des paramètres du modèle.

Estimation des paramètres d'un modèle de mélange

Nous considérons que les données observées $x_1, \cdots, x_n$ ne correspondent qu'à une connaissance partielle des données complètes. Nous verrons que dans le cas de la classification non supervisée cette informaton sera absente et Z_i sera appelée une donnée manquante.

On appelle vraissemblance complète, la vraissemblance calculée à partir des données complètes et elle sera notée $L(y, \theta)$ ou $L(\theta)$.

Estimation par maximum de vraissemblance

Cette méthode propose d'estimer le paramètre θ du modèle par

$$\widehat{\theta}_{MV} = \max L(\theta)$$

où $L(\theta)$ est la vraissemblance complète du modèle.

Les n observations $y_1, \cdots, y_n$, étant supposées indépendantes,on peut exprimer la vraissemblance du paramètre θ par le produit de toutes les densités marginales . Dans le cas du modèle de mélange, le logaritme de vraissemblance s'écrit alors de la façon suivante :

$$L(\theta) = \sum_{j=1}^{n} ln(\sum_{i=1}^{k} \pi_i f(x_j, \theta_i)). \tag{16}$$

Bibliographie

[1] J.P. AURAY et Ed. Alexandre Lacassagne (1990). "Analyse des données multidimensionnelles".tome 4

[2] J. Banfield and A. Raftery. Model-based Gaussian and non-Gaussian clustering. Biometrics, 49 :803 à 821, 1993.

[3] J.P.Benzécri (1980). L'analyse des données, Tome 2 : l'analyse des correspondances, Dunod

[4] J. P. BENZECRI "Histoire et préhistoire de l'analyse des données". Dunod (1983).

[5] C. BIALES. "L'analyse statistique des Données". Chotard et Associés (1988).

[6] J.M.Bouroche (1977). Analyse des données en marketing, Masson.

[7] J. M. BOUROCHE et P. BERTIER. "Analyse des données multidimensionnelles". PUF, 2 èdition (1977).

[8] J. M. BOUROCHE et G. SAPORTA. "L'analyse des données". Collection Que Sais-Je, PUF (1980).

[9] X. BRY. "Analyses factorielles simples". Economica (1995).

[10] X. BRY. "Analyses factorielles multiples". Economica (1996).

[11] F. CAILLIEZ et J. P. PAGES. "Introduction à l'analyse des données". Smash (1976).

[12] G. CELEUX, E. DIDAY, H. RALAMBONDRAINY, Y. LECHEVALLIER et G. GOVAERT. "Classification automatique des données". Dunod (1989).

[13] G. CELEUX et J.P. NAKACHE. "Analyse discriminante sur variables qualitatives". Polytechnica (1994)

[14] P. CIBOIS. "L'analyse factorielle". Coll. Que Sais-Je. PUF (1983).

[15] C.Cohen (1969). On the computation of canonical correlations. Cahiers Centre d'Etudes de Reserche Opérationnelle 11, 121132.

[16] J. P. CRAUSER, Y. HARVATOPOULOS et P. SARNIN. "Guide pratique d'analyse des données". Editions d'organisation (1989).

[17] R. M. CORMACK, A Review of Classification, J.R. Statistic Sol, vol. 134, part 3, 1971.

[18] M.CRUCIANU, J.P.ASSELIN, DE BEAUVILLE R.BON." Méthodes factorielles pour l'analyse des données " Hermes-Lavoisier (2004)

[19] P. DAGNELIE. "Analyse statistique à plusieurs variables". Presses agronomiques de Gembloux (1975).

[20] E. DIDAY et coll. "Optimisation en classification automatique" INRIA (1979).

[21] E. DIDAY, J. LEMAIRE, P. POUGET et F. TESTU. "Eléments d'analyse des données". Dunod (1983).

[22] J.J. DROESBEKE, B. FICHET et P. TASSI (éditeurs). "Modèles pour l'analyse des données multidimensionnelles. Economica (1992).

[23] G. DURU, A. ZIGHED et M.BARDOS. "Analyse discriminante", Dunod (2001)

[24] B. ESCOFIER et J. PAGES. "Analyses factorielles simples et multiples". Dunod (1988).

[25] J.P. FENELON. "Qu'est-ce que l'analyse des données ?". Lefonen (1982).

[26] R.A. Fisher (1936).The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 : 179à 188.

[27] T. FOUCART. "Analyse factorielle des tableaux multiples". Masson (1984).

[28] M.JAMBU. Méthodes de base de l'analyse des données. Eyrolles (1999).

[29] H.Hotelling (1936). Relations between two sets of variants. Biometrika, 28, 321-377.

[30] L.Lebart, A.Morineau et J.P. Fenelon (1979). Traitement des données statistiques. Dunod

[31] M. O. LEBEAUX. "Classification automatique pour l'analyse des données".

[32] I. C. LERMAN. "Les bases de la classification automatique". Gauthier-Villars (1970).

[33] J.MOREAU, P.A.DOUDIN et P.CAZES. L'analyse des correspondances et les techniques connexes. Springer (2000)

[34] K.Pearson (1904). Report on certain enteric fever inoculation statistics. BMJ 3 :1243-1246.

[35] J. M. ROMEDER. "Méthodes et programmes d'analyse discriminante". Dunod (1973).

[36] M. ROUX. "Algorithmes de classification". Masson (1986).

[37] C. Spearman (1904). General Intelligence, Objectively Determined and Measured : The American Journal of Psychology, Vol. 15, No. 2 (Apr., 1904), pp. 201-292.

[38] R. TOMASSONE, M. DANZART, J. J. DAUDIN et J.P. MASSON. "Discrimination et classement". Masson (1988).