Ecole Nationale de Statistique, de Planification et de Démographie

(Université de Parakou)

Classification et Analyse factorielle discriminante

(Notes de cours)

Maxime AGBO

Motivation et Objectif

Nous partons des questions banales suivantes: Pourquoi le Bénin est subdivisé en 12 départements? Pourquoi 12 et non 36? Pourquoi Tchaourou et Parakou sont le même département?

En réalité, pour bien gérer les ressources, il est nécessaire de cibler les actions afin de mieux satisfaire les besoins des populations. Par ailleurs, on sait que 11 millions de Béninois ne signifie pas forcément 11 millions de besoins. On pourrait en avoir moins car plusieurs bénininois peuvent avoir les mêmes besoins. On dit que des béninois peuvent se ressembler. On voit donc que pour satisfaire les populations il est nécessaire de subdiviser le pays en des groupes homogènes (ensemble de personnes ayant les mêmes caractéristiques d'intérêt) selon la distribution des besoins sur l'ensemble du territoire.

D'autre part, il est difficile de subdiviser le Bénin en un nombre élevé de département car cela génère des coûts (chefs lieu, salaire des préfets et du personnel de la préfecture, infrastructure roulante, etc.) Le gouvernement a des contraintes budgétaires. Il faut donc trouver le nombre optimal de départements en fonction du budget disponible.

Enfin, Tchaourou et Parakou sont le même département car ces deux communes ont les mêmes caractéristiques en termes de besoin, de culture, etc.

De façon générale, lorsqu'on a une population, on a besoin de la segmenter de sorte que les individus d'un même segment aient les mêmes caractéristiques d'intérêt. La mèthode permettant de faire cette subdivision est la classification.

Par ailleurs, pour une population donnée, on peut disposer d'information sur le nombre de groupes qui y existent sans connaître ce qui caractérise chaque groupe, c'est-à-dire ce qui gouverne l'appartenance à un groupe. Et pourtant il est important de pouvoir déterminer les caractéristiques de chaque groupe de sorte que, si on prend n'importe quel individu pris au hasard, on puisse dire son groupe probable d'appartenance: C'est de l'Analyse discriminante.

Pour résumer, en classification on ne connaît pas le nombre de groupe qu'on peut avoir; mais en analyse discriminante on connaît le nombre de groupes mais on observe pas ces groupes.

L'objectif de cet ECUE est de se familiariser avec les outils d'analyse descriptive permettant d'identifier des groupes ou associations et de les caractériser: Classification et Analyse factorielle discriminante (AFD). L'accent est mis à la fois sur la théorie et la pratique.

Prerequisites

Statistique descriptive, ACP, AFC, ACM.

Contenu

1. Révision sur les mélanges de population
2. Classification
3. Analyse Factorielle Discriminante

Méthode d'evaluation

Devoir sur table (30%), TP à noter (20%) Examen final (50%). Les étudiants sont encouragés à participer au cours, des notes peuvent être accordées sous forme de bonus.

Références

- Gary Koop, Analysis of Economic Data, 3rd edition, Wiley, 2009.
- Brigitte Le Roux, Analyse géométrique des données multidimensionnelles, Dunod, Paris, 2014.
- Gilbert Saporta, Probabilités, Analyse des données et Statistiques, Editions Technip, 1990.

Logiciels: R et $SPAD$.

1 Chapitre 0: Révision sur les mélanges de populations (en anglais)

Let us suppose that the population of interest $\mathcal{P}$ is composed of m sub-populations $\mathcal{P}_1$, $\mathcal{P}_2$, $\cdots$, $\mathcal{P}_m$; with n_h the size of the sub-population $\mathcal{P}_h$, $h = 1, 2, \cdots, m$. We have $n = \sum_{h=1}^{m} n_h$, where n is the size of the population of interest $\mathcal{P}$.

Example.

- Benin is composed of departments.
- The civil servants population is composed of executive, foreman, laborer.
- Benin population is composed of the women sub-population and the men sub-population.

Let us consider that the variable of interest is X. Our objective is to analyse the relation between the characteristics of X in the sub-populations and that of the population of interest. For example, how to find the mean or the variance of X for the population $\mathcal{P}$ if we know the mean or the variance of X in each sub-population. If we know the average wages in women sub-population and men sub-population, we can know the average wage of Benin civil servants.

Let us consider the following Table 1. X has K occurrences. n_{kh} stands for the

Table 1:

X	$\mathcal{P}_1$	$\cdots$	$\mathcal{P}_h$	$\cdots$	$\mathcal{P}_m$	$\mathcal{P}$
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1m}	$n_{1.}$
x_2	n_{21}	$\cdots$	n_{2h}	$\cdots$	n_{2m}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	$\cdots$	n_{kh}	$\cdots$	n_{km}	$n_{k.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_K	n_{K1}	$\cdots$	n_{Kh}	$\cdots$	n_{Km}	$n_{K.}$
Total	n_1	$\cdots$	n_h	$\cdots$	n_m	n

number of individuals in the sub-population $\mathcal{P}_h$ who present the occurrence x_h. $n_{k.}$ is the total number of individuals presenting the occurrence x_k in the population $\mathcal{P}$.

$$n_{k.} = \sum_{h=1}^{m} n_{kh}$$

1.1 Frequencies

Let f_k be the frequency of the occurrence x_k in the population $\mathcal{P}$ and f_{kh} this frequency in the sub-population $\mathcal{P}_h$.

$$f_k = \frac{n_{k.}}{n} = \frac{1}{n}\sum_{h=1}^{m} n_{kh},$$

$$f_{kh} = \frac{n_{kh}}{n_h} \Rightarrow n_{kh} = n_h \times f_{kh} \Rightarrow f_k = \frac{1}{n}\sum_{h=1}^{m} n_h \times f_{kh}$$

$$\Rightarrow f_k = \sum_{h=1}^{m} P_h \times f_{kh}, \text{ with } P_h = \frac{n_h}{n} = \text{weight of the sub-population } h$$

1.2 Cumulative frequencies

Let $F_h(x)$ be the proportion of individuals whose occurrence is lower than x in sub-population $\mathcal{P}_h$. Hence, the number of individuals whose occurrence is lower than x in the sub-population $\mathcal{P}_h$ is $n_h F_h(x)$. So, the number of individuals whose occurrence is lower than x in the population $\mathcal{P}$ is $\sum_{h=1}^{m} n_h F_h(x)$.

$$\sum_{h=1}^{m} n_h F_h(x)$$

$$F(x) = \frac{1}{n}\sum_{h=1}^{m} n_h F_h(.) = \sum_{h=1}^{m} P_h F_h(x)$$

1.3 The mean

Let $\bar{X}_h$ be the mean in the sub-population $\mathcal{P}_h$ and $\bar{X}$ the mean in the population $\mathcal{P}$.

$$\begin{aligned}\bar{X} &= \sum_{k=1}^{K} f_k x_k = \sum_{k=1}^{K}\left(\sum_{h=1}^{m} P_h f_{kh}\right) x_k \\ &= \sum_{h=1}^{m} P_h \sum_{k=1}^{K} f_{kh} x_k\end{aligned}$$

The mean of the population $\mathcal{P}$ is the weighted average of the sub-populations means $\mathcal{P}_h$, where the weights are the sub-populations weights.

Example. (Homework) Benin is composed of 3 regions (South, Center, North) whose populations are estimated respectively at: 5 billion , 3 billion and 2 billion and the average incomes are respectively at: 70.000, 40.000, 60.000. What is the national average income in Benin?

Example.(Homework) What is the mean when the sub-population have the same size?

Remark 1.1. *Let $\bar{x_m}$ be the highest sub-population mean and $\bar{x_1}$ be the lowest sub-population mean. We have $\bar{x_1} \leq \bar{x} \leq \bar{x_m}$.*

1.4 The median

Suppose that the sub-populations are ranked in ascending order according to their medians, that is, $M_1 \leq M_2 \leq \cdots \leq M_m$. If M is the median of the population $\mathcal{P}$ then we have: $M_1 \leq M \leq M_m$.(Homework)

1.5 The variance

The variance of X in a sub-population $\mathcal{P}_h$ is

$$
\begin{aligned}
\sigma_h^2 &= \sum_{k=1}^{K} f_{kh}(x_k - \bar{x_h})^2 \\
\sigma_h^2 &= \sum_{k=1}^{K} f_{kh}\left[(x_k - \bar{x}) - (x_h - \bar{x})\right]^2 \\
&= \sum_{k=1}^{K} f_{kh}\left[(x_k - \bar{x})^2 + (\bar{x_h} - \bar{x})^2 - 2(x_k - \bar{x})(\bar{x_h} - \bar{x})\right] \\
&= \sum_{k=1}^{K} f_{kh}(x_k - \bar{x})^2 + \sum_{k=1}^{K} f_{kh}(\bar{x_h} - \bar{x})^2 - 2\sum_{k=1}^{K} f_{kh}(x_k - \bar{x})(\bar{x_h} - \bar{x}) \\
&= \sum_{k=1}^{K} f_{kh}(x_k - \bar{x})^2 + \sum_{k=1}^{K} f_{kh}(\bar{x_h} - \bar{x})^2 - 2(\bar{x_h} - \bar{x})\sum_{k=1}^{K} f_{kh}(x_k - \bar{x}) \\
&= \sum_{k=1}^{K} f_{kh}(x_k - \bar{x})^2 + \sum_{k=1}^{K} f_{kh}(\bar{x_h} - \bar{x})^2 - 2(\bar{x_h} - \bar{x}) \\
&= \sum_{k=1}^{K} f_{kh}(x_k - \bar{x})^2 + (\bar{x_h} - \bar{x})^2 - 2(\bar{x_h} - \bar{x})^2 \\
&= \sum_{k=1}^{K} f_{kh}(x_k - \bar{x})^2 + (\bar{x_h} - \bar{x})^2 - (\bar{x_h} - \bar{x})^2
\end{aligned}
$$

In addition

$$\begin{aligned} V(X) = \sigma^2 &= \sum_{k=1}^{K} f_{kh}(x_k - \bar{x})^2 \\ &= \sum_{k=1}^{K}(\sum_{h=1}^{m} P_h f_{kh})(x_k - \bar{x})^2 \\ &= \sum_{h=1}^{m} P_h \left[\sum_{k=1}^{K} f_{kh}(x_h - \bar{x})^2\right] \\ \text{so} \quad V(X) &= \sum_{h=1}^{m} P_h \left[\sigma_h^2 + (\bar{x_h} - \bar{x})^2\right] \\ \sigma^2(X) &= \sum_{h=1}^{m} P_h \sigma_h^2 + \sum_{h=1}^{m} P_h (\bar{x_h} - \bar{x})^2 \end{aligned}$$

The variance of the population has two components. The first one is $\sum_{h=1}^{m} P_h \sigma_h^2$. It is the weighted average of the variances of the sub-populations. It is called *within variance*. The second one is the quantity $\sum_{h=1}^{m} P_h (\bar{x_h} - \bar{x})^2$. it is the variance of the means of the sub-populations. It is called *between variance*. In other words, the variance in the population $\mathcal{P}$ is the sum of the within variance and the between variance.

The within variance allows to appreciate the homogeneity inside the sub-populations. Specifically, within variance helps know how similar or different are the individuals of the same sub-population. When the within variance is low (resp. high) then the sub-populations are homogeneous (resp. heterogeneous).

Concerning the between variance, it allows to appreciate differences (deviation) between the sub-populations. If the between variance is low (resp. high) then the sub-populations are similar (resp. different) from one to another.

2 Chapitre 1: La classification

2.1 Motivation

Il est clair que dans de nombreux situation nous avons besoin de segmenter la population en fonction des caractéristiques des unités statistiques. Avec l'analyse factorielle (ACP ou ACM), le nuage de points (la projection des individus dans les plans factorielles) peut révéler l'existence de groupes. Mais de façon générale, ce n'est pas le cas et l'analyse factorielle étudiée précédemment ne permet plus d'opérer la subdivision de la population. Il faut donc une autre méthode d'analyse des données pour résorber cette limite.

2.2 Principe de la classification

Il s'agit de former des groupes qui soient homogènes autant que possibles, c'est-à-dire il faut que les individus d'un même groupe se ressemblent le plus possible et que les individus de groupes différents soient différents le plus possible. En d'autres termes, les groupes doivent être constituer de sorte que la variance (inertie) inter soit élevée et la variance intra soit faible.

2.3 Définition et approche

La classification est le regroupement des individus d'une population en sous-groupes homogènes. Elle consiste donc à opérer une segmentation de la population (ou partition de la population) à partir des variables observées. On peut dire qu'en fait, pour effectuer la classification, on imagine une ou des variables composites des variables initiales, et on classe les individus selon les valeurs prises par cette variable composite. Il s'agit d'une variable latente, c'est-à-dire une variable qu'on n'observe pas, mais qui permet de définir les classes. Les individus qui ont des valeurs proches pour cette variables seront dans un même groupe. Comme nous l'avons dit plus haut, contrairement à l'analyse discriminante, les classes ne sont pas connues à l'avance, ni leur nombre.

Dans la pratique, la classification utilise des algorithmes et toutes les méthodes de classification partagent un principe et une démarche communs.

1. On compare les individus selon un certain critère, en examinant les ressemblances et les dissemblances;

2. On regroupe les individus qui présentent une similarité par rapport à l'ensemble des variables;

3. Les groupes obtenus doivent être aussi homogènes que possible.

De façon générale, les comparaisons s'effectuent à l'aide d'un critère de distance que l'on définit.

2.4 Notion de distance

La distance est un indicateur permettant de juger de la ressemblance ou non de deux individus ou de deux groupes d'individus. Si nous désignons par $D(A, B)$ la distance entre deux individus A et B, les propriétés suivantes sont vérifiées:

- $D(A, B) = D(B, A)$: symétrie;
- $D(A, B) = 0 \Leftrightarrow A = B$: séparation;
- Si A, B et C sont trois individus, on a $D(A, C) \leq D(A, B) + D(B, C)$: inégalité triangulaire.

Il existe plusieurs tyoes de distance. Dans le cadre de ce cours, nous allons considérer trois types de distance. Supposons que nous ayons une base de données qui comporte p variables X_1, X_2, $\cdots$, X_p, et n individus. Désignons par x_j^i la valeur prise par la variable X_j chez l'individu i.

1. La distance euclidienne:

$$D_e(k, k') = \sqrt{\sum_{j=1}^{p} \left(x_j^k - x_j^{k'}\right)^2}$$

2. Le carré de la distance euclidienne:

$$D_e^2(k, k') = \sum_{j=1}^{p} \left(x_j^k - x_j^{k'}\right)^2$$

3. La distance valeur absolue ou de Manhattan:

$$D_m(k, k') = \sum_{j=1}^{p} \left|x_j^k - x_j^{k'}\right|$$

Après avoir fait le choix du type de distance, il faut aussi définir comment calculer la distance entre un individu et un groupe, et la distance entre deux groupes. Pour ce faire, nous avons les considérations suivantes:

1. La notion de la plus petite distance entre deux éléments de chaque groupe (single linkage ou saut minimal). Voir Figures 1 et 2.
2. La notion de la plus grande distance entre deux éléments pris dans chaque groupe (complete linkage ou saut maximal). Voir Figures 3 et 4.

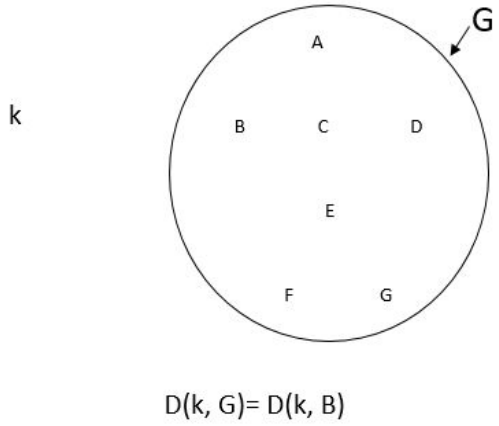

Figure 1: Saut minimal: distance entre un individu et un groupe

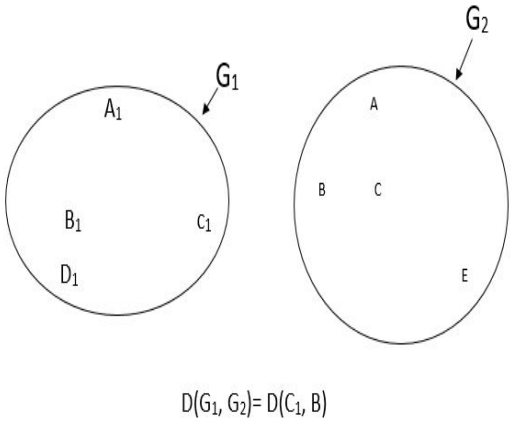

Figure 2: Saut minimal: distance entre deux groupes

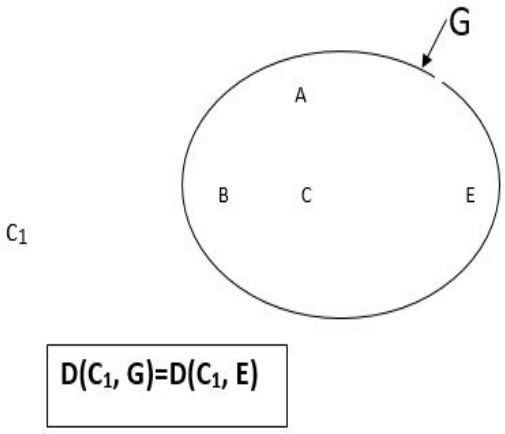

Figure 3: Saut maximal: distance entre un individu et un groupe

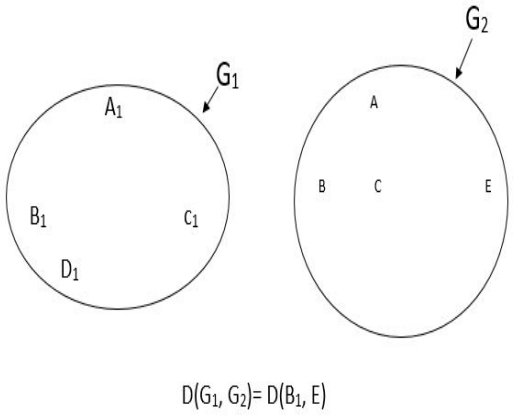

Figure 4: Saut maximal: distance entre deux groupes

3. La notion de la distance moyenne entre les éléments de chaque groupe.

4. La notion de la distance entre les centres de gravité (centres d'inertie ou barycentre ou point moyen, etc.)

2.5 Les différentes méthodes de classification

Il existe plusieurs méthodes de classification:

- la classification hiérarchique (ascendante et descendante);
- la classification non hiérarchique ou le partitionnement;
- la classification mixte.

2.5.1 La classification ascendante hiérarchique (CAH)

La classification ascendante hiérarchique consiste à partir des classes individuelles (singletons) pour faire des regroupements successifs. La technique s'illustre bien à l'aide du diagramme ci-dessous (voir Figure 5 et 6) appelé *arbre de classification* ou *dendogramme*.

De façon spécifique, si on a n individus, la démarche est la suivante:

1. On considère les classes individuelles;

2. On construit la matrice des distances entre les individus pris deux à deux;

3. On regroupe les deux individus les plus proches au sens de la distance choisie;

4. Ce groupe obtenu sera considéré comme un individu. On a maintenant $n-1$ éléments à classifier;

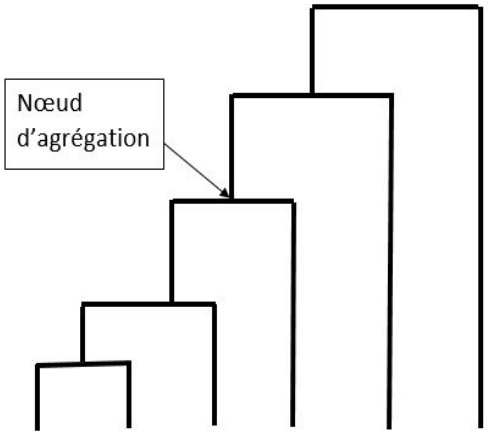

Figure 5: Arbre de classification ou dendogramme

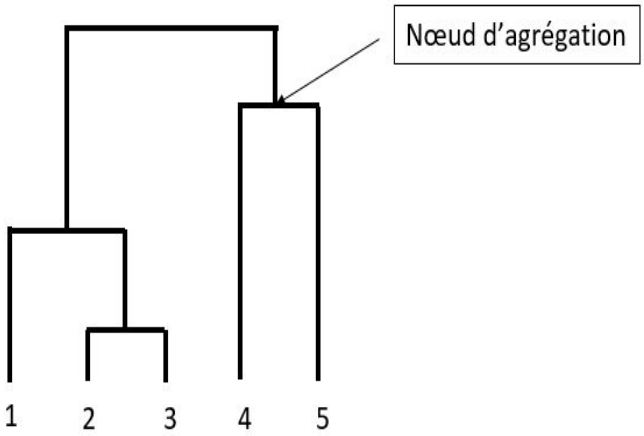

Figure 6: Arbre de classification ou dendogramme

5. On reprend l'étape 2 avec les $n-1$ éléments;

6. on continue le processus jusqu'à ce que tous les individus soient regroupés en une seule classe.

Lorsque l'on finit l'algorithme ci-dessus, il faut obtenir les classes. On calcule alors les indices d'agrégation à chaque noeud d'agrégation (méthode de Ward avec la distance euclidienne). Le principe de ce calcul est qu'à chaque agrégation nous devons avoir un gain (ou perte) minimal d'inertie inter-classe. L'indice d'agrégation d'un noeud qui connecte deux classes est la perte d'inertie inter-classe résultant de leur regroupement. Cette perte d'inertie s'obtient de la façon suivante: soit A et B deux groupes issus de l'ensemble de la population. Soit g_A et g_B les centres de gravité respectifs de A et de B. Le centre de gravité de la population totale est G. On désigne par P_A et P_B les poids respectifs des groupes A et B dans la population totale. Le centre de gravité de

l'agrégation de A et B est

$$g_{AB} = \frac{P_A g_A + P_B g_B}{P_A + P_B}.$$

Lorsque A et B ne sont pas encore regroupés, l'inertie inter-classe est égale à

$$P_A D^2 (g_A, G) + P_B D^2 (g_B, G).$$

Après le regroupement, cette inertie inter-classe devient

$$(P_A + P_B) D^2 (g_{AB}, G).$$

Donc la perte d'inertie inter-classe après regroupement est

$$P_A D^2 (g_A, G) + P_B D^2 (g_B, G) - (P_A + P_B) D^2 (g_{AB}, G).$$

On peut montrer que cette perte est égale à

$$\frac{P_A P_B}{P_A + P_B} D^2 (g_A, g_B).$$

On peut donc reprendre l'algorithme de mise en oeuvre de la CAH comme suit:

1. On part avec les classes individuelles;
2. On cherche les deux classes ayant l'indice d'agrégation minimal. On agrège ces deux classes;
3. Si la classe obtenue est l'ensemble des individus de la population totale, on arrête; sinon on retourne à l'étape 2;

Dans la pratique, on peut construire un histogramme des indices d'agrégation ou la coupure du dendogramme (voir Figures 7, 8 et 9).

Exemple. On considère le tableau de données suivant portant sur cinq individus et deux variables quantitatives X et Y. Avec le critère du saut maximal sur la distance euclidienne, effectuer la CAH et représenter le dendogramme. Extraire une classification à deux classes et donner la décomposition en inertie.

	X	Y
Individu 1	0	0
Individu 2	3	3
Individu 3	9	0
Individu 4	3	6
Individu 5	9	8

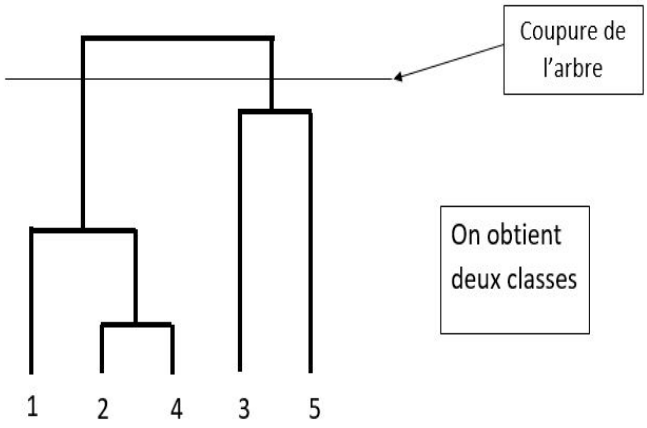

Figure 7: Coupure du dendogramme

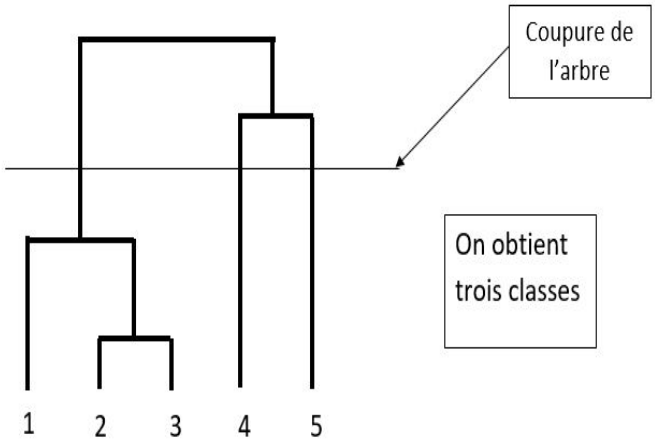

Figure 8: Coupure du dendogramme

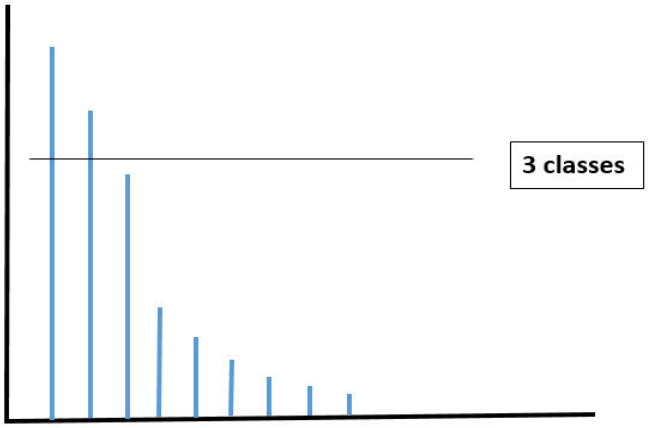

Figure 9: Coupure du dendogramme

Solution. Désignons par I_i l'individu i. On part des classes suivantes: $\{I_1\}$, $\{I_2\}$, $\{I_3\}$, $\{I_4\}$, $\{I_5\}$. La matrice des distances est la suivante.

	$\{I_1\}$	$\{I_2\}$	$\{I_3\}$	$\{I_4\}$	$\{I_5\}$
$\{I_1\}$	0	$\sqrt{18}$	$\sqrt{81}$	$\sqrt{45}$	$\sqrt{145}$
$\{I_2\}$		0	$\sqrt{45}$	$\sqrt{9}$	$\sqrt{61}$
$\{I_3\}$			0	$\sqrt{72}$	$\sqrt{64}$
$\{I_4\}$				0	$\sqrt{40}$
$\{I_5\}$					0

La matrice nous permet de regrouper $\{I_2\}$ et $\{I_4\}$. On a maintenant les classes $\{I_1\}$, $\{I_2, I_4\}$, $\{I_3\}$ et $\{I_5\}$. La nouvelle matrice des distances est

	$\{I_1\}$	$\{I_2, I_4\}$	$\{I_3\}$	$\{I_5\}$
$\{I_1\}$	0	$\sqrt{45}$	$\sqrt{81}$	$\sqrt{145}$
$\{I_2, I_4\}$		0	$\sqrt{72}$	$\sqrt{61}$
$\{I_3\}$			0	$\sqrt{64}$
$\{I_5\}$				0

On peut regrouper $\{I_1\}$ et $\{I_2, I_4\}$. On a maintenant les classes $\{I_1, I_2, I_4\}$, $\{I_3\}$ et $\{I_5\}$. La nouvelle matrice des distances est

	$\{I_1, I_2, I_4\}$	$\{I_3\}$	$\{I_5\}$
$\{I_1, I_2, I_4\}$	0	$\sqrt{81}$	$\sqrt{145}$
$\{I_3\}$		0	$\sqrt{64}$
$\{I_5\}$			0

On regroupe $\{I_3\}$ et $\{I_5\}$. On a maintenant les classes $\{I_1, I_2, I_4\}$ et $\{I_3, I_5\}$. La nouvelle matrice des distances est

	$\{I_1, I_2, I_4\}$	$\{I_3, I_5\}$
$\{I_1, I_2, I_4\}$	0	$\sqrt{145}$
$\{I_3, I_5\}$		0

On regroupe maintenant $\{I_1, I_2, I_4\}$ et $\{I_3, I_5\}$ et l'algorithme s'arrête puisqu'on a maintenant une seule classe qui est la population entière. Le dendogramme est le suivant (Figure 10). Il faut calculer les indices d'agrégation. Soit N_1 le noeud issu de l'agrégation de $\{I_2\}$ et $\{I_4\}$, N_2 le noeud de l'agrégation de $\{I_1\}$ et $\{I_2, I_4\}$, et N_3 le noeud de l'agrégation de $\{I_3\}$ et $\{I_5\}$. Soit respectivement G_1, G_2 et G_3 les centres de gravité des classes formées aux noeuds N_1, N_2 et N_3. On désigne par G le centre de gravité de la population entière. On a $G_1 = (3, 9/2)$, $G_2 = (2, 3)$, $G_3 = (9, 4)$ et

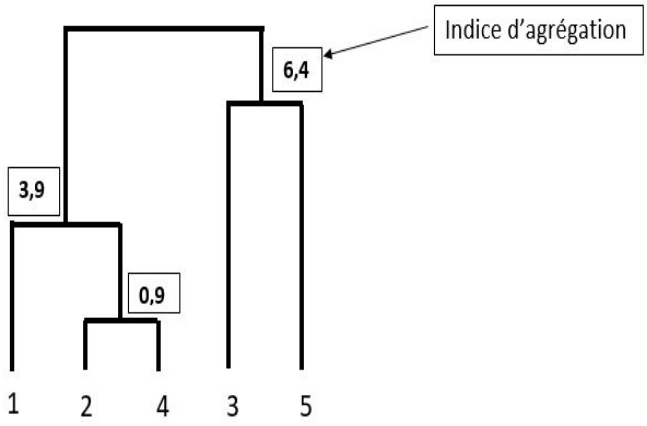

Figure 10: Arbre de classification de l'exemple

$G = (24/5, 17/5)$. L'indice d'agrégation au noeud N_1 est

$$I_{N_1} = \frac{1}{5}D^2(I_2, G) + \frac{1}{5}D^2(I_4, G) - \frac{2}{5}D^2(G_1, G) = 0,9.$$

L'indice d'agrégation au noeud N_2 est

$$I_{N_2} = \frac{1}{5}D^2(I_1, G) + \frac{2}{5}D^2(G_1, G) - \frac{3}{5}D^2(G_2, G) = 3,9.$$

L'indice d'agrégation au noeud N_3 est

$$I_{N_3} = \frac{1}{5}D^2(I_3, G) + \frac{1}{5}D^2(I_5, G) - \frac{2}{5}D^2(G_3, G) = 6,4.$$

En examinant le dendogramme, une partition à deux classes nous donne les classes suivantes: $\{I_1, I_2, I_4\}$ et $\{I_3, I_5\}$. En réalité l'arbre est coupé comme le montre la figure 11.

Les limites de la CAH

La CAH présente les limites suivantes:

- Il y a trop de clacul;
- Les résultats dépendent de la distance utilisée et de la procédure d'utilisation de cette distance.

Remark 2.1. *La classification descendante hiérarchique utilise simplement la démarche inverse de la CAH. Elle part de la population entière et procède par désagrégation jusqu'à obtenir les classes individuelles. Il s'agit d'un clustering divisif. La classification descendante est très peu utilisée.*

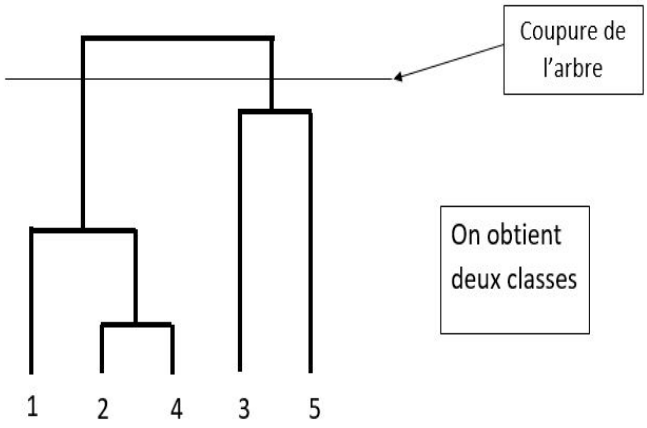

Figure 11: Coupure du dendogramme de l'exemple

2.5.2 La classification non hiérarchique ou le partitionnement

Il s'agit d'obtenir une partition en k classes en k classes en optimisant un critère donné. L'objectif est de minimiser l'inertie intra-classe ou de maximiser l'inertie inter-classe. Pour ce type de classification, le nombre de classes doit être connu à l'avance.

A l'intérieur la méthode de classification non hiérarchique on distingue plusieurs variantes que sont, entre autres:

- la méthode des centres mobiles;
- la méthode des K-means;
- la méthodes des nuées dynamiques;
- etc.

La méthode des centres mobiles. L'algorithme de cette méthode se déroule de la façon suivante:

1. On choisit le nombre de classes de la partition. Soit K ce nombre.
2. On choisit K individus comme centres d'inertie de classes.
3. On met chacun des autres individus dans la classe dont il est plus proche du centre. Pour le déroulement de toute cette étape, le cantre de chaque classe est celui choisi à l'étape 2.
4. On recalcule maintenant le centre d'inertie de chaque classe.
5. On réaffecte les individus à chaque classe en tenant compte des nouveaux centres d'inertie.

6. On retourne à l'étape 4 et le processus se poursuit jusqu'à ce qu'aucun individu ne change plus de classe.

La méthode des K-means. Ici, on utilise la même procédure que celle des centres mobiles à la différence que le centre d'inertie de chaque classe est recalculé après qu'un individu a été affecté à une classe. On n'attend plus que tous les individus soient d'abord affectés avant d'effectuer un recentrage des classes.

La méthode des nuées dynamiques. Pour cette méthode, une classe n'est plus représentée par son centre, mais par un groupe d'individus de cette classe. Ce groupe est appelé *noyau*. Chaque élément du noyau est appelé *étalon* ou *forme forte*.

Les limites de la méthode de partitionnement

- On est obligé de fixer à priori le nombre de classes;
- La convergence de l'algorithme dépend du choix des centres d'inertie ou noyaux initiaux.

2.5.3 La méthode mixte de classification

Cette méthode combine la classification hiérarchique et la partitionnement. Le principle est simple. On réalise d'abord une CAH pour choisir le nombre de classes. On déroule ensuite la méthode de partitionnement en prenant pour classes initiales celles obtenues par la CAH. La méthode mixte permet d'obtenir une classification consolidée.

2.5.4 Mesure de la qualité de la classification

$$Qualité = \frac{Inertie\ inter}{Inertie\ totale} \times 100$$

2.6 Application

(Computer Lab avec l'enseignant. L'étudiant doit installer R et SPAD).

3 Chapitre 1: L'Analyse Factorielle Discriminante (AFD)

3.1 Motivation

La classification nous permet d'dentifier dans une population les différents sous-groupes qui s'y forment selon les ressemblances et les dissemblances. En d'autres termes, la classification nous révèle qu'il y a des groupes ou associations qui se forment et permet de caractériser ces groupes (c'est-à-dire décrire les classes identifiées).

Toutefois, on peut être dans un autre cas de figure où nous savons déjà qu'il y a un certain nombre de groupes au sein de la population, mais nous ne savons que des caractéristiques **non observables** de chaque groupe. Par exemple, prenons un ensemble de clients d'une banque qui ont déposé de dossier de demande de prêt bancaire. La banque sait que dans cet ensemble il y a de bons clients (c'est-à-dire ceux qui vont rembourser leur prêt) et les mauvais clients. Mais quand on prend un client la banque ne peut pas savoir s'il va bien rembourser son prêt. Cette caractérisque n'est pas observable. Alors, la question est savoir comment on peut caractériser les groupes par des caractéristiques observables, de façon à pouvoir prédire le groupe (bon ou mauvais) d'un client quand on connaît ses caractéristiques observables. Ce genre de problème ne se pose pas seulement aux banques. Il se pose aussi à une ambassade qui doit octroyer des visas à des candidats à l'immigration (qui va bien ou mal se comporter dans le pays d'accueil?), à une école qui doit sélectionner des candidats à enroller dans un programme (qui peut supporter ou non le programme?), à une compagnie d'assurance qui conclure une police d'assurance accident (qui est à faible, moyen ou haut risque?), etc.

Contrairement à la classification, l'AFD va au-delà d'une simple description des groupes. Elle est aussi une méthode explicative. En réalité l'AFD consiste à identifier ce qui explique l'appartenance à tel ou tel groupe. Il s'agit d'étudier comment les individus se retrouvent dans tel ou tel groupe. Par quel mécanisme les groupes sont-ils constitués? Pourquoi un individu est dans ce groupe, et non dans l'autre? Comment les variables sont-elles combinées pour que les individus soient dans tel ou tel groupe? En résumé, l'AFD a deux objectifs.

3.2 Les objectifs de l'AFD

L'AFD poursuit deux objectifs que sont:

- un objectif descriptif: il s'agit de caractériser, c'est-à-dire de décrire les groupes. Quelles caractéristiques sont présentées par les individus d'un groupe donné? Qu'est-ce qui différencie (discrimine) les groupes?

- un objectif décisionnel: être capable de prédire le groupe d'appartenance d'un individu dont on connaît les caractéristiques. En d'autres termes, il s'agit de pouvoir décider du groupe auquel on peut affecter un individu dont on observe certaines caractéristiques pertinentes, avec bien sûr un certain risque de se tromper. Ces caractéristiques pertinentes (variables) qui sont utilisées pour atteindre cet objectif sont appelées *prédicteurs*, ou variables indépendantes, ou variables expliquées.

3.3 Présentation de la méthode

Considérons une population pour laquelle nous connaissons une variable qualitative à K modalités y_1, y_2, $\cdots$, y_K. Cette variable qualitative nous permet de définir les groupes au sein de la population. On dira que les individus qui présentent la modalité y_k sont dans le groupe k, $k = 1, 2, \cdots, K$. On a donc K groupes. Cette variable qualitative qui définit les groupes est appelée *variable prédicte*, ou *variable dépendante*, ou *variable expliquée*. En plus de la variable dépendante, on dispose d'information sur P autres variables quantitatives qui seront les prédicteurs. On peut avoir une variable qualitative comme prédicteur. Dans ce cas cette variable doit être orthogonalisée. Désignons par X_1, X_2, $\cdots$, X_P ces prédicteurs.

L'AFD va consister à calculer des fonctions discriminantes, c'est-à-dire des combinaisons linéaires des prédicteurs. Si nous avons K groupes, on calcule $K-1$ fonctions discriminantes. Soit F_1, F_2, $\cdots$, F_{K-1}. On a:

$$F_k = b_{Ok} + b_{1k}X_1 + b_{2k}X_2 + \cdots + b_{Pk}X_P \quad k = 1, 2, \cdots, K-1.$$

Pour l'individu i on va calculer son *score* F_{ik} par la fonction discriminante F_k.

$$F_{ik} = b_{Ok} + b_{1k}X_{1i} + b_{2k}X_{2i} + \cdots + b_{Pk}X_{Pi} \quad k = 1, 2, \cdots, K-1.$$

avec X_{ji} la valeur prise par la variable X_j chez l'individu i. b_{Ok}, b_{1k}, $\cdots$, et

$$b_{Pk}$$

sont les coefficients ou les poids de la fonction discriminante F_k. L'objectif de l'AFD est déterminer ces coefficients afin de pouvoir calculer les scores de chaque individu. On spécifie les fonctions discriminantes (c'est-à-dire calculer les coefficients) de façon à maximiser la variance inter (relativement à la variance intra) dans l'espace des fonctions discriminantes. En d'autres termes, les coefficients recherchés sont ceux qui permettent de maximiser cette variance inter.

Soit u le vecteur des coefficients pour une fonction discriminante donnée. Soit respectivement V, W, et B les matrices des variance-covariances total, intra et inter

dans l'espace des données initiales. On a $V = B + W$. On peut dont écrire:

$$u'Vu = u'Bu + u'Wu.$$

$u'Vu$, $u'Bu$, et $u'Wu$ sont respectivement la variance totale, la variance inter et la variance intra dans l'espace de la fonction discriminante. L'objectif est de maximiser $u'Bu$ et de minimiser $u'Wu$. Pour maximiser $u'Bu$, le programme est

$$\begin{cases} \max\limits_{u} & u'Bu \\ s.c & u'u = 1 \end{cases}$$

Le Lagrangien est

$$\mathcal{L} = u'Bu + \lambda(1 - u'u)$$

La condition du premier ordre donne $Bu = \lambda u$. Donc u est un vecteur de B associé à la valeur propre λ. Pour minimiser $u'Wu$, le programme est

$$\begin{cases} \min\limits_{u} & u'Wu \\ s.c & u'u = 1 \end{cases}$$

Le Lagrangien est

$$\mathcal{L} = u'Wu + \lambda(1 - u'u)$$

La condition du premier ordre donne $Wu = \lambda u$. u est aussi un vecteur propre de W. Ce qui est impossible. On va alors plutôt chercher à résoudre le programme

$$\max_{u}\left(\frac{u'Bu}{u'Vu}\right) \quad \text{ou} \quad \max_{u}\left(\frac{u'Bu}{u'Wu}\right)$$

Le premier programme équivaut à

$$\begin{cases} \max\limits_{u} & u'Bu \\ s.c & u'Vu = 1 \end{cases}$$

On voit facilement que u est un vecteur propre de $V^{-1}B$ associé à la valeur propre λ.

Pour le programme $\max\limits_{u}\left(\frac{u'Bu}{u'Wu}\right)$, la solution est que u est le vecteur propre de $W^{-1}B$.

Remark 3.1. *On peut montrer que les vecteurs propres u de $W^{-1}B$ sont identiques à ceux de $V^{-1}B$. En effet, supposons que $W^{-1}Bu = \theta u$. On sait aussi que $V^{-1}Bu = \lambda u$.*

On peut écrire:

$$\begin{aligned} V^{-1}Bu = \lambda u &\Rightarrow Bu = \lambda Vu \\ &\Rightarrow Bu = \lambda\left(B+W\right)u \\ &\Rightarrow Bu - \lambda Bu = \lambda Wu \\ &\Rightarrow \left(1-\lambda\right)Bu = \lambda Wu \\ &\Rightarrow W^{-1}Bu = \frac{\lambda}{1-\lambda}u \\ &\Rightarrow \theta = \frac{\lambda}{1-\lambda}. \end{aligned}$$

λ *est appelé* ***Lambda de Wilks****.*

Recherche des axes discriminants

Comme P est le nombre de variables indépendantes et K est le nombre de groupes, le nombre d'axes discriminants est $\min(P, K-1)$. Mais généralement, le nombre de groupes est largement inférieur au nombre de prédicteurs. Donc le nombre d'axes discriminants (et alors de fonctions discriminantes) est effectivement $K-1$. On peut facilement montrer que $\frac{u'Bu}{u'Vu} = \lambda$. En d'autres termes, λ est le pouvoir discriminant de l'axe. Même si on distingue $K-1$ fonctions discriminantes, l'on n'a pas besoin de choisir toutes les $K-1$. On va plutôt utiliser les fonctions discriminantes ayant un pouvoir discriminant élevé. On classe alors les axes en fonction des valeurs propres. Généralement, deux fonctions discriminantes peuvent suffir.

Qualité de la discrimination

Pour apprécier la qualité de la discrimination on utilisation un indicateur appelé le *taux d'erreur de classement (TEC).*

$$TEC = \frac{Nombre\ d'individus\ mal\ classés}{Nombre\ total\ d'individus} \times 100$$

En pratique pour obtenir un taux d'erreur de classement efficace, on utilise un échantillon témoin.

3.4 Application

(Computer Lab avec l'enseignant. L'étudiant doit installer R et SPAD).