Support de cours

ANALYSE DES DONNEES I

CHAPITRE 4 : L'ANALYSE DES CORRESPONDANCES MULTIPLES

Iphygénie SARR
Ingénieure statisticienne économiste

Ecole Nationale de la Statistique et de l'Analyse Economique Pierre Ndiaye
Mai 2023

OBJECTIF DU CHAPITRE

Nous consacrerons ce chapitre à l'Analyse des Correspondances multiples (ACM). Cette méthode apparait comme la plus intéressante des extensions de l'Analyse factorielle des correspondances (AFC).

PLAN DU CHAPITRE

01.

Données, tableaux et notations

Données

L'appellation « analyse factorielles des correspondances multiples » se justifie ainsi : alors que l'AFC étudie la relation entre deux caractères I et J observés sur une population donnée, l'ACM étudie les relations entre un nombre quelconque de caractère J_1,..., J_m.

Données

- Supposons que nous considérons un ensemble I composé de N individus, sur lesquels nous observons m caractères qualitatifs J_k avec $k \epsilon [\![1, m]\!]$.
- Supposons que chaque variable qualitative J_k possède K_k modalités et qu'un individu ne peut posséder qu'une et une seule modalité de chacune de ces variables qualitatives.
- Notons $J = \cup_{k=1}^{m} J_k$ l'ensemble des modalités possibles.
- Et $Card\, J = \sum_{k=1}^{m} K_k = K$.

Données

L'ACM constitue une méthode très souple

- Les variables J_k sont nécessairement qualitatives. Cependant, il est toujours possible de transformer une variable quantitative en variable qualitative en divisant son intervalle de variation en classes d'équivalences successives.

- Ainsi le champ des applications possibles de l'ACM apparaît comme très vaste, puisqu'on peut utiliser cette méthode pour analyser des variables d'un type quelconque observées sur une population donnée.

Données

Mais attention…

- Cependant, le découpage du domaine de variation d'une variable quantitative en classes comporte un biais inévitable : celui du choix du nombre de classes (qui ne doit pas être pris à la légère).
- Il faut donc prendre la prendre la précaution de représenter l'histogramme de la distribution d'une variable avant de la découper en classe mais aussi, de se référer à la théorie.

Données

Aussi …

Une enquête un tant soit peu complexe comporte toujours des questions relatives à plusieurs aspects différents. Par exemple, dans une population de ménages, certaines variables seront relatives aux revenus, d'autres au niveau d'instruction, d'autres au statut social, etc.

Bien que l'ACM le permette, la prise en compte simultanée de tous ces aspects risque de conduire à un résultat excessivement compliqué et peu lisible, donc difficilement interprétable. Ainsi l'ACM est certes une méthode souple, mais elle n'est pas un « fourre-tout ».

Il est donc préférable de ne prendre en compte dans une analyse que les variables relatives à un aspect particulier.

Tableaux

De la même manière qu'une ACP ou une AFC, l'ACM utilise comme *input* un tableau de données. Mais pour son cas, deux (02) formes de transcription des données sont possibles :

- un tableau logique encore appelé tableau disjonctif complet ;
- ou un tableau de Burt.

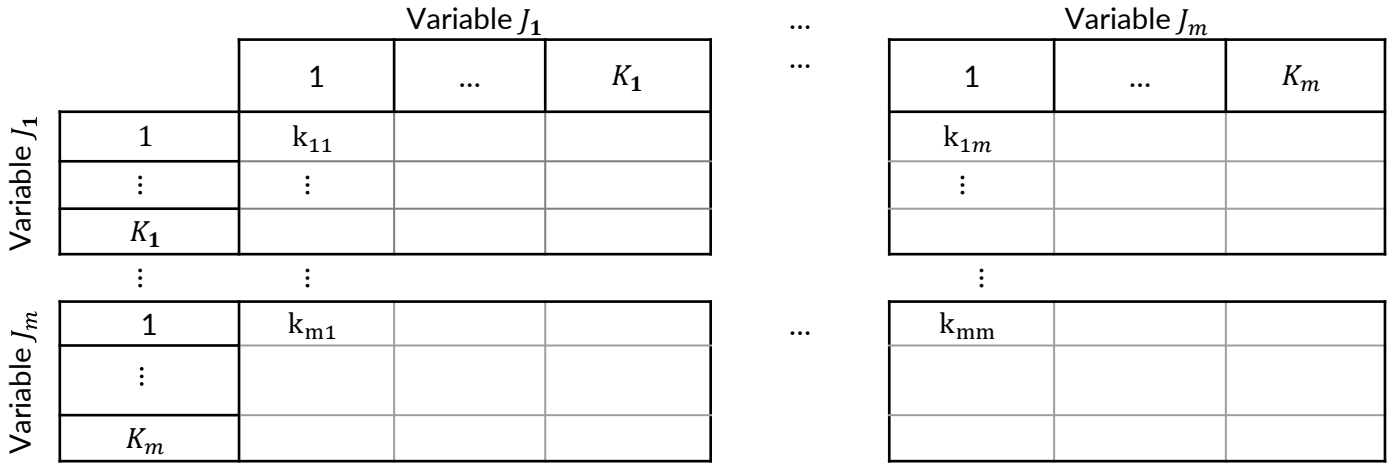
Tableau logique

Il est aussi appelé tableau disjonctif complet. Dans la case (i, j) se trouve le nombre k_{ij} égal à 1 si l'individu i possède la modalité j, égal à 0 sinon.

variables →	Variable J_1			...	Variable J_k			...	Variable J_m			Profil→
modalités →	1	...	K_1	...	1	...	K_k	...	1	...	K_m	
1	k_{11}			...	k_{1j}			...			k_{1K}	m
⋮	⋮			⋮	⋮			⋮			⋮	m
i	k_{i1}			...	k_{ij}			...			k_{iK}	m
⋮	⋮			⋮	⋮			⋮			⋮	m
N	k_{N1}			...	k_{Nj}			...			k_{NK}	m
Profil →	k_1	k_2	k_3	...	k_j	...					k_K	Nm

Tableau de Burt

Il s'agit d'un tableau contenant l'ensemble des tableaux de contingence entre les variables prises deux à deux. Un tableau de Burt est ainsi une juxtaposition de tableaux de contingence.



Tableaux

Une ACM peut donc être réalisé selon au moins deux (02) points de vue : l'un s'appuyant sur l'analyse du tableau disjonctif et l'autre sur le tableau de Burt.

- Dans la première optique, une modalité est vu en tant que variable indicatrice définie sur l'ensemble des individus.
- Dans la deuxième optique, une modalité est vu en tant que classe d'individus dont on connaît la répartition sur l'ensemble des autres modalités.

Tableau logique

	X_1		X_2		X_3		
	1	2	1	2	1	2	3
1	0	1	1	0	0	1	0
2	1	0	0	1	1	0	0
3	0	1	0	1	0	0	1
4	0	1	0	1	0	1	0
5	1	0	0	1	1	0	0
6	0	1	1	0	0	0	1

Tableau de Burt

		X1		X2		X3		
		1	2	1	2	1	2	3
X1	1			0	2	2	0	0
	2			2	2	0	2	2
X2	1					0	1	1
	2					2	1	1

Notations

Tous les résultats que nous établirons dans la suite, sont relatifs à l'analyse des tableaux logiques. Cependant, on montre qu'il existe une équivalence entre l'analyse d'un tableau logique avec celle d'un tableau de Burt Volle (1981).

Vu sous cet angle (i.e. l'analyse d'un tableau logique), l'ACM constitue une l'application de l'analyse factorielle des correspondances.

On pose à partir du tableau logique :

$$f_{ij} = \frac{k_{ij}}{Nm} = \begin{cases} 0 \; si \; l'individu \; i \; ne \; possède \; pas \; la \; modalité \; j \\ \frac{1}{Nm} si \; l'individu \; i \; possède \; la \; modalité \; j \end{cases}$$

Notations

Sur la base de cela, on précise les notations suivantes (analogues à celles obtenues en AFC) :

- $\mathrm{f_i} = \frac{m}{Nm} = \frac{1}{N}$
- $\mathrm{f_j} = \frac{k_j}{Nm}$
- $\mathrm{f_j^i} = \frac{\mathrm{f_{ij}}}{\mathrm{f_i}} = \frac{\mathrm{k_{ij}}}{m} = \begin{cases} 0 \; si \; l'individu \; i \; ne \; possède \; pas \; la \; modalité \; j \\ \frac{1}{m} \; sinon \end{cases}$
- $\mathrm{f_i^j} = \frac{\mathrm{f_{ij}}}{\mathrm{f_j}} = \frac{\mathrm{k_{ij}}}{k_j} = \begin{cases} 0 si \; l'individu \; i \; ne \; possède \; pas \; la \; modalité \; j \\ \frac{1}{k_j} sinon \end{cases}$

Notations

On peut associer au tableau logique deux nuages de points : celui des individus $\mathcal{N}(\mathrm{I})$ et celui des variables $\mathcal{N}(\mathrm{J})$.

Objectifs du point de vue individus : réaliser une typologie des individus en s'appuyant sur une notion de ressemblance.

- Deux individus se ressemblent d'autant plus qu'ils possèdent un grand nombre de modalités en commun.
- Mais en ACM, on est généralement confronté à un grand nombre d'individus. Ainsi, les individus sont étudiés au travers des classes définies par les variables
- Exemple : Dans les enquêtes d'opinion, on peut effectuer une analyse comparative des réponses dans les catégories femmes, jeunes, retraités.

Notations

Objectifs du point de vue modalités : les mêmes objectifs sont visés, faire le bilan des ressemblances entre les modalités.

- Toutefois on parlera d'association entre les modalités et non de ressemblance.
- Deux modalités s'associent d'autant plus qu'elles :
  1. sont présentes ou absentes simultanément chez un grand nombre d'individus.
  2. s'associent beaucoup ou peu aux mêmes modalités.

02.

Analyse du nuage des individus

Métrique

$\mathcal{N}(\mathrm{I})$ est composé des N points X^{i} situés dans l'espace $\mathbb{R}^{\mathrm{K}}$ dotés chacun de la masse $\mathrm{f_i}$ et de coordonnée courante :

$$\mathrm{x_{ij}} = \mathrm{f_j^i}$$

On use dans $\mathcal{N}(\mathrm{I})$ de la métrique du χ^2 centrée sur $\mathrm{f_J}$, de sorte que :

$$\mathrm{d^2}\left(\mathrm{X^i}, \mathrm{X^{i\prime}}\right) = \sum_{\mathrm{j}} \frac{1}{\mathrm{f_j}} (\mathrm{f_j^i} - \mathrm{f_j^{i\prime}})^2$$

$$\mathrm{d^2}\left(\mathrm{X^i}, \mathrm{X^{i\prime}}\right) = \sum_{\mathrm{j}} \frac{\mathrm{Nm}}{k_{\mathrm{j}}} (\frac{\mathrm{k_{ij}}}{m} - \frac{\mathrm{k_{i\prime j}}}{m})^2$$

$$\mathrm{d^2}\left(\mathrm{X^i}, \mathrm{X^{i\prime}}\right) = \frac{1}{m} \sum_{\mathrm{j}} \frac{\mathrm{N}}{k_{\mathrm{j}}} (\mathrm{k_{ij}} - \mathrm{k_{i\prime j}})^2$$

Métrique

Que signifie la distance entre les points X^i et $X^{i\prime}$ de $\mathcal{N}(\mathrm{I})$?

La distance entre ces deux points quantifie la proximité entre les deux individus qu'ils représentent et s'interprète en termes de ressemblance.

Ainsi, l'on dira que deux individus se ressemblent s'ils ont globalement les mêmes réponses ou profils : ils présentent globalement les mêmes modalités.

On le remarque facilement avec l'expression de la distance $\mathrm{d}^2(\mathrm{X}^{\mathrm{i}}, \mathrm{X}^{\mathrm{i}\prime})$ qui croît avec le nombre de modalités qui diffèrent pour les individus i et i'.

Une autre remarque intéressante

Une modalité j intervient dans cette distance avec le poids $\mathrm{N}/k_{\mathrm{j}}$, inverse de sa fréquence. La présence d'une modalité rare éloigne donc son ou ses possesseurs de tous les autres individus.

Inertie du nuage

Le centre de gravité de $\mathcal{N}(\mathrm{I})$ est le point G de coordonnées courantes :

$$\mathrm{G} = \sum_i \mathrm{f_i} X^i$$

$$\text{Avec } \mathrm{g_j} = \sqrt{f_j} = \sqrt{\frac{k_j}{Nm}}$$

L'inertie totale de $\mathcal{N}(\mathrm{I})$ est alors :

$$In_G(I) = \sum_i f_i \left\| X^i - G \right\|^2$$

$$In_G(I) = \sum_{ij} \frac{\left(f_{ij} - f_i f_j\right)^2}{f_i f_j}$$

03.

Analyse du nuage des variables

Métrique

$\mathcal{N}(\mathrm{J})$ est composé des K points Y^{j} situés dans l'espace $\mathbb{R}^{\mathrm{N}}$ dotés chacun de la masse $\mathrm{f_j}$ et de coordonnée courante :

$$\mathrm{y_{ij}} = \mathrm{f_i^j}$$

On use dans $\mathcal{N}(\mathrm{J})$ de la métrique du χ^2 centrée sur $\mathrm{f_I}$, de sorte que :

$$\mathrm{d^2}\left(\mathrm{Y^j}, \mathrm{Y^{j\prime}}\right) = \sum_{\mathrm{i}} \frac{1}{\mathrm{f_i}} (\mathrm{f_i^j} - \mathrm{f_i^{j\prime}})^2$$

$$\mathrm{d^2}\left(\mathrm{Y^j}, \mathrm{Y^{j\prime}}\right) = \sum_{\mathrm{i}} N \left(\frac{\mathrm{k_{ij}}}{k_j} - \frac{\mathrm{k_{ij\prime}}}{k_{j\prime}}\right)^2$$

$$\mathrm{d^2}\left(\mathrm{Y^j}, \mathrm{Y^{j\prime}}\right) = N \sum_{\mathrm{i}} \left(\frac{\mathrm{k_{ij}}}{k_j} - \frac{\mathrm{k_{ij\prime}}}{k_{j\prime}}\right)^2$$

Métrique

Que signifie la proximité ou l'éloignement de deux points Y^j et $Y^{j'}$ de $\mathcal{N}(\mathbf{J})$?

La distance entre ces deux points quantifie la proximité entre les deux modalités qu'ils représentent et s'interprète en termes d'association pour les modalités de variables différentes.

Cette distance croît avec le nombre d'individus possédant une et une seule des deux modalités j et j'.

Et, si ces deux modalités sont possédées par les mêmes individus alors elles sont confondues et la distance qui les sépare est nulle.

Quelques remarques intéressantes

- Suivant notre hypothèse selon laquelle *un individu ne peut posséder qu'une et une seule modalité de chacune de ces variables qualitatives*, les modalités d'une même variable sont obligatoirement éloignées l'une de l'autre.
- Une modalité rare est éloignée de toutes les autres modalités.

Inertie du nuage

Le centre de gravité de $\mathcal{N}(\mathrm{J})$ est le point H de coordonnées courantes :

$$\mathrm{H} = \sum_{j} \mathrm{f_j} \mathrm{Y}^{j}$$

$$\text{Avec } \mathrm{h}_i = \sqrt{f_i} = \sqrt{\frac{1}{N}}$$

L'inertie totale de $\mathcal{N}(\mathrm{J})$ est alors :

$$In_H(J) = \sum_{j} f_j \left\| Y^j - H \right\|^2$$

$$In_H(J) = \sum_{ij} \frac{\left(f_{ij} - f_i f_j\right)^2}{f_i f_j}$$

04.

Cas pratique d’une ACM avec le logiciel R studio

Quelques remarques sur l'interprétation

Globalement l'interprétation de l'ACM se fait suivant le même procédé qu'une analyse factorielle classique. Néanmoins, il est pertinent de préciser certaines singularités.

1. On montre que l'inertie du nuage (individu ou modalité peu importe) est donnée par :

$$In_G(I) = In_H(J) = \frac{K}{m} - 1$$

L'inertie totale dépend uniquement du nombre moyen de modalités par variable $\frac{K}{m}$.

Ainsi en ACM, il est conseillé de ne pas avoir un trop grand nombre de modalités/variables.

Quelques remarques sur l'interprétation

2. Un second point est l'attention qu'il faut porter aux modalités rares. En effet, la contribution d'une modalité à la formation d'un axe est donnée par :

$$CTR_j = \frac{\frac{1}{m}\left(1 - \frac{k_j}{N}\right)}{In_G(I)}$$

Il apparait ainsi que la contribution d'une modalité est d'autant plus grande que son poids est faible. C'est pourquoi, dans une ACM, il faut éviter les modalités de faibles poids car elles peuvent avoir beaucoup d'importance dans l'analyse des axes factoriels.

Quelques remarques sur l'interprétation

3. L'ACM dans son analyse a tendance à faire fi des variables et à mettre le focus sur les modalités. Mais à partir des résultats des contributions des modalités, on peut en déduire celles des variables :

$$CTR_\alpha(J_k) = \sum_{j \in J_k} CTR_\alpha(j)$$

Ce sont ces dernières quantités que l'on considérera pour déterminer les variables qui ont joué le plus grand rôle dans la détermination de l'axe α.

Présentation de la base de données

Population : un échantillon d'électeur des USA

Individu statistique : un électeur

Nombre d'individus : 24

Nombre de variables : 11

Nature des variables : qualitatives

SYNTHESE

Quelques points essentiels à retenir :

- l'ACM est une généralisation de l'analyse factorielle des correspondances ;
- son champ des applications possibles apparaît comme très vaste ;
- cependant elle n'est pas un « fourre-tout » ;
- son interprétation doit toujours être accompagné de l'utilisation des aides à l'interprétation et d'une bonne connaissance des données utilisées.

FIN DU CHAPITRE

MERCI POUR VOTRE ATTENTION !