Analyse des Correspondances Multiples
(ACM)

Planning

01

Introduction

02

Méthodologie

03

Application de la partie théorique + Interpretation des résultats

04

Application avec Python

05

Conclusion

01 Introduction

L’analyse des correspondances multiples (ACM) est une méthode descriptive qui permet d’étudier les liaisons entre plusieurs variables qualitative. Elle généralise l’AFC qui est limitée à l’analyse des liens entre deux variables qualitatives. Elle est souvent utilisée pour interpréter les résultats d’une enquête d’opinion.

L'objectif principal de l'ACM est d'explorer les relations entre les catégories des différentes variables catégorielles.
Elle permet de visualiser graphiquement comment ces catégories sont positionnées dans un espace factoriel défini par les axes de l’ACM.
L'ACM aide à réduire la dimension des données en résumant l'information importante dans un espace de dimension inférieure.

Elle permet de visualiser les contributions des individus et des variables à la variabilité totale des données, comme elle facilite l’interprétation en réduisant la complexité des données catégorielles.
Les résultats visuels et statistiques aident à bien comprendre la structure sous-jacente des données.
L’ACM peut être utilisée pour prendre des décisions éclairées dans des domaines tels que le marketing, la sociologie, la biologie, etc., en mettant en évidence des associations significatives.

L’ACM porte sur une enquête, les individus sont les enquêtés et les variables sont les questions d’un questionnaire.

02 Méthodologie

Pour realiser une Acm, il faut suivre les étapes suivantes:

- ### Préparation des Données :

  Assurez-vous que vos données sont bien formatées et que les variables à analyser sont catégorielles. Si besoin, effectuez des transformations pour convertir les variables appropriées en catégories.

- ### Construction du tableau disjonctif complet

  Un tableau disjonctif complet (TDC) est un type de représentation de données qualitatives utilisé en analyse des données.
  Dans ce tableau, une variable qualitative à K modalités est remplacée par K variables binaires, chacune correspondant à une des modalités.
  K prend 1 si l'individu possède la modalité et 0 sinon.

Exemple:

Une famille est constituée d'un père, d'une mère et d'un jeune garçon. On s’intéresse aux variables "sexe" et "couleur des yeux".
Voici le tableau regroupant ces informations :

Individu	Sexe	Yeux
Père	Masculin	Marron
Mère	Féminin	Noir
Enfant	Masculin	Noir

Le tableau disjonctif complet de cette population prend la forme suivante :

Individu	Sexe M	Sexe F	Yeux M	Yeux N
Père	1	0	1	0
Mère	0	1	0	1
Enfant	1	0	0	1

Construction des tableaux de profils lignes et colonnes

- Profil ligne :
Pour chaque ligne du TDC, on divise chaque cellule par la somme de cette ligne.

Le profil moyen est calculé par cette relation : $\frac{n_k}{n*p}$

avec :
nk : la somme de la colonne k dans le TDC
n : nombre total des individus
p :total des variables/questions

- Profil colonne :
Pour chaque colonne du TDC, on divise chaque cellule par la somme de cette colonne.

Le profil moyen est calculé par cette relation : $\frac{P}{n*p} = \frac{1}{n}$

Exemple:

Profils lignes:

Individu	Sexe M	Sexe F	Yeux M	Yeux N
Père	0,5	0	0,5	0
Mère	0	0,5	0	0,5
Enfant	0,5	0	0	0,5

Profils colonne:

Individu	Sexe M	Sexe F	Yeux M	Yeux N
Père	0,5	0	1	0
Mère	0	1	0	0,5
Enfant	0,5	0	0	0,5

Construction du tableau de Burt

Un tableau de Burt est un type de représentation de données qualitatives utilisé en analyse des données .
C’est une matrice carrée symétrique. Il peut être calculée à partir du TDC:
si l'on note D, la matrice du tableau disjonctif complet, alors le tableau de Burt B est égale à la transposée de D * D.

	Sexe M	Sexe F	Yeux M	Yeux N
Sexe M	2	0	1	1
Sexe F	0	1	0	1
Yeux M	1	0	1	0
Yeux N	1	1	0	2

Calcul des distances

En ACM, le calcul des distances entre individus et entre modalités est une étape clé qui permette d'évaluer la similarité ou la dissimilarité entre individus et modalités, et elles sont utilisées pour représenter graphiquement les résultats de l'analyse.
Les distances permettent de représenter graphiquement la distribution des individus et des modalités dans l'espace factoriel, peuvent aider à interpréter la proximité ou l'éloignement entre individus ou modalités.
Les individus ou modalités proches dans l'espace factoriel sont similaires, tandis que ceux éloignés sont dissimilaires, ils sont utilisées pour calculer la contribution des individus ou modalités à l'inertie totale sur chaque axe. Cela contribue à identifier les éléments qui ont le plus d'influence sur la structure des données.

2 individus avec les mêmes modalités: distance = 0
2 individus avec beaucoup de modalités en commun : distance petite
2 individus avec peu de modalités en commun/ 1 possède une modalité rare : distance grande

On distingue les distances suivantes:
- Distance entre individus:

On a la formule suivante: $$\sum_{k=1}^{m} \frac{1}{\frac{n_k}{n*P}} \left(\frac{x_{\mathrm{ik}}}{P} - \frac{x_{\mathrm{j}k}}{P} \right)^2$$

Avec x_{ik} /p : la profil ligne correspondant à l'individu i et la modalité k.

- Distance entre modalités:

On a la formule suivante: $$\sum_{i=1}^{n} \frac{1}{\frac{1}{n}} \left(\frac{x_{ik}}{n_{\mathrm{k}}} - \frac{x_{\mathrm{jk}}}{n_{\mathrm{k}}} \right)^2$$

- Distance des individus à l'origine:

On a la formule suivante: $$\sum_{k=1}^{m} \frac{1}{\frac{n_k}{n*P}} \left(\frac{x_{\mathrm{ik}}}{P} - \frac{n_k}{n*P} \right)^2$$

- Distance des modalités à l'origine:

On a la formule suivante: $$\sum_{i=1}^{n} \frac{1}{\frac{1}{n}} \left(\frac{x_{ik}}{n_{\mathrm{k}}} - \frac{1}{n} \right)^2$$

Calcul des valeurs et vecteurs propres

C’est une étape fondamentale en ACM .Ces valeurs et vecteurs propres sont associés à la décomposition spectrale de la matrice de données, et ils fournissent des informations essentielles sur la structure de ces données dans l'espace factoriel.

Les valeurs propres représentent la quantité d'inertie (ou de variance) expliquée par chaque axe. Elles indiquent l'importance relative de chaque axe dans la représentation de la variation des données. Elles sont souvent utilisées pour décider combien d'axes à retenir.

Les vecteurs propres définissent l'orientation de chaque axe dans l'espace factoriel. Chaque vecteur propre est associé à un axe et indique comment chaque variable contribue à cet axe.

On utilise la matrice B pour les calculer.

- ## Choix des axes

Il existe plusieurs critères/formules pour choisir le nombre des axes à retenir, Pour nous, on a choisit le calcul du pourcentage cumulé d'inertie expliquée à partir des valeurs propres obtenue dans l'étape précédente d'une façon qu'on divise chaque valeur propre par la somme des valeurs propres puis on multiplie par 100 pour obtenir la proportion relative d'inertie expliquée par chaque axe, puis on choisit les axes qui capture +80% de l'inertie totale.

Calcul des coordonnées

Le calcul des coordonnées des individus et des modalités en ACM est une étape cruciale pour interpréter et visualiser les relations entre les observations et les variables catégorielles.
Les coordonnées permettent de positionner chaque individu et chaque modalité dans un espace à plusieurs dimensions(axes). Ils montrent la position relative des individus et des modalités par rapport aux axes principaux. On peut identifier quelles variables catégorielles contribuent le plus à la position des individus dans l'espace factoriel, et vice versa.

Pour les coordonnées des individus, on multiplie la matrice des profils lignes par le vecteur propre associé à l'axe k .
Pour les coordonnées des modalités, on multiplie la matrice des profils colonnes par le vecteur propre associé à l'axe k.

Calcul du cos^2

Le cos^2 est une mesure utilisée en Analyse des Correspondances Multiples (ACM) pour évaluer la qualité de la représentation des modalités (variables catégorielles) sur les axes factoriels. Le cos^2 est également connu sous le nom de "contribution" ou "qualité de représentation".

Lorsque vous effectuez une ACM, chaque modalité est représentée par des coordonnées dans l'espace factoriel. Le cos^2 mesure la proportion de l'inertie de la modalité qui est expliquée par l'axe considéré.
Plus précisément :

Les cos^2 sont donc des valeurs comprises entre 0 et 1. Une valeur élevée de cos^2 indique que la modalité est bien représentée sur l'axe considéré, tandis qu'une valeur proche de 0 indique que la modalité n'est pas bien représentée par cet axe.

Le cos^2 d’un individu i sur un axe k est calcule comme suit :

$$\frac{(coordonnee\ de\ l'individue\ i\ sur\ l'Axe\ K)^2}{\sum(coordonne\ de\ l'individue\ i\ sur\ l'Axe\ j)^2}$$

Avec la somme concerne tous les axes .

Le cos^2 d’une modalité est calcule comme suit :

$$\frac{(coordonnee\ de\ la\ modalite\ i\ sur\ l'Axe\ K)^2}{\sum(coordonnee\ de\ la\ modalite\ i\ sur\ l'Axe\ j)^{2}}$$

Calcul de la contribution

La contribution mesure la part d'inertie totale expliquée par une modalité (ou individu) sur un axe particulier. C'est une mesure relative qui permet de quantifier l'importance de chaque modalité (ou individu) dans la structuration de l'information sur cet axe.
Elle est souvent exprimée en pourcentage.

Elle est calculée à partir du cos^2 par la formule suivante:

$$\text{Contribution} = \frac{\cos^2 * 100}{Somme\ des\ \cos^2\ sur\ l'axe}$$

03

Application de la partie théorique + Interpretation des résultats

Pour cette analyse, on prend une partie de notre data pour faciliter les calculs:

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94

On transforme les variables quantitatives en variables catégoriales:
On les groupent de la façon suivant:
Pour l’age, on a les modalités suivantes: [19-29] et [30-40]
Pour Annual Income, on a les modalités suivantes: [15-16] et [17-18]
Pour Spending Score, on a les modalités suivantes: [6-50] et [51-100]

On construit le TDC :

TABLEAU DISJONCTIF COMPLET									
ID	MALE	FEMALE	19-29	29-40	15-16	17-18	6-050	50-100	Totale
1	1	0	1	0	1	0	1	0	4
2	1	0	1	0	1	0	0	1	4
3	0	1	1	0	1	0	1	0	4
4	0	1	1	0	1	0	0	1	4
5	0	1	0	1	0	1	1	0	4
6	0	1	1	0	0	1	0	1	4
7	0	1	0	1	0	1	1	0	4
8	0	1	1	0	0	1	0	1	4
Total	2	6	6	2	4	4	4	4	32

On calcul les profils lignes et les profils colonnes, ainsi que les profils moyens

PROFILS LIGNES									
ID	MALE	FEMALE	19-29	29-40	15-16	17-18	6-050	50-100	Total
1	0,25	0	0,25	0	0,25	0	0,25	0	1
2	0,25	0	0,25	0	0,25	0	0	0,25	1
3	0	0,25	0,25	0	0,25	0	0,25	0	1
4	0	0,25	0,25	0	0,25	0	0	0,25	1
5	0	0,25	0	0,25	0	0,25	0,25	0	1
6	0	0,25	0,25	0	0	0,25	0	0,25	1
7	0	0,25	0	0,25	0	0,25	0,25	0	1
8	0	0,25	0,25	0	0	0,25	0	0,25	1
Total	0,5	1,5	1,5	0,5	1	1	1	1	8
Profil moyen	0.0625	0.1875	0.1875	0.0625	0.125	0.125	0.125	0.125	

PROFILS COLONNES									
ID	MALE	FEMALE	19-29	29-40	15-16	17-18	6-050	50-100	Total
1	0,5	0	0,166666667	0	0,25	0	0,25	0	1,16666667
2	0,5	0	0,166666667	0	0,25	0	0	0,25	1,16666667
3	0	0,166666667	0,166666667	0	0,25	0	0,25	0	0,83333333
4	0	0,166666667	0,166666667	0	0,25	0	0	0,25	0,83333333
5	0	0,166666667	0	0,5	0	0,25	0,25	0	1,16666667
6	0	0,166666667	0,166666667	0	0	0,25	0	0,25	0,83333333
7	0	0,166666667	0	0,5	0	0,25	0,25	0	1,16666667
8	0	0,166666667	0,166666667	0	0	0,25	0	0,25	0,83333333
Total	1	1	1	1	1	1	1	1	8
Profil moyen	0,125	0,125	0,125	0,125	0,125	0,125	0,125	0,125	

La distance entre l'individu 1 et les autres individus:

$D^2(1,2) = \frac{1}{0.125}(0.25-0)^2 + \frac{1}{0.125}(0-0.25)^2 = 1$

$D^2(1,3) = \frac{1}{0.1875}(0-0.25)^2 + \frac{1}{0.0625}(0.25-0)^2 = 1{,}33$

$D^2(1,4) = \frac{1}{0.0625}(0.25-0)^2 + \frac{1}{0.1875}(0-0.25)^2 + \frac{1}{0.125}(0-0.25)^2 + \frac{1}{0.125}(0.25-0)^2 = 3.833$

$D^2(1,5) = 3.666$

$D^2(1,6)=2$

$D^2(1,7)=1$

$D^2(1,8)=2$

Interpretation:

Tant que la distance augmente, la resemblance entre les individus diminue, donc d’après les calculs on voit que l’individu 2 ressemble le plus à l’individu 1, par contre l’individu 4 est l’individu qui a plus de difference avec l’individu 1

La distance des individus à l'origine:

$D^2(1,O) = \frac{1}{0.0625}\left((0.25 - 0{,}0625)^2 + (0 - 0{,}0625)^2\right) + \frac{1}{0.1875}\left((0 - 0.1875)^2 + (0.25 - 0{,}1875)^2\right) + \frac{1}{0.125}(2 * (0.25 - 0{,}125)^2 + 2 * (0 - 0{,}125)^2) = 1{,}33$

$D^2(2,O) = 1{,}33$
$D^2(3,O) = 0{,}66$
$D^2(4,O) = 0{,}66$
$D^2(5,O)=1{,}33$
$D^2(6,O)=0{,}77$
$D^2(7,O)=1{,}33$
$D^2(8,O)=1{,}54$

Interpretation:

On voit que l’individu 8 est l’individu qui possède des modalités rare car sa distance à l’origine est la grande.

La distance entre la modalité homme et les autres modalités:

$D^2(M,19\text{-}29)=2{,}887$

$D^2(M,29\text{-}40)=8$

$D^2(M,15\text{-}16)=2$

$D^2(M,17\text{-}18)=6$

$D^2(M,6\text{-}50)=2{,}5$

$D^2(M,50\text{-}100)=2{,}5$

Interpretation:

On voit que les individus qui partagent les modalités(Male et Age ”29-40”) sont plus nombreux que (Male et Age”19-29”)

La distance entre la modalité femme et les autres modalités:

$D^2(F,19\text{-}29)=0.832$

$D^2(F,29\text{-}40)=2.668$

$D^2(F,15\text{-}16)=1,996$

$D^2(F,17\text{-}18)=0,665$

$D^2(F,6\text{-}50)=1,331$

$D^2(F,50\text{-}100)=1,33$

Interpretation:

On voit que les individus qui partagent les modalités(Female et Annual Income”17-18”) sont plus nombreux que (Female et Annual Income”15-16”)

La distance des modalités à l’origine:

$D^2(M,O)=3$

$D^2(F,O)=0,33$

$D^2(19\text{-}29,O)=0.33$

$D^2(29\text{-}40,O)=3$

$D^2(15\text{-}16,O)=1$

$D^2(17\text{-}18,O)=1$

$D^2(6\text{-}50,O)=1$

$D^2(50\text{-}100,O)=1$

Interpretation:

D’après les calculs, on voit que les modalités male et age[29-40] sont rarement possédé par les individus.

Le Tableau de Burt:

		Mtrice					
2	0	2	0	2	0	1	1
0	6	4	2	2	4	3	3
2	4	6	0	4	2	2	4
0	2	0	2	0	2	2	0
2	2	4	0	4	0	2	2
0	4	2	2	0	4	2	2
1	3	2	2	2	2	4	0
1	3	4	0	2	2	0	4

Les valeurs propres:

λ_1=18.45
λ_2=7.34
λ_3=5.12
λ_4=1.65
λ_5=1.005
λ_6=0.33
λ_7=0
λ_8=0

Les vecteurs propres:

$\overrightarrow{v_1}$=(-0.144, -0.511, -0.515, -0.140, -0328, -0.327, -0.301, -0.354)
$\overrightarrow{v_2}$ =(-0.298, 0.331, -0.369, -0.401, -0.422, 0.454, 0.257, -0.225)
$\overrightarrow{v_3}$=(0.273, -0.151, 0.078, 0.201, 0.358, -0.235, 0.637, -0,514)
$\overrightarrow{v_4}$=(0.663, -0.465, 0.057, 0.255, -0.221, 0.419, 0.024, 0.222)
$\overrightarrow{v_5}$=(0, -0.119, 0.461, -0.581, -0.401, 0.281, 0.246, -0.366)
$\overrightarrow{v_6}$=(0.359, 0.359, 0.341, -0.282, -0.282, -0.417, -0.417)
$\overrightarrow{v_7}$=(-0.161, -0.161, 0.605, 0.605, -0.158, -0.158, -0.286, -0.286)
$\overrightarrow{v_8}$=(0.592, 0.592, -0.334, -0.334, -0.085, -0.058, -0.172, -0.172)

Choix des axes:

Alors premièrement, on calcule le pourcentage cumulé d’inertie expliquée (une mesure utilisée dans l’ACM pour évaluer la proportion totale de variabilité(la différenciation des relations entre les catégories des variables catégorielles) des données expliquée par les premiers axes factoriels. L'inertie représente la quantité d'information contenue dans les données, et le pourcentage cumulé d'inertie expliquée donne la proportion cumulée de cette inertie jusqu'à un certain nombre d'axes.).

On a:

Les vps	Inertie expliquée	Pourcentage %	% cumulé
18,45	0,544	54,4	54,4
7,34	0,216	21,6	76
5,12	0,151	15,1	91,1
1,65	0,048	4,8	95,9
1,005	0,029	2,9	98,8
0,33	0,00973	0,973	99,77
0	0		
0	0		

Les coordonnées + cos^2 + contribution:

Axe 1 — Axe 2

Pour les Individus

Axe 1

les Coordonnées	Puissance^2	cos^2	contribution
-0.322	0.103684	0.705582927	0.106303695
-0.335	0.112225	0.510556893	0.076920914
-0.413	0.170569	0.985554894	0.148484498
-0.427	0.182329	0.861790424	0.129838042
-0.319	0.101761	0.798996553	0.120377467
-0.426	0.181476	0.987973977	0.148848959
-0.319	0.101761	0.798996553	0.120377467
-0.426	0.181476	0.987973977	0.148848959

Axe 2

les Coordonnées	Puissance^2	cos^2	contribution
-0.208	0.043264	0.294417073	0.216074221
-0.328	0.107584	0.489443107	0.359204842
-0.05	0.0025	0.014445106	0.010601338
-0.171	0.029241	0.138209576	0.101432727
0.16	0.0256	0.201003447	0.147517475
0.047	0.002209	0.012026023	0.008825961
0.16	0.0256	0.201003447	0.147517475
0.047	0.002209	0.012026023	0.008825961

Pour les Modalités

Axe 1

les Coordonnées	Puissance^2	cos^2	contribution
-1.084	1.175056	0.646253011	0.094644885
-1.097	1.203409	0.585504185	0.085748113
-1.86	3.4596	0.996883371	0.145995314
-1.873	3.508129	0.985898904	0.14438662
-1.075	1.155625	0.806831106	0.118161828
-1.873	3.508129	0.999992874	0.146450707
-1.075	1.155625	0.806831106	0.118161828
-1.873	3.508129	0.999992874	0.146450707

Axe 2

les Coordonnées	Puissance^2	cos^2	contribution
-0.802	0.643204	0.353746989	0.301880179
-0.923	0.851929	0.414495815	0.35372194
-0.104	0.010816	0.003116629	0.002659665
-0.224	0.050176	0.014101096	0.012033576
-0.526	0.276676	0.193168894	0.164846238
-0.005	0.000025	7.12626E-06	6.08139E-06
-0.526	0.276676	0.193168894	0.164846238
-0.005	0.000025	7.12626E-06	6.08139E-06

Interpretation:

Pour l’axe 1:
Selon le cos^2 et la contribution, on voit que les individus 3,6 et 8 ont une cos^2 de 0,98, cela implique qu’il ya une excellente représentation sur cet axe et qu’ils ont grande importance des autres individus.

Pour les modalités, on voit que Annual Income[17k-18k] et Spending Score[51-100] ont une excelente representation sur cet axe, ainsi qu’elles sont importantes des autres modalités sur la structure des données de cet axe.

04 Application avec Python

05 Conclusion

En résumé, l'Analyse des Correspondances Multiples (ACM) est un outil analytique essentiel pour clarifier les liens complexes entre des variables catégorielles. Les résultats obtenus, que ce soit à travers les calculs théoriques ou l'application pratique avec Python, ont fourni des éclairages significatifs sur la distribution des modalités dans un espace factoriel.
Les axes factoriels, les cos^2 et les contributions ont été des indicateurs clés pour comprendre la structure des données.
Cependant, ce sont les graphiques générés par Python qui ont véritablement enrichi notre compréhension. Les visualisations ont permis d'explorer visuellement les relations entre les modalités, offrant une clarté supplémentaire.