TP5 : Analyse des Correspondances Multiples

A. L. N'Guessan, V. Roca & W. Heyse

Le 29 Septembre 2025

Contents

1 Rappels

1.1 Introduction à l'Analyse des Correspondances Multiples

L'**Analyse des Correspondances Multiples** (ACM) est une extension de l'Analyse Factorielle des Correspondances (AFC) appliquée à des tableaux de données où plusieurs variables qualitatives sont croisées. Contrairement à l'AFC, qui est utilisée pour analyser la relation entre deux variables qualitatives, l'ACM permet d'analyser simultanément plusieurs variables qualitatives, en représentant les relations entre leurs modalités dans un espace à faible dimension. L'objectif de l'ACM est de réduire la dimensionnalité des données tout en préservant l'information la plus pertinente possible.

1.2 Fonctionnement Mathématique de l'ACM

1.2.1 Tableau Disjonctif Complet : Construction et Propriétés

L'ACM commence par la construction du **tableau disjonctif complet**. Ce tableau est une matrice binaire qui associe à chaque individu et à chaque modalité de chaque variable une valeur de 0 ou 1.

Construction du tableau disjonctif complet : - Soit un ensemble de N individus décrits par K variables qualitatives, chaque variable X_k ayant J_k modalités. - Le tableau disjonctif complet $\mathbf{Z}$ est une matrice $N \times J$ où $J = \sum_{k=1}^{K} J_k$. Chaque élément z_{ij} de la matrice vaut 1 si l'individu i possède la modalité j, et 0 sinon.

Propriétés du tableau disjonctif complet : - Chaque ligne de la matrice $\mathbf{Z}$ est un profil individu, indiquant les modalités possédées par cet individu. - Chaque colonne correspond à une modalité d'une

Table 1: Tableau de données initial

Variable_1	Variable_2	Variable_3
France	Oui	Rouge
Allemagne	Oui	Vert
France	Non	Bleu
Italie	Oui	Vert
France	Non	Bleu

Table 2: Tableau disjonctif complet

Variable_1_France	Variable_1_Allemagne	Variable_1_Italie	Variable_2_Oui	Variable_2_Non	Variable_3_B
1	0	0	1	0	
0	1	1	1	0	
1	0	0	0	1	
0	0	0	1	0	
1	0	0	0	1	

variable, et la somme des éléments d'une colonne donne l'effectif total de cette modalité.

On note $\mathbf{D}_c$ la matrice contenant sur sa diagonale les effectifs marginaux de chacune des modalités (c'est à dire la somme des lignes du tableau disjonctif complet), avec l'exemple précédant, on aurait :

$$\mathbf{D}_c = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

On définit de manière analogue la matrice $\mathbf{D}_r$ la matrice contenant sur sa diagonale les effectifs marginaux de chacun des individus.

1.2.2 Lien avec l'ACP

L'ACM peut être vue comme une application de l'ACP sur les profils lignes du tableau disjonctif complet après avoir été centré par les profils moyens et avec la métrique du χ^2. On cherche donc les vecteurs/valeurs propres de la matrice :

$$\underbrace{\left(\frac{1}{n}\mathbf{X}_r^\top D_r^{-1}\mathbf{X}_r - g_r g_r^\top\right)}_{\text{TDC centré}} \underbrace{nD_c^{-1}}_{\text{métrique}}$$

avec $\mathbf{X}_r = \mathbf{D}_r^{-1}Z$

Le tableau disjonctif est d'abord centré par le profil moyen des individus, c'est-à-dire que chaque colonne est soustraite par sa moyenne.La construction des axes factoriels suit des étapes similaires à l'ACP.

1.2.3 Correction de Benzécri

Pour éviter la sous-estimation de la variance expliquée par les premiers axes, des corrections spécifiques sont souvent appliquées en ACM. La correction de Benzécri est utilisée pour ajuster la proportion de l'inertie expliquée par les axes factoriels. Elle consiste à recalculer les valeurs propres obtenues pour mieux représenter la variance réelle dans les données. On ne sélectionne que les valeurs propres supérieures à $\frac{1}{p}$ puis, la version corrigée de chaque valeur propre λ_j est donnée par :

$$\lambda_j' = \left(\left(\frac{p}{p-1}\right)\left(\lambda_j - \frac{1}{p}\right)\right)^2$$

1.3 Interprétation des Résultats de l'ACM

On interprétera les résultats de la même manière qu'on interprète les résultats des individus pour une ACP. On se s'intérèssera principalement aux contribution des modalités aux axes et à la projection de toutes les modalités dans le même espace factoriel. Dans cet espace factoriel, les distances entre les modalités reflèterons leur proximité au sens de la distance du χ^2 : si deux modalités sont proches dans l'espace factoriel, elles auront des répartitions proches et seront partagées par les mêmes individus. Pour analyser les contributions des modalités, usuellement, on retiens celles qui ont une contribution sur l'axe k supérieure à $\sqrt{\lambda_k}$.

Pour l'analyse du lien entre les variables initiales et les axes de l'ACM, on pourra aussi utiliser le rapport de corrélation η^2. Ce coefficient mesure le rapport entre la variance a l'intérieur d'une modalité et la variance entre les modalités. Plus il est proche de 1, plus les modalités sont bien séparées selon l'axe et plus il proche de 0, moins les modalités sont séparées.

2 Exercice

On propose d'analyser les résultat d'un questionnaire distribué à de nombreux Data Scientist tout autour du monde.

1. Chargez les données et proposez-en une analyse descrptive. Commentez.
2. A l'aide de la fonction `MCA` du package `FactoMineR`, réalisez l'ACM du jeu de données fournies.
3. Récupérez les valeurs propres et représentez les.
4. Appliquez la correction de Benzecri et représentez les valeurs propres corrigées du jeu de données. Combien d'axes faut-il retenir ?
5. Tracez le graphe des modalités. Colorez les modalités de chaque variable d'une couleur différente.
6. Analysez et commentez les positions relatives des différentes modalités.
7. Pour les composantes retenues et pour les variables d'intérêt, analysez les rapports de corrélation. *Indication :* Vous pouvez vous aider des arguments `habillage` et `addEllipses` de la fonction `fviz_mca_ind` ou de la fonction `plotellipses`.
8. A l'aide des éléments précédemment découverts, interprétez les axes retenus.
9. Ajoutez une ligne au jeu de données avec vos caractéristiques (ou celles que pensez avoir d'ici 3-4 ans) et représentez votre donnée comme individus supplémentaire sur le graphe des modalités et des individus. Par rapport à votre proximité aux différentes modalités, correspondez-vous à l'échantillon représenté ou êtes-vous un outlier ?