TP4 : Analyse Factorielle des Correspondances

A. L. N'Guessan & W. Heyse

2024-2025

Contents

1 Rappels de cours

L'**Analyse Factorielle des Correspondances** (AFC) est une méthode statistique utilisée pour analyser le lien (appelé correspondance) entre **deux variables qualitatives** qui passe par l'analyse de leur tableau de contingence, c'est-à-dire la table qui croise les modalités des deux variables qualitatives. L'AFC est particulièrement utile pour explorer les relations entre les modalités des variables et pour visualiser ces relations dans un espace à faible dimensions, un plan factoriel.

1.1 Fonctionnement Mathématique de l'AFC

1.1.1 Tableau de Contingence

Notons X_1 la première variable qualitative comportant p modalités et X_2 la seconde variable qualitative comportant q modalités. Le point de départ de l'AFC est un **tableau de contingence N** qui répertorie les effectifs d'apparition de chaque combinaison de modalités des deux variables étudiées.

$$\mathbf{N} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1q} \\ n_{21} & n_{22} & \cdots & n_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ n_{p1} & n_{p2} & \cdots & n_{pq} \end{pmatrix} \begin{matrix} n_{1\cdot} \\ n_{2\cdot} \\ \\ n_{p\cdot} \end{matrix} \qquad \begin{matrix} n_{\cdot 1} & n_{\cdot 2} & & n_{\cdot q} \end{matrix}$$

1.1.1.1 Matrice des Effectifs Marginaux À partir du tableau de contingence **N**, on calcule les effectifs marginaux pour les lignes et les colonnes : - **Effectif marginal des lignes** : $n_{i\cdot} = \sum_{j=1}^{q} n_{ij}$ - **Effectif marginal des colonnes** : $n_{\cdot j} = \sum_{i=1}^{p} n_{ij}$

Le total des observations est donné par $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$.

On définit alors les matrices diagonales des effectifs marginaux lignes et colonnes respectivement $\mathbf{D}_r$ et $\mathbf{D}_c$:

$$\mathbf{D}_r = \begin{pmatrix} n_{1\cdot} & & 0 \\ & \ddots & \\ 0 & & n_{p\cdot} \end{pmatrix}, \qquad \mathbf{D}_c = \begin{pmatrix} n_{\cdot 1} & & 0 \\ & \ddots & \\ 0 & & n_{\cdot q} \end{pmatrix}$$

De plus on peut définir les centres de gravité des profils lignes $\mathbf{g}_r$ (resp. colonnes $\mathbf{g}_c$) :

$$\mathbf{g}_r = \begin{pmatrix} \frac{n_{\cdot 1}}{n} \\ \vdots \\ \frac{n_{\cdot q}}{n} \end{pmatrix} \quad \text{et} \quad \mathbf{g}_c = \begin{pmatrix} \frac{n_{1\cdot}}{n} \\ \vdots \\ \frac{n_{p\cdot}}{n} \end{pmatrix}$$

1.1.2 Profils Lignes et Colonnes

Les **profils lignes** et **profils colonnes** sont les distributions de fréquences conditionnelles respectivement par ligne et par colonne.

- **Profil ligne** $\mathbf{X}_r = \mathbf{D}_r^{-1} N$ pour la i-ième ligne : $\mathbf{X}_{ri} = \left(\frac{n_{i1}}{n_{i\cdot}}, \frac{n_{i2}}{n_{i\cdot}}, \ldots, \frac{n_{iJ}}{n_{i\cdot}} \right)$
- **Profil colonne** $\mathbf{X}_c = \mathbf{D}_c^{-1} N^\top$ pour la j-ième colonne : $\mathbf{X}_{cj} = \left(\frac{n_{1j}}{n_{\cdot j}}, \frac{n_{2j}}{n_{\cdot j}}, \ldots, \frac{n_{Ij}}{n_{\cdot j}} \right)$

Le profil ligne permets de comparer la distribution des modalités de la variable X_1 par rapport aux modalités de la variable X_2 (et inversement pour les profils colonnes).

1.1.3 Distance du Chi-2

L'AFC utilise la **distance du** χ^2 pour mesurer l'écart entre les profils observés. Cette distance permet de mettre plus de poids sur les modalités de petits effectif, si on observe de grands écarts sur des modalités peu représentées, ceux-ci ont plus de poids dans le calcul de la distance. Et inversement, on donne moins de poids à des écarts importants qui pourraient être dus au fait que l'on a seulement observé plus de points sur cette modalité.

La distance du Chi-2 entre deux profils ligne $\mathbf{X}_i$ et $\mathbf{X}_{i'}$ est donnée par :

$$d_{\chi^2}(\mathbf{X}_i, \mathbf{X}_{i'}) = \sum_{j=1}^{q} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}} \right)^2$$

La matrice associée a cette métrique pour les profils ligne est $M_r = nD_c^{-1}$.

On défint de la même manière la distance du χ^2 pour les profils colonnes :

$$d_{\chi^2}(\mathbf{X}_j, \mathbf{X}_{j'}) = \sum_{i=1}^{p} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}} \right)^2$$

La matrice associée a cette métrique pour les profils ligne est $M_c = nD_r^{-1}$.

1.1.4 Lien avec l'ACP

On cherche donc a analyser les strucutres entre les différents profils lignes (resp. colonnes) au regard de la distance du χ^2. En ACP, on cherchait a analyser les strucutres entre les différents individus au regard de la distance euclidienne.

L'AFC peut donc être vue comme une application de l'ACP sur les profils lignes (resp. colonnes) pondérés par les masses marginales. Le tableau de contingence normalisé (ou tableau des résidus) est décomposé en valeurs propres et vecteurs propres pour obtenir les axes factoriels.

Appliquons donc la même méthode, on cherches les vecteurs propres et valeurs propres de la matrice :

$$\underbrace{(\mathbf{X}_r - \mathbb{1}_p \mathbf{g}_r)^\top}_{\text{matrice de données centrée}} \underbrace{\frac{D_r}{n}}_{\text{fréquences}} (\mathbf{X}_r - \mathbb{1}_p \mathbf{g}_r) \underbrace{nD_c^{-1}}_{\text{métrique}}$$

Cela revient à faire une ACP de la matrice précédante. En réalité, après simplification, il suffit de s'intérésser à la matrice $N^\top D_r^{-1} N D_c^{-1}$

Dans le cas des profils colonnes, cela revient à regarder les vecteurs propres de la matrice $N D_c^{-1} N^\top D_r^{-1}$.

1.1.5 Lien entre les profils lignes et colonnes

L'ACP des **profils lignes** (appelée analyse directe) permets de projeter les modalités de X_1 sur des facteurs propres. L'ACP des **profils colonnes** (appelée analyse duale) permets de projeter celles de X_2. Dans les deux cas, les valeurs propres sont les mêmes et les vecteurs propres sont tels que si u est un vecteur propre pour $N^\top D_r^{-1} N D_c^{-1}$ alors $\frac{1}{\sqrt{\lambda}} N^\top D_r^{-1} u$ est un vecteur propre de $N D_c^{-1} N^\top D_r^{-1}$.

Cette relation permet de projeter les modalités des deux variables dans le même espace pour avoir une représentation des toutes les modalités dans un espace comparable. De plus elle montre qu'il est équivalent d'étudier les profils lignes ou les profils colonnes.

1.2 En pratique

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.

2 Exercice

On propose d'analyser la répartion des votes par liste électorale et par région aux dernières élections européènenes. Celles-ci se trouvent dans le fichier `data_elections.csv`.

Nom Abrégé	Nom de liste
HUMANITE SOUVERAINE	POUR UNE HUMANITE SOUVERAINE
POUR UNE DEMOCRATIE REELLE : DECIDONS NOUS-MEMES !	POUR UNE DEMOCRATIE REELLE : DECI
LA FRANCE FIERE	LA FRANCE FIERE, MENEE PAR MARION
LFI - UP	LA FRANCE INSOUMISE - UNION POPULA
La FRANCE REVIENT	LA FRANCE REVIENT ! AVEC JORDAN BA
EUROPE ÉCOLOGIE	EUROPE ÉCOLOGIE
FREE PALESTINE	FREE PALESTINE
PARTI ANIMALISTE	PARTI ANIMALISTE - LES ANIMAUX COM
PARTI REVOLUTIONNAIRE COMMUNISTES	PARTI REVOLUTIONNAIRE COMMUNISTE
PARTI PIRATE	PARTI PIRATE
BESOIN D'EUROPE	BESOIN D'EUROPE
PACE	PACE - PARTI DES CITOYENS EUROPÉEN
ÉQUINOXE	ÉQUINOXE : ÉCOLOGIE PRATIQUE ET RE
ECOLOGIE POSITIVE	ECOLOGIE POSITIVE ET TERRITOIRES
LISTE ASSELINEAU-FREXIT	LISTE ASSELINEAU-FREXIT, POUR LE PO
PAIX ET DECROISSANCE	PAIX ET DECROISSANCE
POUR UNE AUTRE EUROPE	POUR UNE AUTRE EUROPE
LA DROITE	LA DROITE POUR FAIRE ENTENDRE LA V
LUTTE OUVRIERE	LUTTE OUVRIERE - LE CAMP DES TRAVA
CHANGER L'EUROPE	CHANGER L'EUROPE
NLP	NOUS LE PEUPLE
URGENCE REVOLUTION !	POUR UN MONDE SANS FRONTIERES NI P
PPL	"POUR LE PAIN, LA PAIX, LA LIBERTÉ !" P
L'EUROPE CA SUFFIT !	L'EUROPE CA SUFFIT !
PRENONS-NOUS EN MAIN	NON ! PRENONS-NOUS EN MAINS
FORTERESSE EUROPE	FORTERESSE EUROPE - LISTE D'UNITE N
REVEIL EUR	RÉVEILLER L'EUROPE
NON À L'UE ET À L'OTAN	NON À L'UE ET À L'OTAN, COMMUNISTES
AR	ALLIANCE RURALE
FRANCE LIBRE	FRANCE LIBRE
EUROPE TERRITOIRES ÉCOLOGIE	EUROPE TERRITOIRES ÉCOLOGIE
LA RUCHE CITOYENNE	LA RUCHE CITOYENNE
GAUCHE UNIE	GAUCHE UNIE POUR LE MONDE DU TRAV
DEFENDRE LES ENFANTS	DEFENDRE LES ENFANTS
EAC	ÉCOLOGIE AU CENTRE
DEMOCRATIE REPRESENTATIVE	DEMOCRATIE REPRESENTATIVE
ESPERANTO	ESPERANTO LANGUE COMMUNE
LIBERTÉ DÉMOCRATIQUE FRANÇAISE	LIBERTÉ DÉMOCRATIQUE FRANÇAISE

1. Chargez les données et donnez une brève analyse descriptive de celles-ci.
2. Contruisez une table de contingence avec les données.
3. Rappelez l'utilité et le fonctionnement d'un test du χ^2. Réalisez le à l'aide de la fonction `chisq.test`. Que pouvez-vous en déduire ?
4. A l'aide de la fonction `CA` disponible dans le package `FactoMineR`, réalisez l'AFC des données.
5. Déterminez le nombre d'axes à interpréter.
6. Analysez les contribution des régions.
7. Faites de même pour les listes.
8. Représentez les modalités des deux variables dans les plans factoriels retenus.

9. Interprétez les axes factoriels retenus.