Analyse factorielle des correspondances

L'analyse factorielle des correspondances (AFC) est une méthode d'analyse exploratoire qui vise à représenter graphiquement les relations entre les modalités de deux variables qualitatives. Elle permet de représenter simultanément les **profils-lignes** (dans $\mathbb{R}^p$) et les **profils-colonnes** (dans $\mathbb{R}^n$) d'un tableau de contingences, dans un espace de faible dimension, tout en préservant au mieux la distance du χ^2. L'objectif de l'AFC est de trouver une représentation bidimensionnelle (voir tridimensionnelle) dans laquelle les proximités géométriques entre points reflètent au mieux les similarités entre les modalités.

> Remarque
>
> L'AFC peut être vue comme une double ACP : une ACP pondérée appliquée aux profils-lignes et aux profils-colonnes, dans leur espaces respectifs avec une métrique adaptée.

1 Notation

On considère un tableau de contingence $K = (k_{ij})$, où k_{ij} est le nombre d'individus appartenant à la classe $i \in \{1, \dots, n\}$ et à la catégorie $j \in \{1, \dots, p\}$. On travaille ensuite avec le tableau des fréquences relatives en normalisant ce tableau. Comme les fréquences sont proportionnelles à la taille d'échantillon n, le tableau des fréquences relatives contient plus d'information. Notons $F = (f_{ij})$, dans lequel

$$f_{ij} = \frac{k_{ij}}{k_{\bullet\bullet}} = \frac{k_{ij}}{\sum_{l=1}^{n}\sum_{m=1}^{p} k_{lm}}.$$

Les marges lignes (resp. colonnes) du tableau correspondent à la somme des colonnes pour chaque ligne (resp. à la somme des lignes pour chaque colonne) :

$$f_{i\bullet} = \sum_{j=1}^{p} f_{ij} = \frac{k_{i\bullet}}{k_{\bullet\bullet}}, \quad 1 \leq i \leq n; \tag{1}$$

$$f_{\bullet j} = \sum_{i=1}^{n} f_{ij} = \frac{k_{\bullet j}}{k_{\bullet\bullet}}, \quad 1 \leq j \leq p. \tag{2}$$

On a $f_{\bullet\bullet} = \sum_{i=1}^{n} f_{i\bullet} = \sum_{j=1}^{p} f_{\bullet j} = 1$.

Exemple

Comme exemple, on va considérer la majeure et le type d'admission des étudiants inscrit au cours STT-2200 à l'automne 2025.

TABLE 1 : Tableau de contingence des étudiants inscrit pour le cours STT-2200 (Automne 2025) croisant leur majeure et leur type d'admission.

	Collège	Université Laval	Autre université	Hors Québec
Actuariat	2	0	0	1
Statistique	2	4	1	0
Bio-info	4	2	0	2
Finance	2	0	0	0
Maths	1	0	0	0
Info	2	1	0	1

Ici, on trouve $k_{\bullet\bullet} = 25$. C'est tout simplement le nombre d'étudiants inscrit au cours. On trouve donc le tableau de fréquences suivant :

TABLE 2 : Tableau de fréquences associé au tableau de contingence précédent.

	Collège	Université Laval	Autre université	Hors Québec	$f_{i\bullet}$
Actuariat	0.08	0	0	0.04	0.12
Statistique	0.08	0.16	0.04	0	0.28
Bio-info	0.16	0.08	0	0.08	0.32
Finance	0.08	0	0	0	0.08
Maths	0.04	0	0	0	0.04
Info	0.08	0.04	0	0.04	0.16
$f_{\bullet j}$	0.52	0.28	0.04	0.16	1

2 Indépendance statistique

Le tableau des fréquences relatives $F = (f_{ij})$ peut être interprété comme une estimation des probabilités conjointes des modalités des deux variables qualitatives. Si les deux variables sont statistiquement indépendantes, on s'attend à ce que la probabilité conjointe s'approche du produit des probabilités marginales :

$$f_{ij} \approx f_{i\bullet} f_{\bullet j}, \quad i \in \{1, \dots, n\},\ j \in \{1, \dots, p\}.$$

Pour tester si les écarts observés entre f_{ij} et $f_{i\bullet}f_{\bullet j}$ sont significatifs, on utilise le test du χ^2 d'indépendance :

$$T = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(k_{ij} - \mathbb{E}(k_{ij})\right)^2}{\mathbb{E}(k_{ij})} = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(k_{ij} - \frac{k_{i\bullet}k_{\bullet j}}{k_{\bullet\bullet}}\right)^2}{\left(\frac{k_{i\bullet}k_{\bullet j}}{k_{\bullet\bullet}}\right)}.$$

Sous l'hypothèse d'indépendance, cette statistique suit approximativement une loi du χ^2. Si les variables sont indépendantes, la statistique T doit être proche de 0.

3 Profils-lignes et profils-colonnes

Pour analyser les structures dans le tableau de contingence, on introduit la notion de profil. Chaque ligne du tableau peut être vue comme un profil-ligne

$$L_i = \left(\frac{k_{i1}}{k_{i\bullet}}, \dots, \frac{k_{ip}}{k_{i\bullet}}\right) = \left(\frac{f_{i1}}{f_{i\bullet}}, \dots, \frac{f_{ip}}{f_{i\bullet}}\right).$$

Le profil-ligne représente la répartition des modalités i de la première variable parmi les modalités de la seconde.

De même, chaque colonne du tableau peut être vue comme un profil-colonne

$$C_j = \left(\frac{k_{1j}}{k_{\bullet j}}, \dots, \frac{k_{nj}}{k_{\bullet j}}\right) = \left(\frac{f_{1j}}{f_{\bullet j}}, \dots, \frac{f_{nj}}{f_{\bullet j}}\right).$$

Le profil-colonne représente la répartition des modalités j de la deuxième variable parmi les modalités de la première.

On peut ensuite s'intéresser au profil-ligne moyen (resp. profil-colonne moyen) obtenus comme la moyenne pondérée des profils-lignes (resp. profils-colonnes). Autrement dit, ils correspond aux fréquences marginales colonnes (resp. fréquences marginales lignes). Le profil-ligne moyen est donné par

$$\left(\sum_{i=1}^{n} f_{i\bullet}\frac{f_{i1}}{f_{i\bullet}}, \dots, \sum_{i=1}^{n} f_{i\bullet}\frac{f_{ip}}{f_{i\bullet}}\right) = \left(f_{\bullet 1}, \dots, f_{\bullet p}\right),$$

et le profil-colonne moyen est donné par

$$\left(\sum_{j=1}^{p} f_{\bullet j}\frac{f_{1j}}{f_{\bullet j}}, \dots, \sum_{j=1}^{p} f_{\bullet j}\frac{f_{nj}}{f_{\bullet j}}\right) = \left(f_{1\bullet}, \dots, f_{n\bullet}\right).$$

Si les variables sont indépendantes, tous les profiles sont égaux à leur profils moyens respectifs. Autrement dit, pour tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, p\}$,

$$\left(\frac{f_{i1}}{f_{i\bullet}}, \dots, \frac{f_{ip}}{f_{i\bullet}}\right) = \left(f_{\bullet 1}, \dots, f_{\bullet p}\right) \quad \text{et} \quad \left(\frac{f_{1j}}{f_{\bullet j}}, \dots, \frac{f_{nj}}{f_{\bullet j}}\right) = \left(f_{1\bullet}, \dots, f_{n\bullet}\right).$$

Ainsi, plus les profils s'éloignent de leurs moyennes, plus les variables montrent une dépendance.

Pour mesurer la différence entre deux profils-lignes, on utilise la distance du χ^2 pondérée par les fréquences marginales :

$$d^2(L_i, L_{i'}) = \sum_{j=1}^{p} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2 .$$

On peut faire de même pour la différence entre deux profils-colonnes :

$$d^2(C_j, C_{j'}) = \sum_{i=1}^{n} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2 .$$

On peut écrire cela sous forme matricielle. Notons $D_n = \mathrm{diag}(f_{i\bullet})$ la matrice diagonale des poids des lignes et $D_p = \mathrm{diag}(f_{\bullet j})$ la matrice diagonale des poids des colonnes. La matrice $D_n^{-1}F$ a pour lignes les profils-lignes et la matrice $D_p^{-1}F^\top$ a pour lignes les profils-colonnes. la distance du χ^2 entre deux profils-lignes L_i et $L_{i'}$ s'écrit alors

$$d^2(L_i, L_{i'}) = (L_i - L_{i'})^\top D_p^{-1} (L_i - L_{i'}),$$

et de manière analogue pour deux profils-colonnes C_j et $C_{j'}$

$$d^2(C_j, C_{j'}) = (C_j - C_{j'})^\top D_n^{-1} (C_j - C_{j'}).$$

Ces distances sont à la base de la représentation géométrique dans l'analyse des correspondances, où l'on cherche une projection des profils dans un espace de faible dimension qui conserve au mieux ces distances.

4 Estimation des éléments propres

L'analyse des profils-lignes s'appelle l'analyse directe. On considère les profils-lignes contenus dans la matrice $D_n^{-1}F \in \mathbb{R}^{n\times p}$. On projette les profils-lignes dans un espace muni de la métrique du χ^2 sur les colonnes, définie par

$$\langle x, y \rangle = x^\top D_p^{-1} y.$$

L'analyse des profils-colonnes s'appelle l'analyse duale. On considère les profils-colonnes contenus dans la matrice $D_p^{-1}F^\top \in \mathbb{R}^{p\times n}$. On projette les profils-colonnes dans un espace muni de la métrique du χ^2 sur les lignes, définie par

$$\langle x, y \rangle = x^\top D_n^{-1} y.$$

Pour l'analyse directe, on cherche le premier axe factoriel, i.e. la direction $u \in R^p$ qui maximise la variance projetée des profils-lignes, sous contrainte que u soit normé. On cherche donc

$$\max_u u^\top D_p^{-1} F^\top D_n^{-1} F D_p^{-1} u, \quad \text{s.c.} \quad u^\top D_p^{-1} u = 1.$$

Ce problème d'optimisation revient à chercher le premier vecteur propre de la matrice

$$S = F^\top D_n^{-1} F D_p^{-1}.$$

La matrice S joue un rôle analogue à la matrice de covariance dans l'ACP. Le premier vecteur propre u_1 vérifie donc la relation

$$S u_1 = F^\top D_n^{-1} F D_p^{-1} u_1 = \lambda_1 u_1,$$

avec λ_1 la valeur propre associée à u_1. Les vecteurs propres de la matrice S donnent les axes factoriels dans l'espace des colonnes. Les coordonnées des profils-lignes sur le premier axe factoriel sont obtenues par la relation

$$\Phi_1 = D_n^{-1} F D_p^{-1} u_1.$$

On obtient les autres couples de valeurs propres et vecteurs propres, ainsi que les coordonnées des profils-lignes sur les axes factoriels associés de manière similaire.

L'analyse duale se fait de façon similaire. On cherche le premier vecteur propre de la matrice

$$T = F D_p^{-1} F^\top D_n^{-1}.$$

Le premier vecteur propre v_1 vérifie donc la relation

$$T v_1 = F D_p^{-1} F^\top D_n^{-1} v_1 = \mu_1 v_1,$$

avec μ_1 la valeur propre associée à v_1. Les vecteurs propres de la matrice T donnent les axes factoriels dans l'espace des lignes. Les coordonnées des profils-colonnes sur le premier axe factoriel sont obtenues par la relation

$$\Psi_1 = D_p^{-1} F^\top D_n^{-1} v_1.$$

On obtient les autres couples de valeurs propres et vecteurs propres, ainsi que les coordonnées des profils-colonnes sur les axes factoriels associés de manière similaire.

Propriété

Les matrices S et T ont les mêmes $r = \min(n-1, p-1)$ premières valeurs propres positives. Cela garantit une représentation cohérente des lignes et des colonnes dans le mêmes espace

réduit. Pour $k = 1, \dots, r$, les relations entre les vecteurs propres u_k et v_k sont

$$u_k = \frac{1}{\sqrt{\lambda_k}} F^\top D_n^{-1} v_k \quad \text{et} \quad v_k = \frac{1}{\sqrt{\lambda_k}} F D_p^{-1} u_k.$$

Preuve

En partant de l'équation

$$T v_1 = F D_p^{-1} F^\top D_n^{-1} v_1 = \mu_1 v_1,$$

en multipliant à gauche par $F^\top D_n^{-1}$, on obtient :

$$F^\top D_n^{-1} F D_p^{-1} F^\top D_n^{-1} v_1 = \mu_1 F^\top D_n^{-1} v_1.$$

Ainsi le vecteur $F^\top D_n^{-1} v_1$ est un vecteur propre de la matrice $F^\top D_n^{-1} F D_p^{-1}$ associée à la valeur propre μ_1. Comme λ_1 est la plus grande valeur propre de $F^\top D_n^{-1} F D_p^{-1}$, on en déduit que $\mu_1 \leq \lambda_1$. En procédant de la même manière, en partant de $S u_1 = \lambda_1 u_1$, on déduit que $\lambda_1 \leq \mu_1$. Donc $\lambda_1 = \mu_1$. On peut ensuite faire de même pour les r premières valeurs propres. On en déduit aussi les relations entre les valeurs propres.

Remarque

En centrant les profils, on peut projeter les profils-lignes et les profils-colonnes dans un même repère, facilitant ainsi l'interprétation géométrique conjointe.

5 Centre de gravité et inertie

Dans les sorties des logiciels de statistique, le nuage des points issus d'une AFC est généralement centré en $(0, 0)$. Cette convention reflète une analyse relative aux centres de gravité des profils-lignes et des profils-colonnes. Ce centrage est à la fois pratique et interprétable. En effet, il fait apparaître les distances entre les modalités par rapport à leur moyenne poindérée, i.e. par rapport au comportement moyen dans la population.

Chaque modalité (ligne ou colonne) est associée à un poids, correspondant à sa fréquence marginale : le poids de la ie ligne est $f_{i\bullet}$ et le poids de la je colonne est $f_{\bullet j}$. Le centre de gravité des lignes est la moyenne pondérée des profils-lignes :

$$G_L = (g_1, \dots, g_p)^\top, \quad \text{où} \quad g_j = \sum_{i=1}^{n} f_{i\bullet} \frac{f_{ij}}{f_{i\bullet}} = \sum_{i=1}^{n} f_{ij} = f_{\bullet j}, j \in \{1, \dots, p\}.$$

De même, le centre de gravité des colonnes est

$$G_C = (f_{1\bullet}, \dots, f_{n\bullet})^\top.$$

Pour recentrer les profils autour du centre de gravité, on soustrait leur valeur moyenne :

$$\frac{f_{ij}}{f_{i\bullet}} - g_j = \frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} = \frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet}}.$$

Ce centrage garantit que chaque profil-ligne $i \in \{1, \dots, n\}$ est moyenné à zéro :

$$\sum_{j=1}^{p} \frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet}} = 0.$$

L'AFC ne se fait donc plus sur la matrice S mais plutôt sur une matrice centrée $S^\star = (s^\star_{jj'})$, où

$$s^\star_{jj'} = \sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i\bullet} f_{\bullet j}\right)\left(f_{ij'} - f_{i\bullet} f_{\bullet j'}\right)}{f_{i\bullet} f_{\bullet j'}}.$$

Par définition, la trace de la matrice $S^\star$ donne l'inertie totale :

$$\operatorname{tr}(S^\star) = \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i\bullet} f_{\bullet j}\right)^2}{f_{i\bullet} f_{\bullet j}}.$$

Celle-ci correspond à la statistique du χ^2 normalisée que l'on utilise pour tester l'indépendance entre les variables.

Propriété

On a que, pour tout $j, j' \in \{1, \dots, p\}$, $s^\star_{jj'} = s_{jj'} - f_{\bullet j}$, où

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i\bullet} f_{\bullet j'}}.$$

Preuve

La propriété précédente entraine que les matrices S et $S^\star$ one les mêmes vecteurs propres pour les p premières dimensions, ce qui permet d'effectuer l'analyse factorielle sur la version centrée.

6 Coordonnées factorielles

On a que, pour tout $k = 1, \dots, r$,

$$\Phi_k = D_n^{-1} F D_p^{-1} u_k \quad \text{et} \quad \Psi_k = D_p^{-1} F^\top D_n^{-1} v_k.$$

Or, on a aussi vu les relations entre les vecteurs propres u_k et v_k,

$$u_k = \frac{1}{\sqrt{\lambda_k}} F^\top D_n^{-1} v_k \quad \text{et} \quad v_k = \frac{1}{\sqrt{\lambda_k}} F D_p^{-1} u_k.$$

On en déduit donc les relations entre les coordonnées factorielles des profils-lignes et celles des profils-colonnes :

$$\Phi_k = \frac{1}{\sqrt{\lambda_k}} D_n^{-1} F \Psi_k \quad \text{et} \quad \Psi_k = \frac{1}{\sqrt{\lambda_k}} D_p^{-1} F^\top \Phi_k.$$

On peut maintenant examiner ces relations sur chacune des composantes :

$$[\Phi_k]_i = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\bullet}} [\Psi_k]_j \quad \text{et} \quad [\Psi_k]_j = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{\bullet j}} [\Phi_k]_i \,,$$

où $[\Phi_k]_i$ désigne la coordonnée du profil-ligne L_i sur le ke axe factoriel et $[\Psi_k]_j$ désigne la coordonnée du profil-colonne C_j sur le même axe factoriel. Ces relations expriment, à un facteur $1/\sqrt{\lambda_k}$ près, que chaque profil-ligne est au barycentre des projections des profils-colonnes affectés du poids de la colonne j dans la ligne i et que chaque profil-colonne est au barycentre des projections des profils-lignes affectés du poids de la ligne i dans la colonne j.

> Remarque
>
> Ainsi, en AFC, nous avons une double représentation barycentrique. Sur les axes factoriels, chaque point d'un nuage est au barycentre des points de l'autre nuage.