

Cours d'Analyse des Données Multidimensionnelles

Chapitre 2 : Analyse Factorielle des Correspondances (AFC)

Zaineb Smida

5ème année GEA
INSA Lyon

3 décembre 2024

Généralités

- Objectif identique à celui de l'ACP : identifier un petit nombre de dimensions pour simplifier et interpréter un ensemble de données peu lisibles au premier abord.
- **Différence :**
 - l'AFC ne concerne pas le même type de données que l'ACP.
 - l'AFC est une méthode privilégiée de description de variables qualitatives.
 - lorsque deux variables qualitatives sont mesurées sur n individus, les données sont stockées dans un *tableaux de contingence* : résultat d'un tri croisé entre 2 variables **qualitatives** X et Y .
 - l'AFC traite des *tableaux de contingence* et permet la description graphique et numérique de ces tableaux.

Exemple

- *Un chef de produit désire cibler la clientèle d'une nouvelle lessive écologique. Il voudrait notamment savoir quelle est la tranche d'âge la plus réceptive à ce produit.*
- *Echantillon de 391 personnes.*
- *Tri croisé entre 6 classes d'âge et une variable "Achat" à 4 modalités (systématiquement, la plupart du temps, occasionnellement, jamais).*

Tableau de contingence = Tableau **variable/variable** avec 2 variables qualitatives (classe d'**âge** et fréquence d'**achat**) à respectivement $L = 6$ et $C = 4$ modalités

Achat Âge	Syst	Lpdt	Occas	Jamais
15 à 19 ans	6	6	24	9
20 à 24 ans	2	25	37	6
25 à 34 ans	5	17	25	9
35 à 44 ans	12	29	37	3
45 à 59 ans	3	45	36	12
60 ans et plus	11	19	9	4

Exemple

Voir les notes de cours : Tableau de contingence

Question : le comportement d'achat de produits écologiques **est-il lié** à l'âge ?

- le test du χ^2 d'indépendance permet de tester cette hypothèse.
- l'AFC propose une analyse graphique pour approfondir l'analyse de la liaison lorsque l'hypothèse d'indépendance est rejetée.

Notations :

- n_{ij} : effectif **conjoint**,
- $n_{i.} = \sum_{j=1}^C n_{ij}$, $n_{.j} = \sum_{i=1}^L n_{ij}$: effectifs **marginaux**.

Remarque

*Contrairement au tableau individus/variables de l'ACP, les lignes et colonnes en AFC jouent un rôle symétrique (**pas de notion d'individus et de variables** mais uniquement des **modalités**).*

- La mesure du χ^2 : permet de mesurer l'écart entre le tableau de contingence observé et un tableau avec indépendance parfaite des variables X et Y

$$\chi^2 = \sum_{i=1}^L \sum_{j=1}^C \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

avec $e_{ij} = \frac{n_{i.} \times n_{.j}}{n} \quad \forall i, \forall j$: effectif attendu pour la cellule (i, j) si indépendance parfaite entre X et Y

- Test du χ^2 d'indépendance :
 - H_0 : X et Y sont indépendantes
 - On rejette H_0 au niveau α (5% en général) si

$$\chi^2 > \chi_{(L-1)(C-1), 1-\alpha}^2$$

où $\chi_{(L-1)(C-1), 1-\alpha}^2$ est le quantile d'ordre $1 - \alpha$ d'une loi de χ^2 à $(L - 1)(C - 1)$ degrés de liberté.

Exemple

*Voir les notes de cours : voir le **Tableau : Test du χ^2 d'indépendance.***

$$\chi_{obs}^2 = 49,551$$

$$p\text{-value} = 1,425 \times 10^{-5} < 5\%.$$

Donc on rejette H_0 . L'âge et le comportement d'achat de lessive écologique sont donc liés.

*Représentation graphique de la liaison à l'aide d'un diagramme en colonnes juxtaposées : voir **Figure : Diagramme 1 des profils lignes** et **Figure : Diagramme 2 des profils lignes**.*

Analyse des contributions au χ^2

- Supposons que le test précédent ait rejeté H_0 et que l'on conclut donc que **les deux variables sont liées**. On cherche alors à expliquer **cette liaison**.
- On peut s'intéresser aux cellules (i, j) qui ont **le plus contribué** à la mesure du χ^2
 \hookrightarrow cellules (i, j) qui ont donné des $c_{ij} \equiv \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$ **élevés**.
- Les c_{ij} sont appelés **contributions** au χ^2

$$\hookrightarrow \chi^2 = \sum_{i=1}^L \sum_{j=1}^C c_{ij}$$

Analyse des contributions au χ^2 :

- \Leftrightarrow commenter les grands c_{ij}
- \Leftrightarrow expliquer d'où provient l'écart à l'indépendance et dans quel sens est cet écart.

Exemple

Voir les notes de cours : Tableau : Résultats des contributions au χ^2
 Seuil = $\chi^2/L \times C = 49,55/(6 \times 4) = 2,06$.

On repère les **4 plus fortes contributions**, puis pour ces 4 couples de modalités les effectifs attendus e_{ij} (**Tableau : Effectifs théoriques**) aux effectifs observés n_{ij} (tableau Stat. Des. 2 OU BIEN : **Tableau de contigence**) :

- 60 ans et plus \times Syst : obs $>$ att : sur-représentation des acheteurs systématiques de lessive écologique de plus de 60 ans
- 60 ans et plus \times Occas : obs $<$ att : sous-représentation
- 15-19 ans \times Lpdt : obs $<$ att : sous-représentation
- 45-59 ans \times Syst : obs $<$ att : sous-représentation

Analyse intéressante mais :

- lecture difficile du tableau,
- ne permet pas de comparer les lignes entre elles ou les colonnes entre elles.

Nuages de points associés à un tableau de contingence

Tableaux des profils lignes et des profils colonnes

- **Profils-lignes** : vecteurs lignes $\frac{n_{ij}}{n_{i.}}$ pour i fixé et $j = 1, \dots, C$
- ⇒ Tableau des L profils-lignes qui donne pour chaque ligne (modalité X_i pour $i = 1, 2, \dots, L$) la répartition des modalités des colonnes (**distribution conditionnelle** de Y sachant $X = X_i$).
- ⇒ On obtient un tableau $L \times C$ tel que **les sommes en ligne sont égales à 100 %** et l'on considère que l'on a L "individus" (les L profils-lignes) et C "variables" (les C modalités de Y).
- ⇒ Les profils-lignes forment un nuage de L points dans \mathbb{R}^C . On affecte à chacun de ces points un poids proportionnel à sa fréquence marginale ($n_{i.}/n$).

Exemple

Voir les notes de cours : voir Tableau : Profils lignes et Figure : Diagramme 1 des profils lignes.

- Une ligne : distribution, pour une classe d'âge donnée, des fréquences de consommation de produits écologiques.
- Première ligne : comment se répartissent les 15-19 ans en matière de consommation.
- Somme en ligne de 100%. Voir Figure : Diagramme 2 des profils lignes.

Par exemple 13% des 15-19 ans achètent systématiquement de la lessive écologique.

Nuages de points associés à un tableau de contingence

Tableaux des profils lignes et des profils colonnes

- **Profils-colonnes** : vecteurs colonnes $\frac{n_{ij}}{n_{.j}}$ pour j fixé et $i = 1, \dots, L$
- ⇒ Tableau des C profils-colonnes qui donne pour chaque colonne la répartition des modalités des lignes (somme en colonne de 100 %)
- ⇒ On obtient un tableau avec C “individus” et L “variables”.
- ⇒ Les profils-colonnes forment un nuage de C points dans \mathbb{R}^L . On affecte à chacun de ces points un poids proportionnel à sa fréquence marginale ($n_{.j}/n$).

Exemple

*Voir les notes de cours : voir **Tableau : Profils colonnes** et **Figure : Diagramme 1 des profils colonnes** .*

- *Une colonne : distribution, pour une fréquence de consommation donnée, des classes d'âge.*
- *Première colonne : comment se répartissent les consommateurs systématiques en matière de classe d'âge.*
- *Somme en colonne de 100%. Voir **Figure : Diagramme 2 des profils colonnes**.*

Par exemple 28 % des acheteurs systématiques de lessive écologique ont 60 ans et plus.

Point moyen des profils-lignes et des profils-colonnes

- Le point moyen (le centre de gravité du nuage de points) des profils-lignes est noté $g_L \in \mathbb{R}^C$, de composantes

$$g_{Lj} = \sum_{i=1}^L \frac{n_{ij}}{n_{i.}} \times \frac{n_{i.}}{n} = \frac{n_{.j}}{n}, \quad j = 1, \dots, C$$

⇒ C'est le **profil-ligne marginal** (distribution marginale de Y en fréquence).

- Le point moyen des profils colonnes est noté $g_C \in \mathbb{R}^L$, de composantes

$$g_{Ci} = \frac{n_{i.}}{n}, \quad i = 1, \dots, L$$

⇒ C'est le **profil-colonne marginal** (distribution marginale de X en fréquence).

Remarque

Dans le cas de l'indépendance parfaite,

$$\forall i, \forall j, \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} = g_{Lj} \text{ et } \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} = g_{Ci}$$

⇒ les deux nuages sont réduits chacun à un seul point (leur centre de gravité).

- L'étude de la forme de ces nuages de points permet de rendre compte de la structure des écarts à l'indépendance.

⇒ il faut choisir **une métrique pour chacun des espaces.**

Distance du χ^2

1 Entre deux profils-lignes i et i'

$$d_{\chi^2}^2(i, i') = \sum_{j=1}^C \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2$$

$\hookrightarrow d_{\chi^2}^2$ est appelée distance du χ^2

- $d_{\chi^2}^2$ est différent de la distance euclidienne à cause du terme de pondération $\frac{n}{n_{.j}}$.
- La métrique usuelle favorise les colonnes à forts effectifs ($n_{.j}$ grand) pour lesquelles les fortes variations sont fréquentes.
- La distance du χ^2 évite ce phénomène en pondérant plus faiblement les colonnes à fort effectif.

- l'inertie totale du nuage de points des profils-lignes est définie par

$$I_L = \sum_{i=1}^L \frac{n_{i.}}{n} d_{\chi^2}^2(i, g_L)$$

↪ c'est une mesure de la variabilité des profils-lignes

Remarque

On peut définir *l'inertie du profil-ligne i* par : $I(i) = \frac{n_{i.}}{n} d_{\chi^2}^2(i, g_L)$. D'où

$$I_L = \sum_{i=1}^L I(i).$$

Proposition

$$I_L = \frac{\chi^2}{n}$$

2. Entre deux profils-colonnes j et j'

$$d_{\chi^2}^2(j, j') = \sum_{i=1}^L \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}} \right)^2$$

⇒ Inertie du nuage de points des profils-colonnes :

$$I_C = \sum_{j=1}^C \frac{n_{.j}}{n} d_{\chi^2}^2(j, g_C) = \sum_{j=1}^C I(j)$$

Proposition

$$I_C = \frac{\chi^2}{n} = I_L$$

Etapes de l'AFC

Calculs de type ACP

- L'AFC est une méthode de type "ACP".
- On réalise 2 "ACP" : une sur le **tableau des profils lignes**, une sur le **tableau des profils colonnes** afin de réduire la dimension des profils.
- Pour chaque tableau on fait une ACP sur la **matrice d'inertie** (ou matrice de variance-covariance) : non donnée dans ce cours.
- On calcule les **valeurs propres** et les **vecteurs propres** des deux matrices.

Exemple

Voir les notes de cours : Tableau : Profils lignes et Tableau : Profils colonnes.

$L=6$ profils-lignes sont des vecteurs de \mathbb{R}^4 . $C=4$ profils-colonnes sont des vecteurs de \mathbb{R}^6 .

Etapes de l'AFC

Propriétés de l'AFC

- Les résultats de l'ACP sur le tableau des profils colonnes peuvent être obtenus à partir des résultats de l'ACP sur le tableau des profils lignes.
- On obtient **les mêmes valeurs propres** dans les deux ACP.
- On obtient un **nombre de valeurs propres** égal à $\min(L, C) - 1$ avec $1 > \lambda_1 > \lambda_2 > \dots > \lambda_{\min(L, C)-1}$.
- On obtient pour chaque ACP $\min(L, C) - 1$ **axes principaux**.
L'inertie associée à l'axe principal i vaut λ_i .
- Comme en ACP, **l'inertie** (du nuage des profils lignes ou colonnes) **est la somme des valeurs propres** :

$$I_C = I_L = \sum_{\ell=1}^{\min(L, C)-1} \lambda_{\ell}$$

On a $\lambda_{\ell} = \sum_{i=1}^L \frac{n_{i.}}{n} \text{Coord}_{c_{\ell}}^2(X_i)$ où $\text{Coord}_{c_{\ell}}(X_i)$ est la coordonnée de la modalité X_i sur l'axe c_{ℓ} .

Exemple

Voir les notes de cours : voir Tableau : AFC-Valeurs propres pour les valeurs propres (1ère colonne)

- $L = 6, C = 4, \min(L, C) = 4, \min(L, C) - 1 = 3$ valeurs propres donc 3 nouveaux axes peuvent être construits.

- $$I = \sum_{i=1}^{\min(L,C)-1} \lambda_i = 0,12$$

Choix de la dimension

- On obtient les mêmes valeurs propres pour les lignes et les colonnes \Rightarrow on choisit une seule fois la dimension.
 - Centrer et réduire les variables n'a aucun sens en AFC \Rightarrow la valeur 1 ne représente rien \Leftrightarrow le critère de Kaiser n'a pas de sens.
- \Rightarrow il faut utiliser le critère de la part d'inertie expliquée.

Exemple

Voir les notes de cours :

- voir **Tableau : AFC-Valeurs propres** (3ème colonne) \Rightarrow On choisit de retenir $k = 2$ axes avec plus de 84% d'inertie expliquée.
- voir **Tableau : AFC-Coordonnées des profils-lignes** et **Tableau : AFC-Coordonnées des profils-colonnes** pour les coordonnées des nouveaux axes retenus.

Interprétation des axes : les contributions

- En AFC, on dispose de variables qualitatives
 - ⇒ pas de notion de corrélations
 - ⇒ pas de graphique des corrélations
- Pour interpréter un axe principal, on utilise la notion de contribution des modalités (de X ou de Y) à l'inertie de cet axe :
 Pour l'ACP sur le tableau des profils-lignes, la contribution (en pourcentage) de la modalité X_i de X à l'inertie de l'axe principal c_ℓ vaut :

$$CRT_\ell(X_i) = \frac{\frac{n_{i.}}{n} \text{Coord}_{c_\ell}^2(X_i)}{\lambda_\ell} * 100$$

La somme des contributions de toutes les modalités X_i , $i = 1, \dots, L$, à un axe vaut 100.

↔ on considère qu'une modalité X_i a contribué à l'axe c_ℓ si

$$CRT_\ell(X_i) \geq \frac{100}{L}.$$

Exemple

Voir les notes de cours :

voir **Tableau : AFC-Contributions des profils-lignes** pour les contributions des profils-lignes (*seuil* = $100/6 = 16,67$)

- 1^{er} axe : 60 ans et plus, 20-24 ans.
- 2^{ème} axe : 15-19 ans, 45-59 ans.

Voir **Tableau : AFC-Contributions des profils-colonnes** pour les contributions des profils-colonnes (*seuil* = $100/4 = 25$)

- 1^{er} axe : **Syst.**
- 2^{ème} axe : **LPDT.**

Représentation graphique

- On a retenu k (petit) axes principaux ou axes factoriels sur lesquels on projette simultanément toutes les modalités de X et de Y .
- Sur le graphique obtenu, la distance entre deux profils-lignes (resp. colonnes) correspond à la distance du χ^2 entre ces deux profils.
- Il faut donc projeter deux nuages de points.
 - On peut les projeter **successivement** : on obtient deux graphiques. Sur le graphique des profils lignes, la distance entre deux profils lignes correspond à la distance du χ^2 entre ces deux profils.
 - On peut les projeter **simultanément** (ce qui est fait avec R).

Interprétation

- Comme en ACP, pour un axe retenu, il ne faut interpréter que les profils ayant **bien contribué à l'inertie de l'axe** et étant **bien représentés sur l'axe**.

↪ même mesure qu'en ACP pour **la qualité de représentation sur l'axe c_ℓ** :

$$\cos_\ell^2(X_i) = \frac{\text{Coord}_{c_\ell}^2(X_i)}{\min(L, C) - 1 \sum_{r=1} \text{Coord}_{c_r}^2(X_i)}$$

- Il faut interpréter successivement la proximité des profils- lignes puis celle des profils-colonnes.
- Deux profils-lignes proches représentent deux modalités de X avec des répartitions (distributions conditionnelles) des modalités de Y assez semblables.
- Sur un axe donné, un profil-ligne proche (resp. éloigné) d'un profil-colonne correspond à une attraction (resp. répulsion) du couple de modalités (ligne, colonne).

Exemple

Voir les notes de cours :

Qualité de représentation des profils (Cos2) : Voir les tableaux : ACP-Cos2 des profils lignes et ACP-Cos2 des profils colonnes

- **Axe 1** ($\cos^2 > 0,5$) : 20-24 ans, 60 ans et plus, **Syst**, **Occas**.
- **Axe 2** ($\cos^2 > 0,25$) : 15-19 ans, 25-34 ans, 45-49 ans, **Lpdt**, **jamais**.

Profils à la fois bien représentés et de forte contribution

- **Axe 1** : 20-24 ans, 60 ans et plus, **Syst**.
- **Axe 2** : 15-19 ans, 45-49 ans, **Lpdt**.

Interprétation du graphique simultané : Voir AFC-Graphique des modalités (profils)

- **Axe 1** : sur-représentation **des acheteurs systématiques** de lessive écologique parmi **les plus de 60 ans** au contraire **des 20-24 ans**.
- **Axe 2** : sous-représentation des **15-19 ans** achetant **la plupart du temps** de la lessive écologique au contraire des **45-59 ans**.

Conclusion

Justifications de l'AFC :

- Données sous formes de *tableau de contingence*, i.e. *deux variables qualitatives*.
- *Beaucoup de modalités* pour les deux variables (grand tableau).
- Deux variables qualitatives *dépendantes* (au sens du *test du χ^2 d'indépendance*).
- Permet de retrouver les résultats de l'analyse des contributions.