Support de cours

ANALYSE DES DONNEES I

CHAPITRE 3 : L'ANALYSE FACTORIELLE DES CORRESPONDANCES

Iphygénie SARR
Ingénieure statisticienne économiste

Ecole Nationale de la Statistique et de l'Analyse Economique Pierre Ndiaye
Mai 2024

OBJECTIF DU CHAPITRE

L’objectif de ce chapitre est de décrire l’Analyse factorielle des Correspondances (AFC). Elle est une méthode d’analyse factorielle qui s’applique à des bases de données formée de deux (02) variables qualitatives.

PLAN DU CHAPITRE

01. Tableau et notations

02. Analyse directe du nuage $\mathcal{N}(\mathrm{I})$

03. Analyse duale du nuage $\mathcal{N}(\mathrm{J})$

04. Cas pratique d'une AFC sur R studio

01.

Tableau et notations

Tableau de contingence

- L’analyse factorielle des correspondances est particulièrement adaptée à l’étude des tableaux de contingence, à laquelle elle fournit un outil d’analyse puissant.
- On appelle tableau de contingence un tableau qui donne la ventilation d’une population selon deux caractères qualitatifs que l’on croise.

Tableau de contingence

Modalité *i* \ Modalité *j*		k_{ij}		
				k_i
		k_j		k

Tableau de contingence

Par exemple : le tableau qui ventile la population sénégalaise recensées en 2013 selon les deux caractères « région » et « classe d'âge ».

- La case courante du tableau (ligne i et colonne j) contient le nombre k_{ij}.
- Les marges du tableau contiennent les distributions marginales k_I et k_J.
- La somme des éléments du tableau est égale à k.

Tableau de contingence

- On reconnait un tableau de contingence à ceci : en calculant des sommes en ligne ou en colonne, on obtient des quantités qui ont un sens.
- Ainsi les tableaux non contingents indiquent les valeurs d'un ensemble hétéroclite de variables observées sur une population donnée pour lesquelles la somme des diverses variables relatives à une même modalité n'a pas de sens.
- Sur des tableaux de contingence, on peut donc calculer des fréquences en divisant la valeur d'une case par la valeur de la somme du tableau.

Notations

- Considérons une population E d'effectif k répartie selon deux caractères qualitatifs I et J, possédant respectivement N et K modalités.
- Le tableau de contingence qui donne la ventilation de E selon le croisement $I \times J$ des deux caractères est un tableau à N lignes et K colonnes, de terme courant k_{ij}.
- k_{ij} étant le nombre d'individus ayant simultanément les modalités i et j des variables (qualitatives) I et J.

Notations

- Nous noterons les fréquences empiriques calculées à partir des k_{ij} par la relation $f_{ij} = \frac{k_{ij}}{k}$.
- f_{ij} est la fréquence du couple (i, j) dans la population considérée.
- Sur la base de cela, on précise les notations suivantes :

1. $f_i = \sum_j f_{ij}$
2. $f_j = \sum_i f_{ij}$
3. $f_j^i = \frac{f_{ij}}{f_i}$
4. $f_i^j = \frac{f_{ij}}{f_j}$

Notations

- Les nombres f_j^i ou f_i^j sont particulièrement intéressants.
- f_j^i représente la fréquence de la classe j dans la classe i.
- C'est-à-dire la proportion des individus possédant la modalité j parmi ceux possédant la modalité i.
- Ainsi pour chaque modalité i prise à part, celui-ci est associé à un vecteur à K coordonnées : $\left(f_1^i, f_2^i, \dots, f_K^i\right)^T$
- Et vice versa…

Notations

- On peut ainsi associer à tout tableau de contingence deux nuages de points $\mathcal{N}(\mathrm{I})$ et $\mathcal{N}(\mathrm{J})$.
- Un nuage $\mathcal{N}(\mathrm{I})$ composé des N points X^{i} situés dans l'espace $\mathbb{R}^{\mathrm{K}}$ dotés chacun de la masse f_{i} et de coordonnée courante : $\mathrm{x}_{\mathrm{ij}} = \mathrm{f}_{\mathrm{j}}^{\mathrm{i}}$
- Un nuage $\mathcal{N}(\mathrm{J})$ composé des K points Y^{j} situés dans l'espace $\mathbb{R}^{\mathrm{N}}$ dotés chacun de la masse f_{j} et de coordonnée courante : $\mathrm{y}_{\mathrm{ij}} = \mathrm{f}_{\mathrm{i}}^{\mathrm{j}}$

Notations

- Pour distinguer les deux nuages $\mathcal{N}(\mathrm{I})$ et $\mathcal{N}(\mathrm{J})$, il est convenu d'appeler l'un (quelconque) des deux nuages le « nuage direct » ; et l'autre le « nuage dual ».
- A partir de cela, il devient possible d'établir l'ensemble des résultats de l'analyse factorielle des correspondances établis dans le chapitre 1 du cours à partir des nuages ainsi construits.

02.

Analyse directe du nuage $\mathcal{N}(\mathbf{I})$

Nuage direct

Considérons $\mathcal{N}(\mathrm{I})$ comme étant le nuage direct. L'objectif d'une telle représentation sera aussi de mettre en évidence les relations entre les « individus » qui sont ici des modalités. C'est-à-dire :

- comparer les profils lignes entre eux ;
- détecter les écarts d'indépendance entre les modalités ;
- mettre en évidence les modalités de X^{i} et $\mathrm{X}^{\mathrm{i}'}$ qui s'attirent et qui se repoussent.

Nuage direct

Pour ce faire, nous considérons dans l'espace $\mathbb{R}^K$ la métrique du χ^2 centrée sur f_J, de sorte que :

$$d^2(X^i, X^{i\prime}) = \sum_j \frac{1}{f_j}(f_j^i - f_j^{i\prime})^2$$

La distance entre deux points $\boldsymbol{X^i}$ et $\boldsymbol{X^{i\prime}}$ de $\mathcal{N}(\mathbf{I})$ est la distance entre les deux distributions f_J^i et $f_J^{i\prime}$ calculée selon la métrique du χ^2 centrée sur f_J. On peut donc interpréter cette distance comme une distance entre deux distributions, ou, entre deux structures.

Nuage direct

- Cependant, nous avons aussi le choix : soit continuer à travailler avec les nuages tels qu'ils sont définis ci-dessus (donc avec la métrique χ^2), soit modifier les coordonnées des points de telle sorte que l'on puisse utiliser la métrique euclidienne canonique.
- L'avantage décisif de cette transformation nous permet de nous ramener au chapitre 1 établi dans le cas où la métrique utilisée est la métrique euclidienne canonique.

Nuage direct

- Nous nous effectuons donc la transformation suivante sur les points du nuages $\mathcal{N}(\mathrm{I})$
- $\mathcal{N}(\mathrm{I})$ est désormais composé des N points X^{i} situés dans l'espace $\mathbb{R}^{\mathrm{K}}$ dotés chacun de la masse f_{i} et de coordonnée courante : $\mathrm{x}_{\mathrm{ij}} = \frac{1}{\sqrt{\mathrm{f}_{\mathrm{j}}}} \mathrm{f}_{\mathrm{j}}^{\mathrm{i}}$
- On remarquera qu'avec les coordonnées ainsi définies, et en utilisant la métrique euclidienne canonique, on retrouve la même distance que celle définie avec le χ^2.

Inertie du nuage

Le centre de gravité d'un tel nuage $\mathcal{N}(I)$ est le point G tel que :

$$G = \sum_i f_i X^i$$

$$g_j = \sum_i f_i \frac{1}{\sqrt{f_j}} f_j^i$$

$$g_j = \sum_i \frac{f_{ij}}{\sqrt{f_j}} = \frac{f_j}{\sqrt{f_j}}$$

$$g_j = \sqrt{f_j}$$

Inertie du nuage

L'inertie totale de $\mathcal{N}(\mathrm{I})$ est alors :

$$In_G(I) = \sum_i f_i \left\| X^i - G \right\|^2$$

Or $\left\| X^i - G \right\|^2 = \sum_j \left(x_{ij} - g_j \right)^2$

$$\left\| X^i - G \right\|^2 = \sum_j \left(\frac{f_{ij}}{f_i \sqrt{f_j}} - \sqrt{f_j} \right)^2$$

$$\left\| X^i - G \right\|^2 = \sum_j f_j \left(\frac{f_{ij} - f_i f_j}{f_i f_j} \right)^2$$

Inertie du nuage

D'où :

$$In_G(I) = \sum_i f_i \sum_j f_j \left(\frac{f_{ij} - f_i f_j}{f_i f_j} \right)^2$$

$$In_G(I) = \sum_{ij} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

L'inertie de $\mathcal{N}(\mathrm{I})$ est donc égale à l'écart d'indépendance entre les variables I et J mesuré par le coefficient de corrélation du χ^2.

03.

Analyse duale du nuage $\mathcal{N}(\mathbf{J})$

Nuage dual

L'analyse du nuage $\mathcal{N}(J)$ composé des K points Y^j situés dans l'espace $\mathbb{R}^N$ dotés chacun de la masse f_j et de coordonnée courante : $y_{ij} = f_i^j$ se fait soit :

- en reprenant le même procédé que dans l'analyse direct ;
- en utilisant des formules de transition.

04.

Cas pratique d’une AFC sur R studio

Présentation de la base de données

Une base de données contenant les fréquences d'execution de 13 tâches ménagères au sein d'un couple :

- **Variable I :** acteurs
- **Variable J :** tâches
- **Nombre de modalités de la variable I :** 4
- **Nombre de modalité de la variable J :** 13

SYNTHESE

Quelques points essentiels à retenir :

- l'AFC s'intéresse à deux variables qualitatives en d'autres termes aux bases de données étant sous la forme de tableaux de contingence ;
- de manière pratique on peut apporter des modifications à ces données brutes afin de se ramener aux résultats obtenus lors d'une analyse factorielle quelconque ;
- c'est ce parti pris que l'on a eu à adopter dans ce chapitre ;
- la lecture de ses nuages doit toujours être accompagné de l'utilisation des aides à l'interprétation.

FIN DU CHAPITRE

MERCI POUR VOTRE ATTENTION !