



HAL
open science

Analyse des Correspondances Multiples

Denis Laloë

► **To cite this version:**

Denis Laloë. Analyse des Correspondances Multiples. Maitrise. Analyse des correspondances, Clermont-Ferrand, France. 2023, pp.16. <hal-04443661>

HAL Id: hal-04443661

<https://hal.inrae.fr/hal-04443661v1>

Submitted on 7 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Analyse des Correspondances Multiples

Février 2023

Denis Laloë
GABI - GiBBS

Février 2023

Introduction

Individus décrits par plusieurs variables qualitatives (facteurs)

Tableau

- Lignes : individus / observations.
- Colonnes : facteurs (sexe, CSP, pays,...)

Objectif

- Objectif: trouver des scores qui maximisent l' "explication" de la variation des facteurs
 - R^2 d'un modèle score = facteur
 - η^2 : rapport de corrélation

Extension de l'AFC à > 2 facteurs

- n individus
- K facteurs
- $p = \sum_{k=1} K p_k$

Un exemple: les prix Nobel

Alhuzali T, Beh EJ, Stojanovski E (2022) Multiple correspondence analysis as a tool for examining Nobel Prize data from 1901 to 2018. PLoS ONE 17(4): e0265929.

Prix Nobel décrits par quatre facteurs

- 785 individus $n = 785$
- 4 facteurs $K = 4$
 - Période : 1901-1940, 1941-1980, 1981-2018 : $p_1 = 3$
 - Genre : Masculin/Féminin $p_2 = 2$
 - Pays : BI, CA, DE, FR, IT, JP, RU, US : $p_3 = 8$
 - Discipline : Ch, Ec, Li, Me, Pc, Ph : $p_4 = 6$
 - $p = 19$

Table de données

Période	Genre	Pays	Discipline
1901-1940	M	CA	Me
1901-1940	M	CA	Me
1901-1940	M	FR	Ch
1901-1940	M	FR	Ch
1901-1940	M	FR	Ch
1901-1940	M	FR	Ch

Table des données : tableau disjonctif \mathbf{Z}

$$\mathbf{Z} = [\mathbf{Z}_1 | \mathbf{Z}_2 | \dots | \mathbf{Z}_K]$$

P1901-	P1941-	P1981-	F	M	BI	CA	DE	FR	IT	JP	RU	US	Ch	Ec	Li	Me
1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1
1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1
1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0

Table des données : du tableau disjonctif au tableau de Burt

$$B = Z^T Z$$

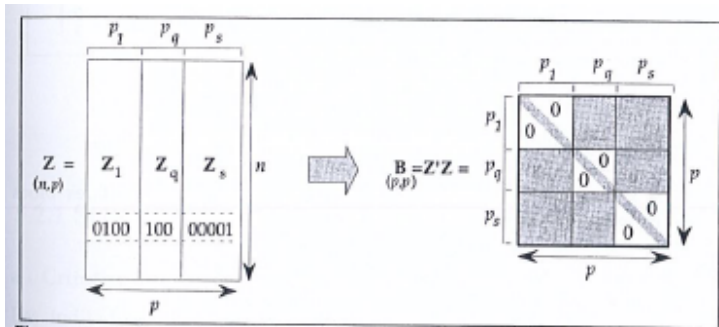


Figure 5.1 – 3. Construction du tableau des faces de l'hypercube (tableau de Burt) B à partir du tableau disjonctif complet Z

Construction des tables avec R

ade4

- acm.disjonctif
- acm.burt

FactoMineR

- tab.disjonctif

Distances entre points

- entre deux individus : $d^2(i, i') = \frac{1}{K} \sum_{j=1}^p \frac{n}{z_j} (z_{ij} - z_{i'j})^2$
 - 2 individus avec mêmes modalités: $d=0$
 - Plus de poids pour les modalités rares (z_j petit)
- entre deux modalités : $d^2(j, j') = \sum_{i=1}^n \left(\frac{z_{ij}}{z_j} - \frac{z_{ij'}}{z_{j'}} \right)^2$

Schéma de dualité

$$(\mathbf{X}, \mathbf{Q}, \mathbf{D}) \Leftrightarrow \begin{array}{ccc} \langle p \rangle & \xrightarrow{\mathbf{Q}} & \langle p \rangle \\ \mathbf{X}' \uparrow & & \downarrow \mathbf{X} \\ \langle n \rangle & \xleftarrow{\mathbf{D}} & \langle n \rangle \end{array}$$

- $\mathbf{X}(\mathbf{n}, \mathbf{p})$: matrice de données (transformées)
- $\mathbf{Q}(\mathbf{p}, \mathbf{p})$: matrice d'un produit scalaire, carrée symétrique
- $\mathbf{D}(\mathbf{n}, \mathbf{n})$: matrice d'un produit scalaire, carrée symétrique

Transformation du tableau disjonctif

Notations

- Individu i , facteur k , modalité j intra-facteur k
- $\delta_{ikj} = 1$ si l'individu i a la modalité j pour le facteur k , 0 sinon
- n_{kj} : nombre d'individus à modalité j pour le facteur k
- f_{kj} : fréquence de la modalité j pour le facteur k

Structure du tableau

- Somme de la colonne kj , facteur k , modalité j intra-facteur k : n_{kj}
- Somme des lignes intra-facteur : 1

Transformation du tableau disjonctif. Schéma de dualité

Triplet

- Table : $\frac{\delta_{ikj} - f_{kj}}{f_{kj}}$; Somme pondérée intra-bloc=0
 - : centrage : Somme pondérée intra-bloc=0
 - : centrage : Somme par colonne =0
 - : Réduction : Division par f_{kj}
- Poids des lignes : $\frac{1}{n}$; Total = 1
- Poids des colonnes : $\frac{n_{kj}}{Kn} = \frac{f_{kj}}{K}$; Total = 1

Un exemple de table

Table $\frac{\delta_{ikj} - f_{kj}}{f_{kj}}$

	Période.1901.1940	Période.1941.1980	Période.1981.2018	Genre.F	Genre.M
1	4.567376	-1.000000	-1.000000	-1.000000	0.0579515
200	-1.000000	1.875458	-1.000000	-1.000000	0.0579515
400	-1.000000	-1.000000	1.115903	-1.000000	0.0579515
743	4.567376	-1.000000	-1.000000	17.25581	-1.0000000

Table des fréquences des genres

Var1	Freq
F	0.0547771
M	0.9452229

On vérifie que la somme pondérée par les fréquences = 0 :
 (17.25581 * 0.0547771 - 0.9452229)

Quelques propriétés

Relations barycentriques lignes/colonnes

- Positionnement des modalités au barycentre des individus correspondants (à un facteur près) :
Coordonnée moyenne d'individus avec la modalité j du facteur k , sur l'axe i : $\sqrt{\lambda_i} C_{kj}(i)$

Rang et inerties

- Rang $\leq \min(n - 1, p - K)$
- Inertie d'une modalité : $l_j = \frac{1}{K} \left(1 - \frac{n_{.j}}{n}\right)$: les modalités rares contribuent plus à l'inertie
- Inertie d'un facteur : $l_k = \frac{1}{K} (p_q - 1)$: les facteurs à grand nombre de modalités contribuent plus à l'inertie. Eviter les facteurs à trop grand nombre de modalités, et les modalités rares.
- Inertie totale : $I = \frac{p}{K} - 1$

Lien entre ACM sur tableau disjonctif et AC sur tableau de Burt

	ACM sur tableau disjonctif	AC sur tableau de Burt
Valeurs propres	λ_i	$\mu_i = \lambda_i^2$
Coordonnées des modalités	$C(k, i)$	$\sqrt{\lambda_i} C(k, i)$

Lien entre facteurs et axes

Rapport de corrélation η^2

- y : score quantitatif
- f : variable qualitative
- η^2 : R^2 du modèle linéaire $y=f+e$

Lien entre score des lignes et modalités

- l'ACM maximise la moyenne des rapports de corrélation entre scores des axes et facteurs. L'inertie d'un axe est la moyenne de ces rapports de corrélation.
- $\frac{1}{K} \sum_{i=1}^K \eta^2(q^{(j)}, y)$

Aide à l'interprétation

L'étude des rapports de corrélation ("graphe des représentations" dans FactoMineR) est une aide à l'interprétation, surtout quand le nombre de modalités est grand et rend l'interprétation difficile.

Variables supplémentaires

- N'interviennent pas dans le calcul
- Enrichissement de l'interprétation
- Possibilité de tests
- Variable quantitative \Rightarrow Corrélacion avec les axes,
- Variable qualitative \Rightarrow utilisation des relations barycentriques.