ANALYSE DES CORRESPONDANCES MULTIPLES

Pierre-Louis Gonzalez

Technique de description de données qualitatives

n individus décrits par p variables qualitatives

χ_1 χ_2 χ_p

à m_1 m_2 ….. m_p modalités

L'A.C.M. décrit les relations deux à deux entre p variables qualitatives à travers une représentation des groupes d'individus correspondant aux diverses modalités.

Cette méthode est particulièrement bien adaptée à l'exploration d'enquêtes.

I - PRÉSENTATION FORMELLE

1. Données et notations

Chaque individu est décrit par les numéros des catégories où il est classé pour les p variables. Les données brutes se présentent sous forme d'un tableau à n lignes et p colonnes.

Les éléments de ce tableau sont des codes arbitraires sur lesquels aucune opération arithmétique n'est licite.

La forme mathématique utile pour les calculs est alors le **tableau disjonctif des indicatrices des p variables** obtenu en juxtaposant les p tableaux d'indicatrices de chaque variable χ_i

$n=5 \quad p=3 \quad m_1=3 \quad m_2=2 \quad m_3=3$

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 2 & 1 \\ 3 & 1 & 2 \end{bmatrix} \longrightarrow \left[\begin{array}{ccc|cc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{array}\right]$$

codage réduit

codage disjonctif

$$X = \left(X_1 \mid X_2 \mid X_3 \right)$$

- **La somme des éléments de chaque ligne de X est égale à p : nombre de variables**
- **La somme des éléments d'une colonne X donne l'effectif marginal de la catégorie correspondante.**

n

individus

3	1
1	5
2	3
3	1

2 variables qualitatives

Tri croisé

N

AFC

Tableau de contingence

0	0	1	
1	0	0	
0	1	0	
0	0	1	

ACM

Tableau disjonctif

2. Une propriété remarquable de l' AFC si p = 2

Lorsqu'il n'y a que deux variables qualitatives χ_1 et χ_2

à m_1 et m_2 modalités, **l'AFC formelle du tableau disjonctif** X est équivalente à **l'AFC du tableau de contingence N**

$$X = (X_1 \mid X_2) \qquad N = X_1' X_2$$

Cette propriété est à l'origine de la méthode étudiée ici

L'AFC formelle du tableau X revient à chercher les valeurs propres et les vecteurs propres du produit des deux tableaux de profils associés à X

Bien que fournissant des axes identiques à l'analyse des correspondances de N, les inerties associées et les parts d'inertie sont très différentes et ne peuvent être interprétées sans précaution.

Ainsi le passage (théorique) d'une analyse des correspondances sur le tableau disjonctif associé au tableau étudié au chapitre précédent conduit aux résultats suivants:

ACM		AFC	
$\mu_1 = 0{,}6$	8%	$\lambda_1 = 0{,}040$	83,7%
$\mu_2 = 0{,}54$	7%	$\lambda_2 = 0{,}005$	11,5%
$\mu_3 = 0{,}52$	7%	$\lambda_3 = 0{,}001$	2,4%

Les valeurs propres qui étaient très séparées dans l'AFC de N, ne le sont plus dans l'ACM du tableau disjonctif X

En AFC l'inertie est égale au Khi-deux associé au tableau de contingence divisé par le nombre d'individus

En ACM l'inertie est égale au nombre moyen de modalités diminué de 1

II PRINCIPE DE L'A.C.M.

1. Principe de l'A.C.M.

On réalise l'A.C.P. des profils-lignes avec la métrique du Khi-deux (comme en A.F.C.).

Les points représentatifs des catégories dans les graphiques factoriels doivent être considérés comme des barycentres.

2. Inertie totale du nuage de points

$$\text{Inertie} = \left[\frac{1}{p} \sum_{i=1}^{p} m_i \right] - 1$$

= nombre moyen de catégories diminué d'une unité.

En général, vu la nature des données, **les inerties portées par les premiers axes sont faibles.**

Les valeurs propres seront notées μ

3. Formules de transition

Coordonnée d'une modalité

z = vecteur à n composantes des coordonnées des individus sur un axe.

a = vecteur à $\sum_{i=1}^{p} m_i$ composantes des coordonnées des catégories des variables sur un axe

$$\underline{a} = \frac{1}{\sqrt{\mu}} D^{-1} X' \underline{Z}$$

À $\frac{1}{\sqrt{\mu}}$ près, la coordonnée d'une catégorie i est égale à la moyenne arithmétique des coordonnées des n_i individus de cette catégorie.

Exemple Axe 1

$$\text{Coord(TA2)} = \frac{1}{\sqrt{\mu_1}}\frac{1}{5}\,[\text{coord(Box)} + \text{coord (Cock)} + \text{coord(Dalm)} + \text{coord(EpBr)} + \text{coord (Labra)}]$$

Coordonnée d'un individu

$$z = \frac{1}{\sqrt{\mu}} \frac{1}{p} X \underline{a} = \frac{1}{\sqrt{\mu}} \frac{1}{p} \sum_{j=1}^{p} X_j \underline{a}_j$$

À $\frac{1}{\sqrt{\mu}}$ près, la cordonnée d'un individu est égale à la moyenne arithmétique des coordonnées des catégories auxquelles il appartient.

Exemple Axe 1

$$\text{Coord(Beauceron)} = \frac{1}{\sqrt{\mu_1}} \frac{1}{6} \quad [\text{coord(TA3)} + \text{coord (PO2)} + \ldots$$

$$+\ldots\ldots\ldots + \text{coord (AG2)}]$$

III PRATIQUE DE L'ANALYSE DES CORRESPONDANCES MULTIPLES

L'interprétation des résultats d'une A.C.M. se fait « grosso modo » comme en analyse des correspondances sur tableau de contingence et comme en en A.C.P.

Néanmoins :

On prendra garde ici au fait que **les pourcentages d'inertie n'ont qu'un intérêt restreint**

La sélection et l'interprétation des axes factoriels se feront essentiellement à l'aide :

- **des contributions des variables actives**
- **des valeurs tests associées aux variables supplémentaires.**

Rappelons une fois encore la signification des proximités entre points-colonnes sur un plan factoriel : il s'agit d'une **proximité, en projection, de points moyens de catégories représentant plusieurs individus.**

1. Les contributions

1.1 Contributions à un axe factoriel

Une catégorie j d'effectif n_j qui a pour coordonnée a_j sur un axe factoriel fournit une contribution égale à:

$$CTR(j) = \frac{\frac{n_j}{np}(a_j)^2}{\mu}$$

En A.C.M. les modalités d'une même variable χ_i ont des contributions qui peuvent être cumulées.

$$CTR(\chi_i) = \sum_{j=1}^{m_i} CTR\ (j)$$

1.2 Contributions à l'inertie totale

Une catégorie est d'autant plus éloignée du centre que son effectif est faible.

$$d^2(j,g) = \frac{n}{n_j} - 1$$

L'inertie totale apportée par cette modalité vaut :

$$I(j) = \frac{1}{p}\left(1 - \frac{n_j}{n}\right)$$

Elle décroît en fonction de l'effectif.

Il convient donc d'éviter de travailler avec des catégories d'effectif trop faible qui risquent de perturber les résultats de l'analyse (absence de robustesse).

L'inertie totale d'une variable vaut : $$\text{Inertie}(\chi_i) = \frac{m_i - 1}{p}$$

La contribution à l'inertie totale est d'autant plus importante que son nombre de modalités est élevé.

On recommande généralement pour cette raison d'éviter des disparités trop grandes entre les nombres de catégories des variables.

2) Règles d'interprétation

2.1 Nombre d'axes

On peut remarquer que si toutes les variables étaient indépendantes, toutes les valeurs propres seraient identiques et égales à 1/p.

Néanmoins le critère consistant à interpréter les axes d'inertie

> 1/p est en général peu utilisable en pratique.

Il est préférable d'utiliser la formule de Benzecri.

2.2 Formule de taux d'inertie corrigé (Benzecri 1979)

En ACM les taux d'inertie sont des mesures pessimistes de la qualité d'une représentation. On peut y remédier en utilisant la formule de Benzecri

$$\text{taux}(\mu) = \left(\frac{p}{p-1}\right)^2 \left(\mu - \frac{1}{p}\right)^2 \quad \text{pour} \quad \mu > \frac{1}{p}$$

p représente le nombre de variables actives

μ représente la valeur propre issue de l'ACM.

2.3 Interprétation des axes

Interprétation sommaire : On peut rechercher les variables dont la contribution cumulée est supérieure à 1/p (parfois exprimé en %). Cela ne permet pas de donner une «signification» des axes

Interprétation détaillée : On cherche les modalités dont la contribution est supérieure au poids.

CTR(j) > poids

D'où:

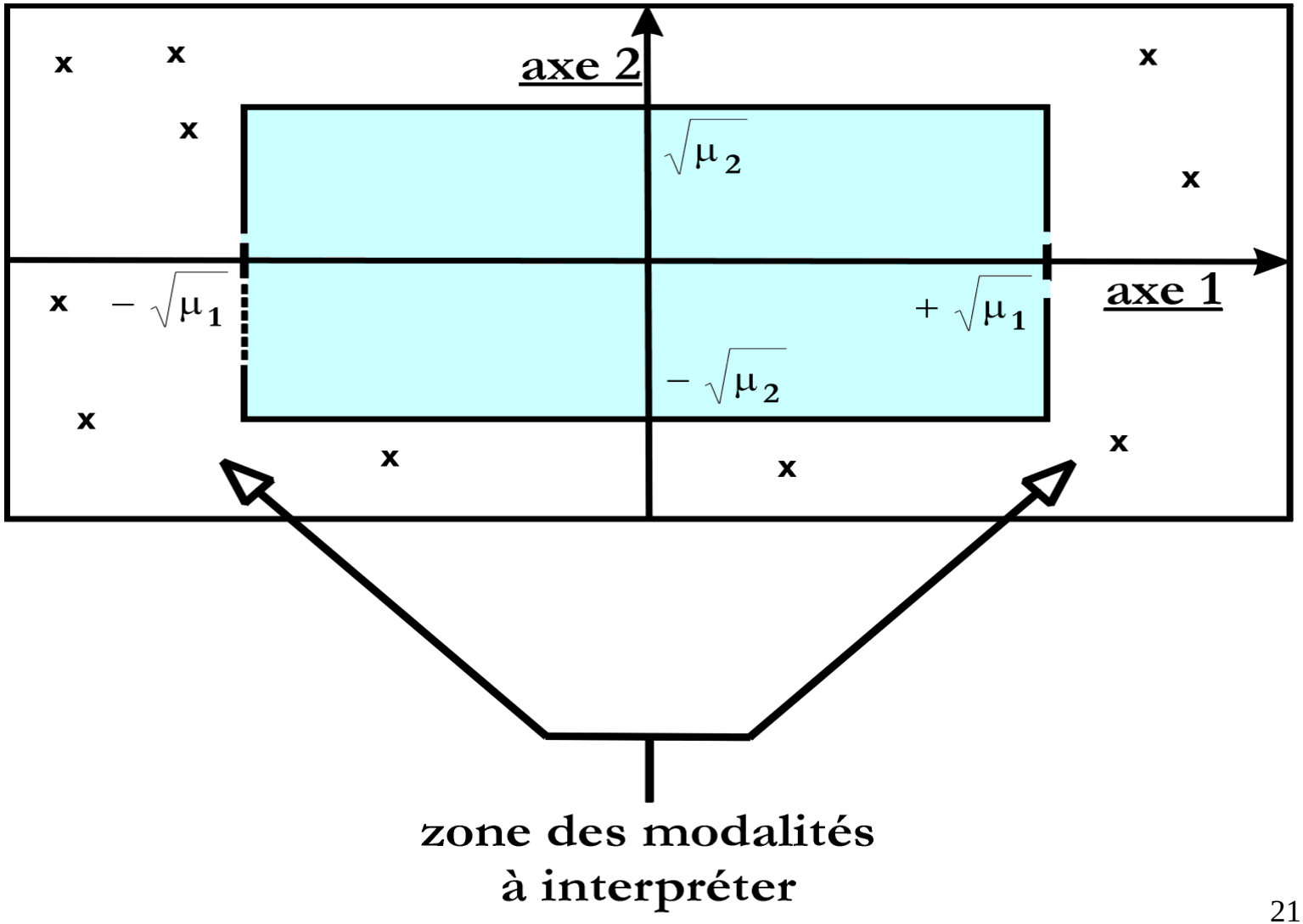
$$\frac{\frac{n_j}{np}(a_j)^2}{\mu} > \frac{n_j}{np}$$

On en déduit

$$|a_j| > \sqrt{\mu}$$

Cette expression montre que l'on interprète essentiellement les modalités éloignées de l'origine.

Exemple: Plan 1-2



3. L'USAGE DE VARIABLES SUPPLÉMENTAIRES

3.1 Les deux groupes de variables

- Les **variables actives** sont celles qui déterminent les axes

- Les **variables supplémentaires** ne participent pas au calcul des valeurs propres et vecteurs propres.

Elles peuvent être représentées sur les plans factoriels selon le principe barycentrique pour les variables qualitatives : **chaque catégorie est le point moyen d'un groupe d'individus.**

Le choix des variables supplémentaires obéit à diverses préoccupations.

- **Réduire** la taille du tableau à diagonaliser
- **Conforter** l'interprétation des axes par des variables n'ayant pas servi à les déterminer
- Enfin, il est possible de mettre en variables supplémentaires des **variables quantitatives** qui ne pourraient être actives (à moins de les rendre qualitatives par découpage en classes).

Dans ce dernier cas, la plupart des logiciels se bornent à indiquer leurs corrélations avec les composantes factorielles.

3.2 Les tests associés

Pour pouvoir interpréter la liaison entre un axe factoriel et une **variable supplémentaire**, il faut pouvoir juger de **« l'intensité »** de cette liaison.

- Si la variable supplémentaire est **numérique** $\underline{X}$, on vérifiera si la **corrélation** $r(\underline{Z},\underline{X})$ dépasse un seuil critique

  par exemple: $\dfrac{2}{\sqrt{n+2}}$

- Si la variable supplémentaire est **qualitative** à m modalités, on testera la valeur du **rapport de corrélation** par un test de Fisher-Snedecor (analyse de la variance à un facteur).

Valeur-test

Lebart et Morineau ont introduit cette notion pour chaque modalité d'une variable qualitative, afin de juger si le point représentatif d'une modalité est significativement différent de la moyenne générale.

C'est le cas si: $|\text{V.T.}| > 2$

Principe de calcul de la valeur-test

- Soit a_i la coordonnée d'une modalité d'une variable supplémentaire, d'effectif n_i , sur un axe d'inertie égale à μ

- Si les n_i individus de cette catégorie étaient pris au hasard parmi les n individus de l'échantillon (sans remise), la moyenne des coordonnées des n_i individus concernés serait une variable aléatoire centrée, de variance égale à : $\frac{\mu}{n_i} \times \frac{n - n_i}{n - 1}$

- Avec les conventions habituelles de la représentation simultanée :

$$a_i = \frac{1}{\sqrt{\mu}}$$ (moyenne des coordonnées)

La quantité $\dfrac{(\text{moyenne des coordonnées}) - o}{\text{écart} - \text{type}}$

$$= VT = a_i \sqrt{n_i} \sqrt{\frac{n-1}{n-n_i}}$$

mesure donc en **nombre d'écart-type l'éloignement du point représentatif d'une modalité par rapport à l'origine.**

Si n_i assez grand, on comparera cette valeur à celle d'une variable de Laplace Gauss LG (0,1) en raison du théorème central-limite.

On considère donc comme « significatives » d'un axe les catégories qui ont une valeur-test supérieure en valeur absolue à 2 (au seuil 5 %).

Remarques :

Cette pratique permet un dépouillement rapide des résultats :

- En principe, le calcul des valeurs-tests n'est légitime que pour des variables supplémentaires n'ayant pas servi à la détermination des axes.

- Leur utilisation pour des variables actives ne doit être considérée qu'à titre indicatif.

4) À propos du découpage en classes

La pratique qui consiste à découper en classes des variables numériques, donc à les rendre qualitatives, pour ensuite effectuer une analyse des correspondances multiples se justifie par le fait qu'il s'agit d'une **analyse non linéaire des données.**

Sous réserve d'avoir suffisamment d'observations par classe on peut ainsi utiliser des **liaisons non linéaires** entre variables qui ne seraient pas apparues en A.C.P. ordinaire où l'on travaille avec la matrice R des corrélations linéaires.

IV. LA CLASSIFICATION DE DONNÉES QUALITATIVES

Les n individus à classer sont décrits par des variables qualitatives

1. Données de présence - absence

On utilise un des indices de dissimilarité déduit des indices de similarité proposés qui combinent de diverses manières les quatre nombres suivants associés à un couple d'individus.

a = nombre de caractéristiques communes

b = nombre de caractéristiques possédées par i et pas par j

c = nombre de caractéristiques possédées par j et pas par i

d = nombre de caractéristiques que ne possèdent ni i, ni j.

Les indices compris entre 0 et 1 sont aisément transformables en dissimilarité par complémentation à 1.

Jaccard $\frac{a}{a+b+c}$

Dice ou Czekanowski $\frac{2a}{2a+b+c}$

Ochiaï $\frac{a}{\sqrt{(a+b)(a+c)}}$

Russel et Rao $\frac{a}{a+b+c+d}$

Rogers et Tanimoto $\frac{a+d}{a+d+2(b+c)}$

2. Individus décrits par des variables qualitatives à m_1 m_2 ... m_p modalités

On utilise la représentation disjonctive complète et la distance du Khi-deux entre lignes du tableau.

$$d^2_{\chi^2}(i,i') = \sum_j \frac{np}{n_{.j}} \left(\frac{x_{ij} - x_{i'j}}{p} \right)^2$$

Elle traduit le fait que deux individus ayant en commun une modalité rare sont plus proches que deux individus ayant en commun une modalité fréquente.

On utilise alors la **méthode de Ward** (puisque la distance du Khi-deux est euclidienne) sur le tableau des distances.

Autre solution :

Classification hiérarchique sur le tableau des coordonnées factorielles des n individus après A.C.M. de X.

Les deux approches sont équivalentes si on utilise tous les facteurs de l'A.C.M. soit $\sum m_i - p$

en conservant la normalisation de chaque axe à $\sqrt{\mu}$

Cette approche permet toutefois de ne sélectionner que certains facteurs et donc de ne pas prendre en compte une information résiduelle considérée comme du bruit.