

# Variables qualitatives : analyse des correspondances

Jean-Marc Lasgouttes

<http://ana-donnees.lasgouttes.net/>

## L'analyse factorielle des correspondances

**But** On cherche à décrire la liaison entre deux variables qualitatives.

**Exemple** on peut regarder la répartition de la couleur des yeux en fonction de la couleur des cheveux.

**Différence avec l'ACP** l'ACP se fait dans un cadre différent ; les variables sont quantitatives et donc

- il est possible de faire des opérations mathématiques sur les valeurs des variables ;
- par contre, il n'est en général pas possible de compter les individus qui ont une caractéristique donnée (taille=1, 83m)

**Pourquoi deux variables ?** le cas de plus de deux variables est l'analyse de correspondance multiples, traité plus tard dans le cours.

## Partie I. Les données qualitatives

### Variables qualitatives

Soit  $\mathcal{X}$  une variable qualitative. On dispose d'un échantillon de  $n$  individus sur lesquels la variable est mesurée.

**Modalités (ou catégories)** les valeurs que peut prendre une variable qualitative ; si la variable a  $m$  modalités (valeurs possibles), on note  $x_i$ ,  $1 \leq i \leq m$ , ces modalités, ou plus simplement  $i$ .

**Effectif** le nombre d'occurrence de la modalité  $i$  dans l'échantillon ; on le note  $n_i$ , et on a  $\sum_{i=1}^m n_i = n$ .

**Profil** c'est l'ensemble des valeurs  $n_i/n$  ; la somme du profil sur les modalités est 1.

### Tableau de contingence

Soient  $\mathcal{X}_1$  et  $\mathcal{X}_2$  deux variables qualitatives à  $m_1$  et  $m_2$  modalités respectivement décrivant un ensemble de  $n$  individus.

**Définition** le tableau de contingence est une matrice à  $m_1$  lignes et  $m_2$  colonnes renfermant les effectifs  $n_{ij}$  d'individus

tels que  $\mathcal{X}_1 = i$  et  $\mathcal{X}_2 = j$ .

$$\mathbf{N} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & & \\ & & \ddots & \\ \vdots & & & n_{ij} & \vdots \\ & & & & n_{m_1 m_2} \end{pmatrix}$$

La constitution de ce tableau est aussi appelé un « tri croisé ».

### Marges et profils

**Marge en ligne** c'est la somme  $n_{i.} = \sum_{j=1}^{m_2} n_{ij}$ , c'est-à-dire l'effectif total de la modalité  $i$  de  $\mathcal{X}_1$ .

On définit aussi le profil marginal des lignes  $n_{i.}/n$ .

**Marge en colonne** c'est la somme  $n_{.j} = \sum_{i=1}^{m_1} n_{ij}$ , c'est-à-dire l'effectif total de la modalité  $j$  de  $\mathcal{X}_2$ .

On définit aussi le profil marginal des colonnes  $n_{.j}/n$ .

**Deux lectures possibles** selon la variable que l'on privilégie, on peut définir

- le tableau des *profils-lignes*  $n_{ij}/n_{i.}$ , qui représente la fréquence de la modalité  $j$  conditionnellement à  $\mathcal{X}_1 = i$  ; la somme de chaque ligne est ramenée à 100%.
- le tableau des *profils-colonnes*  $n_{ij}/n_{.j}$ , qui représente la fréquence de la modalité  $i$  conditionnellement à  $\mathcal{X} = j$  ; la somme de chaque colonne est ramenée à 100%.

### Propriétés des profils

**Moyenne** la moyenne des profils-lignes (avec poids correspondant aux profils marginaux des lignes) est le profil marginal des colonnes :

$$\sum_{i=1}^{m_1} \frac{n_{i.}}{n} \times \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n},$$

et de même pour les colonnes  $\sum_{j=1}^{m_2} \frac{n_{.j}}{n} \times \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$ .

**Indépendance empirique** lorsque tous les profils lignes sont identiques, il y a indépendance entre  $\mathcal{X}_1$  et  $\mathcal{X}_2$ , puisque la connaissance de  $\mathcal{X}_1$  ne change pas la répartition de  $\mathcal{X}_2$ . On a pour tout  $j$

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \cdots = \frac{n_{m_1 j}}{n_{m_1.}} = \frac{n_{1j} + \cdots + n_{m_1 j}}{n_{1.} + \cdots + n_{m_1.}} = \frac{n_{.j}}{n}$$

et donc  $n_{ij} = \frac{n_{i.} n_{.j}}{n}$ .

# Partie II. Géométrie de nuages de profils

## Analyse des correspondances de deux variables : les données

**Effectifs** on a un tableau de contingence  $\mathbf{N}$  à  $m_1$  lignes et  $m_2$  colonnes résultant du croisement de deux variables qualitatives  $\mathcal{X}_1$  et  $\mathcal{X}_2$  à  $m_1$  et  $m_2$  modalités respectivement. On note  $\mathbf{D}_1$  et  $\mathbf{D}_2$  les matrices diagonales des effectifs marginaux

$$\mathbf{D}_1 = \begin{pmatrix} n_{1.} & & & 0 \\ & n_{2.} & & \\ & & \ddots & \\ 0 & & & n_{m_1.} \end{pmatrix}$$

$$\mathbf{D}_2 = \begin{pmatrix} n_{.1} & & & 0 \\ & n_{.2} & & \\ & & \ddots & \\ 0 & & & n_{.m_2} \end{pmatrix}$$

**Profils** le tableau des profils des lignes  $n_{ij}/n_{i.}$  est donné par  $\mathbf{D}_1^{-1}\mathbf{N}$  et celui des profils des colonnes  $n_{ij}/n_{.j}$  par  $\mathbf{N}\mathbf{D}_2^{-1}$ .

## Représentation géométrique des profils

**Nuage de points** les profils-lignes forment un nuage de  $m_1$  points de  $\mathbb{R}^{m_2}$ . Chaque point est affecté d'un poids égal à sa fréquence marginale  $n_{i.}/n$ , et la matrice des poids est donc  $\frac{1}{n}\mathbf{D}_1$ .

**Centre de gravité** c'est le profil marginal des colonnes car

$$\mathbf{g}_\ell = \frac{1}{n}(\mathbf{D}_1^{-1}\mathbf{N})'\mathbf{D}_1\mathbf{1}_{m_1} = \left(\frac{n_{.1}}{n}, \dots, \frac{n_{.m_2}}{n}\right)'$$

**Profils-colonnes** les lignes du tableau  $\mathbf{D}_2^{-1}\mathbf{N}'$  forment un nuage de  $m_2$  points de  $\mathbb{R}^{m_1}$ , avec matrice de poids  $\frac{1}{n}\mathbf{D}_2$  et centre de gravité

$$\mathbf{g}_c = \left(\frac{n_{1.}}{n}, \dots, \frac{n_{m_1.}}{n}\right)'$$

## Comment étudier ces données

**Cas indépendant** en cas d'indépendance empirique, on a

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \text{ et } \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$$

Les deux nuages sont donc réduits à leurs centres de gravité.

**Dimension des nuages** comme les profils somment à 1, les  $m_1$  profils-lignes sont situés dans le sous-espace  $W_1$  de dimension  $m_2 - 1$  défini par  $\sum_{j=1}^{m_2} x_j = 1$  et  $x_j \geq 0$ .

**ACP** l'étude de la forme des nuages au moyen de l'analyse en composantes principales permettra de rendre compte de la structure des écarts à l'indépendance.

# Partie III. L'ACP : une ACP sur un nuage de profils

## La métrique du $\chi^2$

**Profils-lignes** la norme du  $\chi^2$  de la différence de profils-lignes  $\mathbf{e}_i - \mathbf{e}_{i'}$  est définie par

$$\|\mathbf{e}_i - \mathbf{e}_{i'}\|_{\chi^2} = \sum_{j=1}^{m_2} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2,$$

ce qui revient à utiliser la métrique diagonale  $n\mathbf{D}_2^{-1}$ .

**Inertie** l'inertie totale du nuage des profils-lignes par rapport à  $\mathbf{g}_\ell$  est

$$\begin{aligned} I_{\mathbf{g}_\ell} &= \sum_{i=1}^{m_1} \frac{n_{i.}}{n} \|\mathbf{e}_i - \mathbf{g}_\ell\|_{\chi^2}^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{n_{i.}}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}\right)^2 \\ &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{1}{n_{i.}n_{.j}} \left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2 \end{aligned}$$

Cette inertie mesure l'écart à l'indépendance.

## Pourquoi la métrique du $\chi^2$ ?

**Pondération** la pondération  $n/n_{.j}$  permet de donner des importances comparables aux différentes « variables ».

**Équivalence distributionnelle** si deux colonnes  $j$  et  $j'$  de  $\mathbf{N}$  ont le même profil, il est logique de les regrouper en une seule d'effectif  $n_{ij} + n_{ij'}$ ; on a alors quand  $n_{ij}/n_{.j} = n_{ij'}/n_{.j'}$

$$\begin{aligned} \frac{n}{n_{.j}} \left[\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}\right]^2 + \frac{n}{n_{.j'}} \left[\frac{n_{ij'}}{n_{i.}} - \frac{n_{.j'}}{n}\right]^2 \\ = \frac{n}{n_{.j} + n_{.j'}} \left[\frac{n_{ij} + n_{ij'}}{n_{i.}} - \frac{n_{.j} + n_{.j'}}{n}\right]^2 \end{aligned}$$

La distance entre les profils-ligne est donc inchangée.

**Profils-colonnes** on définit la distance entre deux profils-colonnes  $\mathbf{e}_j$  et  $\mathbf{e}_{j'}$  comme

$$\|\mathbf{e}_j - \mathbf{e}_{j'}\|_{\chi^2} = \sum_{i=1}^{m_1} \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right)^2,$$

soit une métrique de matrice  $n\mathbf{D}_1^{-1}$ . On remarque que  $I_{\mathbf{g}_\ell} = I_{\mathbf{g}_c}$ .

## ACP des deux nuages de profils

**Comment ?** deux possibilités en dualité exacte

	données	métrique	poids
Profils-lignes	$\mathbf{X} = \mathbf{D}_1^{-1}\mathbf{N}$	$\mathbf{M} = n\mathbf{D}_2^{-1}$	$\mathbf{D} = \frac{\mathbf{D}_1}{n}$
Profils-colonnes	$\mathbf{X} = \mathbf{D}_2^{-1}\mathbf{N}'$	$\mathbf{M} = n\mathbf{D}_1^{-1}$	$\mathbf{D} = \frac{\mathbf{D}_2}{n}$

**Centrage** Il n'est pas très utile, puisque la différence des inerties est 1 :

$$I_g = \frac{d^2}{n} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - \frac{n_i \cdot n_j}{n})^2}{n_i \cdot n_j}$$

$$= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{n_{ij}^2}{n_i \cdot n_j} - 1 = I_0 - 1,$$

**Approche** On effectue donc une ACP non centrée  
 — on remplace la variance par  $\mathbf{X}'\mathbf{D}\mathbf{X}$ .  
 — on élimine la valeur propre 1 associée à l'axe principal  $\mathbf{g}$  (admis)

### Calcul de l'ACP (profils-lignes)

**Facteurs principaux** ils sont vecteurs propres de

$$\mathbf{M}\mathbf{X}'\mathbf{D}\mathbf{X} = (n\mathbf{D}_2^{-1})(\mathbf{D}_1^{-1}\mathbf{N})' \frac{\mathbf{D}_1}{n} (\mathbf{D}_1^{-1}\mathbf{N}) = \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}.$$

et on a donc pour chaque axe principal  $k$

$$\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k = \lambda_k \mathbf{u}_k$$

**Composantes principales** la composante principale associée au facteur  $\mathbf{u}_k$  est  $\mathbf{a}_k = \mathbf{X}\mathbf{u}_k = \mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k$  ; elle est vecteur propre de la matrice  $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$  car

$$\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}_k = \mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k$$

$$= \lambda_k \mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k = \lambda_k \mathbf{a}_k$$

**Analyse des profils-colonnes** on échange les indices 1 et 2 et on transpose  $\mathbf{N}$ .

### Comparaison lignes-colonnes

	ACP profils-lignes	ACP profils-colonnes
Facteurs principaux	Vecteurs propres de $\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}$	Vecteurs propres de $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$
Composantes principales	Vecteurs propres de $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$ normalisés par $\text{var } \mathbf{a}_k = \mathbf{a}_k' \frac{\mathbf{D}_1}{n} \mathbf{a}_k = \lambda_k$	Vecteurs propres de $\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}$ normalisés par $\text{var } \mathbf{b}_k = \mathbf{b}_k' \frac{\mathbf{D}_2}{n} \mathbf{b}_k = \lambda_k$

**Comparaison** les deux analyses conduisent aux mêmes valeurs propres et les facteurs principaux de l'une sont les composantes principales de l'autre (à un facteur près).

## Partie IV. Aspects pratiques

### Interprétation des résultats

**Coordonnées des points** Les coordonnées des points-lignes et points-colonnes s'obtiennent en cherchant les vecteurs propres des produits des deux tableaux de profils. Ce sont les grandeurs principales à obtenir.

**Projection des nuages** il est possible de projeter les deux nuages de points sur le même représentations. On justifiera plus tard le sens de cette représentation et son interprétation.

**Cercle des corrélations** il n'a aucun intérêt ici, puisque les véritables variables sont qualitatives.

**(non) effet de taille** comme les composantes variables sont centrées ( $\sum_{i=1}^{m_1} n_i \cdot a_{ik} = \sum_{j=1}^{m_2} n_j \cdot b_{jk} = 0$ ), ont sait que les coordonnées des  $\mathbf{a}_k$  et  $\mathbf{b}_k$  ne peuvent être toutes de même signe ; il n'y a donc jamais d'effet de « taille ».

### Contributions à l'inertie

**Contribution des profils-lignes** On sait que  $\lambda_k = \sum_{i=1}^{m_1} \frac{n_i}{n} (a_{ik})^2$ , où  $a_{ik}$  est la coordonnée du profil-ligne  $i$  sur la  $k$ -ième composante principale de l'ACP sur les profils-lignes. On définit donc la contribution de la modalité  $i$  à l'axe principal  $k$  comme

$$\frac{n_i}{n} \cdot \frac{(a_{ik})^2}{\lambda_k}.$$

On considérera les modalités ayant l'influence la plus importante (typiquement  $> \alpha n_i/n$ ,  $\alpha = 2$  ou  $3$ ) comme constitutives des axes ; on regardera aussi le signe de la coordonnée.

Il n'y a pas ici de modalités sur-représentées, puisqu'on ne peut pas les retirer.

**Contribution des profils-colonnes** pour les mêmes raisons, la contribution de la modalité  $j$  de  $\mathcal{X}_2$  à l'axe  $k$  est

$$\frac{n_j}{n} \cdot \frac{(b_{jk})^2}{\lambda_k}.$$

### Qualité de la représentation

**Profils-lignes** l'AFC est une ACP, et on peut donc mesurer la qualité de la représentation de la modalité  $i$  (son profil-ligne) par un sous-espace factoriel. La qualité (le  $\cos^2$  de l'angle entre le point et sa projection) s'écrit encore, pour le plan formé des  $q$  premiers axes :

$$\frac{\sum_{k=1}^q (a_{ik})^2}{\sum_{k=1}^{m_2} (a_{ik})^2}.$$

Comme pour l'ACP,  $> 0.8$  signifie « bien représenté » et  $< 0.5$  veut dire « mal représenté ». Les valeurs sont souvent données en pourcents.

**Profils-colonne** Le principe est le même, mais la formule devient, pour la modalité  $j$  :

$$\frac{\sum_{k=1}^q (b_{jk})^2}{\sum_{k=1}^{m_1} (b_{jk})^2}.$$