Partie II

Analyse Factorielle des Correspondances

Introduction

<u>But</u> : Description de la liaison "correspondance" entre deux variables.

<u>Exemple</u> :
i) La répartition des couleur des yeux en fonction de la couleur des cheveux
ii) Abondance d'une espéce j dans un milieu i
iii) Nombre de fois que le personnage i a utilisé le mot j

<u>AFC vs ACP</u> : l'ACP se fait dans un cadre des variables numériques "quantitatives". Donc il est possible de faire des opération mathématiques et étudier la corrélation entre les variables.

Données qualitatives

On pose $[n] := \{1, \cdots, n\}$.
On dispose d'un échantillon de n individus sur lesquels une variable qualitative X est mesurée.

<u>Modalités</u> : ce sont les valeurs que une variable qualitative peut prendre.

<u>Effectif</u> : Supposons que X a m modalités qu'on note par $i \in [m]$. L'effectif est le nombre d'occurence de la modalité i et sera noté par n_i et nous avons $\sum_{i=1}^{m} n_i = n$

<u>Profil</u> : C'est l'ensemble des valeurs $\frac{n_i}{n}$, $i \in [m]$. La somme des profiles sur une modalité est egale à 1.

Tableau de Contingence

A la différence de l'ACP les données de l'AFC sont organisées en tableaux appellés tableaux de contingence (ou aussi tableau de dépendance ou croisé).

Definition

Un tableau de contingence est un tableau d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur la même population de n individus.

Les données

- Soient X_1 et X_2 deux variables qualitatives à m_1 et m_2 modalités respectivement décrivant les n individus.

$$
\begin{array}{cc|ccccc|}
 & & & & X_2 & & \\
 & & 1 & \cdots & j & \cdots & m_2 \\
\hline
 & 1 & n_{11} & \cdots & n_{1j} & \cdots & n_{1m_2} \\
 & \vdots & \vdots & \ddots & \vdots & & \vdots \\
X_1 & i & n_{i1} & \cdots & n_{ij} & \cdots & n_{im_2} \\
 & \vdots & \vdots & & \vdots & \ddots & \vdots \\
 & m_1 & n_{m_1 1} & \cdots & n_{m_1 j} & \cdots & n_{m_1 m_2} \\
\hline
\end{array}
$$

- n_{ij} est l'efféctif des individus ayant la modalité i et j.

Les données

Davantage que le tableau de contingence, c'est le tableau des fréquences relatives qu'on va considr̀er dans ce chapitre. Les féquences sont données par : $f_{ij} = \frac{n_{ij}}{n}$.

Marge en ligne : c'est la somme $f_{i.} = \sum_{j=1}^{m_2} f_{ij}$.

Marge en colonne : c'est la somme $f_{.j} = \sum_{i=1}^{m_1} f_{ij}$.

- Nous avons ainsi :

$$\sum_{i=1}^{m_1} f_{i.} = \sum_{j=1}^{m_2} f_{.j} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} f_{ij} = 1 \qquad (1)$$

Les données

				X_2			
		1	$\cdots$	j	$\cdots$	m_2	marge
	1						
	$\vdots$			$\vdots$			$\vdots$
X_1	i		$\cdots$	f_{ij}	$\cdots$		$f_{i\bullet}$
	$\vdots$			$\vdots$			$\vdots$
	m_1						
marge			$\cdots$	$f_{\bullet j}$	$\cdots$		$\Sigma = 1$

Figure – Tableau des fréquences relatives

Liaison entre les variables

Comme l'AFC considère un tableau de contingence ou celui des fréquences, relatives, nous ne pouvons plus définir la liaison entre les deux variable par les coefficients de corrèlation comme pour l'ACP.

Definition

Il y a indépendance entre les deux variables considérées si :

$$f_{ij} = f_{i.}f_{.j} \quad \forall i \in [m_1], \forall j \in [m_2]. \tag{2}$$

Nous disons qu'il y a liaison entre les deux variables, ou que ces variables sont liées si elles ne sont pas indépendantes.

Nous dirons alors que les deux variables :

- s'attirent si $f_{ij} > f_{i.}f_{.j}$,
- se répulsent si $f_{ij} < f_{i.}f_{.j}$.

Liaison entre les variables

Sous l'hypothèse d'indépendance nous avons deux lectures possibles :

Profiles-lignes : On considère le tableau comme un ensemble des lignes : $\frac{f_{ij}}{f_{i.}} = f_{.j} \quad \forall i \in [m_1], j \in [m_2]$.
Le terme $f_{.j}$ s'interprète comme le pourcentage de population totale possédant la modalité j, et le terme $\frac{f_{ij}}{f_{i.}}$ représente ce même pourcentage dans la sous population possédant la modalité i.

Profiles-colonnes : On considère le tableau comme un ensemble des colonnes : $\frac{f_{ij}}{f_{.j}} = f_{i.} \quad \forall i \in [m_1], j \in [m_2]$.
Par raison de symétrie on peut de la même façon expliquer les terme dans cette équation.

Liaison entre les variables

$$
\begin{array}{cc|ccccc|c|}
 & & & & X_2 & & & \\
 & & 1 & \cdots & j & \cdots & m_2 & \\
\hline
 & 1 & & & & & & \\
 & \vdots & & & \vdots & & & \\
X_1 & i & \frac{f_{i1}}{f_{i.}} & \cdots & \frac{f_{ij}}{f_{i.}} & \cdots & \frac{f_{im_2}}{f_{i.}} & 1 \\
 & \vdots & & & \vdots & & & \\
 & m_1 & & & & & & \\
\hline
\end{array}
$$

Figure – Tableau des Profils-lignes

Liaison entre les variables

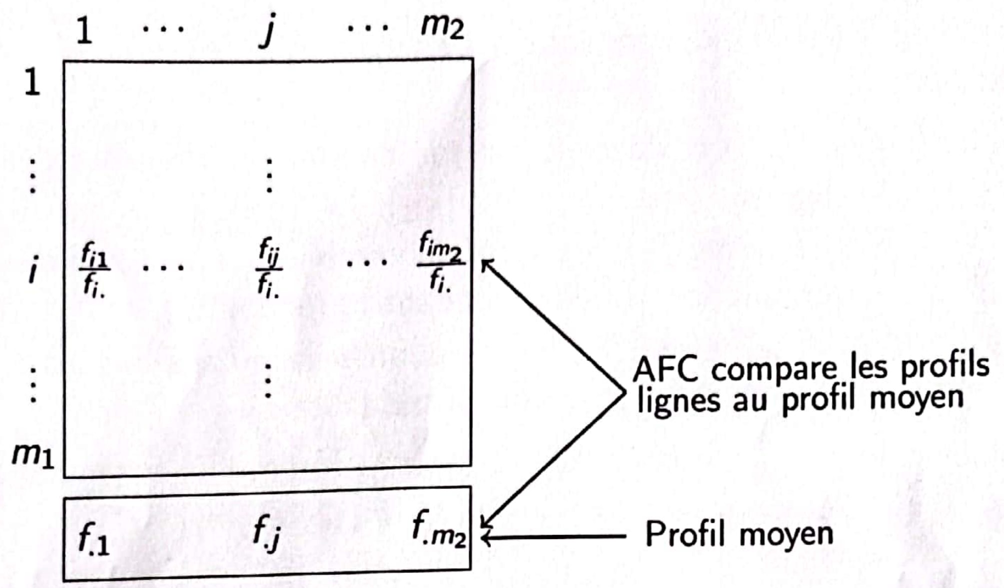

Figure – Approche de l'écart à l'indépendance

Exemple

- Prenons l'exemple de répartition de 592 femmes selon la couleur des yeux et des cheveux (proposé par Cohen en 1980).

	Brun	châtain	roux	blond	Total
maron	68	119	26	7	220
noisette	15	54	14	10	93
vert	5	29	14	16	64
bleu	20	84	17	94	215
Total	108	286	71	127	592

- Le tableau de contingence ci-dessus donne le nombre de femmes possédant à la fois une des quatre modalités de la couleur des cheveux et une des quatre modalités de la couleur des yeux.

Exemple

- Le tableau des fréquences correspondant permet de ne plus tenir compte du nombre de femmes total.

	Brun	châtain	roux	blond	Total
maron	11,4	20,1	4,3	1,1	37,1
noisette	2,5	9,1	2,3	1,6	15,7
vert	0,8	4,8	2,3	2,7	10,8
bleu	3,3	14,1	2,8	15,8	36,3
Total	18,2	48,3	11,9	21,4	100

- Ainsi on peut se demander s'il y a indépendance entre la couleur des cheveux et la couleur des yeux ou encore quelles sont les associations entre les couleurs.

Exemple

- Le tableau ci-desous est le tableau de profiles-lignes exprimé en pourcentage arrondis.

	Brun	châtain	roux	blond	Profil moyen
maron	30,9	54	11,8	3,1	100
noisette	16,1	58	15	10,7	100
vert	7,8	45,3	21,8	25	100
bleu	9,3	39	7,9	43,7	100
Profil moyen	18,2	48,3	11,9	21,4	100

- La fraction $\frac{f_{ij}}{f_{i.}}$ représente la fréquence pour une femme d'avoir les cheveux de couleur j sachant qu'elle a les yeux d'une couleur i.

La distance du χ^2

- On considère l'ensemble des lignes :

$$L_i = \left(\frac{f_{i1}}{f_{i.}}, \frac{f_{i2}}{f_{i.}}, \cdots, \frac{f_{ij}}{f_{i.}}, \cdots, \frac{f_{im_2}}{f_{i.}} \right), \forall i \in [m_1]$$

avec la pondèration $f_{i.}$.

- Pour deux individus quelconques i et i' on considère :

$$d^2_{\chi^2}(L_i, L_{i'}) = \sum_{j=1}^{m_2} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

- C'est une modification de la distance euclidienne, qui tient compte des écarts entre deux probabilités d'avoir un caractère chez deux individus en prenant en considèration la probabilité que l'individu ait tout les caractères étudiés.

Interprétation geométrique des Profils

- L'ensemble $\{P_i, i \in [m_1]\}$ avec :
$P_i = \left(\frac{f_{i1}}{f_{i.}\sqrt{f_{.1}}}, \frac{f_{i2}}{f_{i.}\sqrt{f_{.2}}}, \cdots, \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}, \cdots, \frac{f_{im_2}}{f_{i.}\sqrt{f_{.m_2}}} \right)$
peut être considéré comme un nuage de points dans $\mathbb{R}^{m_2}$
- Plus généralement, la distance du χ^2 est la même que la distance euclidienne entre les points P_i et $P_{i'}$ dans $\mathbb{R}^{m_2}$:
$d^2(P_i, P_{i'}) = \sum_{j=1}^{m_2} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}} \right)^2$

Projection du nuage sur une axe

- Le nuage $\{P_i, i\}$ sera projeter orthogonalement sur une axe de vecteur unitaire u de façon que la perte d'information soit minimale.
- Soit V la matrice de variance-covariance du nuage. Comme en ACP on cherche à maximiser $u'Vu$ sous la contraint $u'u = 1$.
- Ce qui revient à trouver la valeur propre maximale de V.
- On pose $\lambda_{ij} = \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}$ alors la matrice de variance-covariance est $V = (\sigma_{ij})_{(i,j)\in[m_1]\times[m_2]}$ avec $\sigma_{jk} = \sum_{i=1}^{m_1} f_{i.}\left(\lambda_{ij} - \sqrt{f_{.j}}\right)\left(\lambda_{ik} - \sqrt{f_{.k}}\right)$

Projection du nuage sur une axe

- la moyenne arithmétique pondrés est : $\bar{x}_j = \sqrt{f_{.j}}$
- On a :

$$
\begin{aligned}
\sigma_{jk} &= \sum_{i=1}^{m_1} f_{i.}\left(\lambda_{ij} - \sqrt{f_{.j}}\right)\left(\lambda_{ik} - \sqrt{f_{.k}}\right)\\
&= \sum_{i=1}^{m_1} f_{i.}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}}\right)\left(\frac{f_{ik}}{f_{i.}\sqrt{f_{.k}}} - \sqrt{f_{.k}}\right)\\
&= \sum_{i=1}^{m_1} \left(\frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.k}}}\right)\left(\frac{f_{ik} - f_{i.}f_{.k}}{\sqrt{f_{i.}f_{.k}}}\right)
\end{aligned}
$$

- Soit $r_{ij} := \left(\frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.k}}}\right)$ pour $(i,j) \in [m_1] \times [m_2]$ et $R = (r_{ij})_{(i,j)\in[m_1]\times[m_2]}$.

Exercice 3.
Au cours d'une enquête sur un échantillon de taille 60, on a obtenu le tableau de contingence suivant:

	M_1	M_2
M_1	10	10
M_2	5	15
M_3	15	5

1) Donner le tableau des probabilité relatives et le tableau marginal

2) Dans l'espace $\mathbb{R}^2$, on considére un nuage $\mathcal{B}(I)$ des points P_i, avec $i \in I$.

   a) Donner les points P_i du nuage $\mathcal{B}(I)$.

   b) Calculer la distance χ^2 entre les les differents points de $\mathcal{B}(I)$.

3) a) Déterminer la matrice des variance co-variance W ou la matrice R.

   b) Déterminer les valeurs propres de la matrice W.

   c) En deduire la variabilité totale du nuage $\mathcal{B}(I)$

4) On projette, maintenant, le nuage $\mathcal{B}(I)$ orthogonalement sur un axe, et on note $C(I)$ le nuage projeté. Donner la variabilité totale de nuage projeté $C(I)$.

5) Calculer la variabilité expliquée par la projection du nuage $\mathcal{B}(I)$.