# CHAPITRE 10

# ANALYSE FACTORIELLE DES CORRESPONDANCES

# Plan

1. Les données
2. L'AFC est une AC particulière
3. Les représentations graphiques
4. Les aides à l'interprétation
5. Une application de l'AFC

# 1. Les données

Les données peuvent se présenter sous 2 formes équivalentes :

- sous la forme de 2 tableaux disjonctifs complets $X_1$ et $X_2$. Chacun de ces 2 tableaux décrivent les modalités d'une variable qualitative
- sous la forme du tableau de contingence $(X_1)'X_2$ (ou bien $(X_2)'X_1$).

• Le nombre de colonnes de $X_1$, c'-à-d le nombre de modalités de la première variable qualitative est $m_1$.

• Le nombre de colonnes de $X_2$, c'-à-d le nombre de modalités de la seconde variable qualitative est $m_2$.

• $X_1$ et $X_2$ possèdent $n$ lignes, chaque ligne correspondant à un individu.

• Pour ne pas alourdir les notations, nous supposerons que $m_2 > m_1$.

• Le tableau de contingence est évidemment plus maniable en pratique que les tableaux disjonctifs complets, surtout si les individus sont nombreux. Aussi, les données sont présentées le plus souvent sous la forme d'un tableau de contingence.

• Les tableaux disjonctifs complets servent surtout à exposer le principe de l'AFC.

# 2. L'AFC est une AC particulière

## 2.1. L'AC entre 2 variables qualitatives

• L'AFC est l'AC des 2 tableaux $X_1$ et $X_2$.

• À l'étape $k$, le problème de l'AC est la détermination de $z_1^k$ et de $z_2^k$ de telle manière que $R^2(z_1^k, z_2^k)$ ait une valeur maximale (cf Cours 9).

• $z_1^k$ et $z_2^k$ sont resp. les vecteurs propres d'ordre $k$ de $P_1P_2$ et $P_2P_1$.

• $z_1^k = X_1a_1^k$ est une quantification de la première variable qualitative, le codage des modalités étant donné par les lignes de $a_1^k$.

• De façon symétrique, le codage des modalités de la seconde variable qualitative est donné par $a_2^k$ et $z_2^k = X_2a_2^k$.

• En reprenant les notations de l'AC (cf Cours 9), c'-à-d en notant $V_{ij}$ la matrice $\frac{1}{n}(X_i)'X_j$, les facteurs vérifient les équations suivantes :

$$\begin{aligned}(V_{11})^{-1}V_{12}a_2^k &= R(z_1^k, z_2^k)a_1^k\\(V_{22})^{-1}V_{21}a_1^k &= R(z_1^k, z_2^k)a_2^k\end{aligned}$$

ou encore

$$\begin{aligned}(V_{11})^{-1}V_{12}(V_{22})^{-1}V_{21}a_2^k &= R^2(z_1^k, z_2^k)a_1^k\\(V_{22})^{-1}V_{21}(V_{11})^{-1}V_{12}a_1^k &= R^2(z_1^k, z_2^k)a_2^k.\end{aligned}$$

• $V_{12}$ est le tableau de dimension $m_1 \times m_2$ des fréquences relatives des 2 variables qualitatives.

• $(V_{11})^{-1}$ est la matrice diagonale des inverses des fréquences relatives des modalités de la variable qualitative numéro 1.

• L'élément à l'intersection de la ligne $i$ et de la colonne $j$ du tableau $(V_{11})^{-1}V_{12}$ est égal à $\dfrac{f_{ij}}{f_{i.}}$.

• $(V_{11})^{-1}V_{12}$ est le tableau des profils des lignes.

• De façon similaire, $(V_{22})^{-1}V_{21}$ est le tableau de dimensions $m_2$ et $m_1$ dont l'élément à l'intersection de la ligne $j$ et de la colonne $i$ est $\dfrac{f_{ij}}{f_{.j}}$.

• $(V_{22})^{-1}V_{21}$ est le tableau transposé des profils des colonnes.

• Les facteurs et les variables canoniques possèdent la propriété suivante :

**Les variables canoniques obtenues à partir des indicatrices non centrées sont centrées.**

**Les facteurs canoniques sont centrés.**

**Rappel :**

• Centrer une variable consiste à projeter cette variable orthogonalement au vecteur $u_n$, dont les $n$ coordonnées sont égales à 1.

• Il n'est donc pas nécessaire de centrer les colonnes des tableaux $X_1$ et $X_2$ avant d'effectuer l'AC.

• L'AC effectuée à partir des variables non centrées fournit les mêmes résultats aux valeurs triviales près.

## 2.2. Les corrélations canoniques et le Khi-deux

Le lien entre les corrélations canoniques de l'AFC des 2 variables qualitatives et la distance du Khi-deux qui mesure la liaison entre ces variables qualitatives est le suivant :

**À la valeur propre triviale près, la somme des carrés des corrélations canoniques de l'AFC de 2 variables qualitatives est égale à $\chi^2$ divisé par $n$.**

$$\sum_{k=2}^{m_1} R^2(z_1^k, z_2^k) = \frac{\chi^2}{n}.$$

Par conséquent :

**Le Khi-deux est une mesure globale de la liaison entre 2 variables qualitatives.**

**L'AFC décrit plus précisément cette liaison, en la décomposant en plusieurs relations entre des couples de variables canoniques.**

**L'intensité de chacune de ces relations est mesurée par le coefficient de corrélation canonique.**

Comme en AC, l'interprétation des résultats porte sur les étapes pour lesquelles la corrélation entre les variables canoniques est forte, c'-à-d les premières étapes.

# 3. Les représentations graphiques

• En AC, pour décrire les résultats de l'étape $r$, on représente sur le même axe les composantes principales des 2 tableaux, soit $z_1^r$ et $z_2^r$.

• La relation qui lie composante canonique et facteur est ici particulièrement simple.

• Considérons par exemple, la composante d'ordre $r$ du premier tableau :

$$z_1^r = X_1 a_1^r.$$

$z_1^r$ prend $n$ valeurs, une par individu, mais seules $m_1$ de ces valeurs sont différentes et correspondent aux $m_1$ valeurs prises par le facteur $a_1^r$.

• Les graphiques seront alors constitués par $m_1 + m_2$ points, qui représentent les modalités des 2 variables qualitatives.

• Notons une différence de normalisation des composantes entre l'approche AC et l'approche ACP :

– les composantes canoniques ont une variance égale à 1

– les compoosantes principales ont une variance égale à la valeur propre de l'étape dont elles sont issues.

• C'est cette dernière convention qui est le plus souvent admise et que nous utiliserons ici pour les représentations graphiques.

# 4. Les aides à l'interprétation

Comme en ACP, 2 types d'aides à l'interprétation sont calculés pour permettre de mieux comprendre les graphiques de l'AFC :

- les contributions des modalités à la variance
- la qualité de représentation des modalités.

## 4.1. Les contributions des modalités à la variance

La variance d'une composante principale donnée, $a_1^r$ par exemple, est égale à la valeur propre d'ordre $r$.

Cette variance peut être calculée de la façon suivante :

$$\text{Var}[a_1^r] = \sum_{i=1}^{m_1} f_{i.}(a_{1i}^r)^2 = R^2(z_1^r, z_2^r).$$

Donc :

**La contribution de la modalité $i$ de la première variable qualitative à la variance est égale à :**

$$\frac{f_{i.}(a^r_{1i})^2}{R^2(z^r_1, z^r_2)}.$$

**La contribution de la modalité $j$ de la seconde variable qualitative à la variance est égale à :**

$$\frac{f_{.j}(a^r_{2j})^2}{R^2(z^r_1, z^r_2)}.$$

• Pour chacune des 2 variables qualitatives, la somme des contributions des modalités est égale à 100%.

• Les modalités qui contribuent fortement à la variance d'une composante principale “expliquent” cette composante principale.

## 4.2. La qualité de représentation des modalités

• La modalité $i$ est représentée par un point de $\mathbb{R}^{m_2}$.

• Comme en ACP, on peut mesurer la qualité de représentation de ce point sur l'axe $r$.

• Le centre du nuage a pour $j$ième coordonnée, dans le cas de la première analyse

$$\sum_{i=1}^{m_1} f_{i.} \frac{f_{ij}}{f_{i.}} = \sum_{i=1}^{m_1} f_{ij} = f_{.j}$$

et le carré de la distance de $i$ au centre du nuage est égal à :

$$d^2(i, O) = \sum_{j=1}^{m_2} \frac{1}{f_{.j}} \left( \frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 .$$

• Le cosinus carré, qui mesure la qualité de représentation de $i$, est égal à

$$\frac{(a_{2j}^r)^2}{\displaystyle\sum_{i=1}^{m_1} \frac{1}{f_{i.}} \left( \frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2},$$

où $a_{2j}^r$ désigne la $j$ième coordonnée de la composante principale d'ordre $r$ du second tableau.

• Les modalités bien représentées sont "expliquées" par la composante principale. Pour un individu donné, la somme des qualités de représentation pour l'ensemble des axes est égale à 100%.

# 5. Quelques applications de l'AFC

• Vous pourrez appliquer toute la théorie de l'analyse factorielle des correspondances décrite dans les paragraphes précédents en faisant les exercices qui vous ont été distribués.

• On ne vous demande pas de savoir faire les calculs à la main mais seulement de savoir interpréter les informations que renvoie les logiciels comme Minitab.

• C'est d'ailleurs ce dernier point qui pourra être demandé à l'examen et en pratique.