Republique Algerienne Démocratique & Populaire.

Ministère De L'Enseignement Supérieur Et de La Recherche Scientifique

Université Hassiba Benbouali - Chlef

Faculté des Sciences Exactes & Informatique

Département de Mathématiques

Module : Analyse de données

Chap03 : Analyse des Correspondances Multiples

Cours déstiné aux étudiants de Master Mathématiques
(en cours de rédaction)

Présenté par Dr :HAMEL Elhadj [1]

2020/2021

1. e-mail : hamel_2@yahoo.fr

Table des matières

Chapitre 1

Analyse des Correspondances Multiples

1.1 Introduction

L'analyse factorielle des correspondances multiples (ACM ou AFCM) est la généralisation de l'analyse des correspondances au cas de plusieurs variables. Elle consiste donc à représenter les modalités de variables qualitatives dans un espace euclidien dans lequel les distances du χ^2 entre deux modalités d'une même variable sont préservées au mieux. On considère donc dans cette section p variables qualitatives observées simultanément sur n individus de poids identiques $\frac{1}{n}$.
On peut faire remonter les principes de cette méthode à Guttman (1941), mais aussi à Burt (1950) ou à Hayashi (1956). D'autres types d'extension ont été proposés par Benzécri (1973), Escofier-Cordier (1965), et par Masson (1974) qui s'appuie notamment sur les travaux de Carroll (1968), Horst (l961) et Kettenring (1971)

1.2 Les domaines d'application

Cette analyse très simple est non plus adaptée aux tableaux de contingence de l'AFC, mais aux *tableaux disjonctifs complets* que nous décrivons ci-dessous. Ces tableaux sont des tableaux logiques pour des variables codées. Les propriétés de tels tableaux font de l'ACM une méthode spécifique aux règles d'interprétation des représentations simples. *Elle permet donc l'étude des liaisons entre plus de deux variables qualitatives, ce qui étend le spectre d'étude de l'AFC.*
L'ACM est donc très bien adaptée au traitement d'enquêtes lorsque les variables sont qualitatives (ou rendues qualitatives). Il est également possible de n'appliquer cette méthode plusieurs fois en ne prenant en compte que quelques variables.

1.2.1 Les objectifs

Les objectifs que cette méthode spécifique, l'ACM, doit remplir sont les mêmes que ceux de l'ACP ou de l'AFC. Il s'agit d'obtenir une typologie des lignes et des colonnes et relier ces deux typologies. Nous aurons ici trois familles d'éléments à étudier, les individus, les variables et les modalités des variables. Cette étude se fait par la définition de ressemblances et liaisons pour ces trois familles que nous détaillons dans la section suivante. Afin d'établir un bilan des

ressemblances entre individus, comme en ACP nous cherchons à répondre à des questions du type :

— Quels sont les individus qui se ressemblent ?
— Quelles sont ceux qui sont différents ?
— Existe-t-il des groupes homogènes d'individus ?
— Est-il possible de mettre en évidence une typologie des individus ?

Les mêmes types de questions se posent pour les variables et les modalités.

1.3 Notations et définitions

1.3.1 Les données

L'ACM permet l'étude de tableaux décrivant une population de n individus et p variables qualitatives. Une variable qualitative (ou nominale) peut être décrite par une application de l'ensembles des I individus dans un ensemble fini non structuré, par exemple non ordonné. Ces variables qualitatives peuvent être codées par un codage condensé qui attribue une valeur à chaque modalité. Par exemple les modalités pour la couleur d'un vin peuvent être 1 pour le rouge, 2 pour le blanc et 3 pour le rosé. Les données peuvent donc être représentées sous la forme d'une matrice X décrite par le ci-dessous,

$$\begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & \dots & j & \dots & p \end{array} \\ \begin{pmatrix} & & & & \\ & & \vdots & & \\ & \dots & x_{ij} & \dots & \\ & & \vdots & & \\ & & & & \end{pmatrix} \end{array}$$

à la fois le nombre de variables et l'ensemble des variables et x_{ij} est le codage condensé de l'individu i pour la variable j.

Les x_{ij} représentant une codification, en prendre la moyenne n'a aucun sens. Ces données ne peuvent donc pas être traitées par l'ACP ou l'AFC précédemment étudiées. Ce tableau présente donc des spécificités dont l'analyse factorielle doit tenir compte par une méthode spécifique.

Exemple 1.3.1.

	Bac	*coleur des yeux*	*groupe sanguin*	*sexe*
ind1	*Science*	*Noir*	*O*	*M*
ind2	*science*	*Bleu*	*O*	*F*
ind3	*Lettre*	*Noir*	*A*	*M*
ind4	*Economie*	*Marron*	*AB*	*M*
ind5	*Langues*	*Marron*	*B*	*F*

Exemple 1.3.2. *Soit le tableau de donnéees suivant où cinq individus sont décrits à l'aide de deux variables qualitatives X et Y possédant respectivement deux (x_1 et x_2) et trois (y_1, y_2 et*

y_3) modalités.

	1	*2*	*3*	*4*	*5*
X	x_1	x_2	x_2	x_1	x_1
Y	y_2	y_3	y_1	y_3	y_2

.

1.3.2 Tableau disjonctif complet

Une représentation des données est *le tableau disjonctif complet*. Il représente les individus en ligne(n individus), alors que les colonnes représentent les modalités (m modalités) des p variables (et non plus les variables) . Ainsi, à l'intersection de la ligne i avec la colonne s, la valeur $\mathcal{K}_{is}$ vaut 1 si l'individu i possède la modalité k et 0 sinon.

$$\begin{cases} \mathcal{K}_{is} = 1, & \text{si l'individu i possède la modalité s ;} \\ \mathcal{K}_{is} = 0, & \text{si non.} \end{cases}$$

$$\mathcal{K} = \begin{array}{c|ccccc|c} & 1 & \dots & s & \dots & m & \mathcal{K}_{i.} \\ \hline 1 & & & & & & . \\ \vdots & & & \vdots & & & \\ i & & \dots & \mathcal{K}_{is} & \dots & & p \\ \vdots & & & \vdots & & & \\ n & & & & & & \\ \hline \mathcal{K}_{.s} & & & n_s & & & np \end{array}$$

Ce tableau porte le nom *de disjonctif complet*, car l'ensemble des valeurs $\mathcal{K}_{is}$ d'un même individu pour les modalités d'une même variable, comporte la valeur 1 une fois (complet) et une fois seulement (disjonctif). Chaque modalité s est relié à une variable j. Nous avons ainsi trois familles d'éléments les individus, les variables et les modalités.
Notons m_j le nombre des modalités de la variable j et $m = m_1 + m_2 + ... + m_j + ... + m_p$ le nombre des modalités de toutes les variables confondues $\{1, ..., j, .., p\}$. Nous avons donc les égalités suivantes :

- les totaux en ligne sont constants et égaux au nombre p de variables.
- es totaux en colonne correspondent au nombre dŠindividus possédant la modalité s, noté n_s.
- le total général vaut ici np.

Exemple 1.3.3. *Construire le tableau disjonctif complet associé au tableau de l'exemple 2 : n=5 ,p=2 ,m=2+3=5 de terme général k_{is}.*

$$\mathcal{K} = \left(\begin{array}{c|ccccc|c} & x_1 & x_2 & y_1 & y_2 & y_3 & marge \\ \hline 1 & 1 & 0 & 0 & 1 & 0 & 2 \\ 2 & 0 & 1 & 0 & 0 & 1 & 2 \\ 3 & 0 & 1 & 1 & 0 & 0 & 2 \\ 4 & 1 & 0 & 0 & 0 & 1 & 2 \\ 5 & 1 & 0 & 0 & 1 & 0 & 2 \\ \hline marge & 3 & 2 & 1 & 2 & 2 & np = 10 \end{array}\right)$$

— **Matrice des fréquences F :** A partir de TDC on déduit le tableau une matrice de fréquence F (associée à un TDC) de terme générale $f_{is} = \frac{\mathcal{K}_{is}}{np}$.

Propriété 1.3.1. — *$f_{is} = \dfrac{1}{np}$ si l'individu i possède la modalité s et $f_{is} = 0$ sinon.*

$$\begin{cases} f_{is} = \dfrac{1}{np}, & \text{si l'individu i possède la modalité s ;} \\ f_{is} = 0, & \text{si non.} \end{cases}$$

— *le poids des lignes (indvidus) est constant et vaut $\dfrac{1}{n}$.*

— *le poids des colonnes (modalités) vaut $\dfrac{n_s}{np}$. Il est d'autant plus grand que la modalité est fréquente.*

$$F = \begin{array}{c|ccccc|c} & 1 & \dots & s & \dots & m & \\ \hline 1 & & & & & & . \\ \vdots & & & \vdots & & & \\ i & & \dots & f_{is} = \dfrac{\mathcal{K}_{is}}{np} & \dots & & \frac{1}{n} \\ \vdots & & & \vdots & & & \\ n & & & & & & \\ \hline & & & \frac{n_s}{np} & & & \end{array}$$

On notera aussi :

— $r = (\dots, \frac{1}{n}, \dots)^t \in \mathbb{R}^n$ *le vecteur des poids des individus.*

— $c = (\dots, \frac{n_s}{np}, \dots)^t \in \mathbb{R}^m$ *le vecteur des poids des modalités.*

— D_r *= diag(r)*

— D_c *= diag(c)*

On déduit de ce tableau de fréquences une matrice de profil-lignes **L** et une matrice des profil-colonnes **C**.

— **Matrice des profil-lignes L** de terme générale $\ell_{is} = \dfrac{f_{is}}{f_{i.}} = \dfrac{\mathcal{K}_{is}}{p}$

$$L = \begin{array}{c|ccccc|c} & 1 & \dots & s & \dots & m & \\ \hline 1 & & & & & & . \\ \vdots & & & \vdots & & & \\ i & & \dots & \ell_{is} = \dfrac{\mathcal{K}_{is}}{p} & \dots & & \\ \vdots & & & \vdots & & & \\ n & & & & & & \\ \hline c^t & & & \frac{n_s}{np} & & & \end{array}$$

On a :

— $L = D_r^{-1}.F$

— Profil ligne moyen :$\overline{L} = c$.

— **Matrice des profil-colonnes C** : de terme générale $c_{is} = \dfrac{f_{is}}{f_{.s}} = \dfrac{\mathcal{K}_{is}}{n_s}$

$$C = \begin{array}{c|ccccccc|c} & 1 & \dots & & s & & \dots & m & \\ \hline 1 & & & & & & & & \\ \vdots & & & & \vdots & & & & \\ i & & \dots & c_{is} = \dfrac{\mathcal{K}_{is}}{n_s} & & \dots & & & . \\ \vdots & & & & \vdots & & & & \\ n & & & & & & & & \\ \hline & & & & & & & & \end{array}$$

On a :

— $L = F.D_c^{-1}$

— Profil ligne moyen :$\overline{C} = r$.

— **Deux nuages de points pondérés :**

On considère dans la suite les deux nuages centrés de point suivants :

— Le nuage des n point-individus centrés de $\mathbb{R}^m$ c'est à dire les n lignes de la matrice des profil-lignes centrés $L = D_r^{-1}(F - rc^t)$. Chaque point-individu est pondéré par $\frac{1}{n}$.

— Le nuage des m point-modalités centrés de $\mathbb{R}^n$ c'est à dire les m lignes de la matrice des profil-colonnes centrés $C = (F - rc^t)D_c^{-1}$. Chaque point-modalité est pondéré par $\dfrac{n_s}{np}$.

1.3.3 Tableau de Burt

On construit le tableau de Burt, à partir du tableau disjonctif complet $\mathcal{K}$, le tableau symétrique B d'ordre (m,m) qui rassemble les croisements deux à deux de toutes les variables :

$$B = \mathcal{K}^t \times \mathcal{K}$$

B est appelé *tableau de contingence de Burt* associé au tableau disjonctif complet $\mathcal{K}$.

On peut écrire le terme générale de tableau $B = (B_{ss'})_{s,s'=1\dots m} = \sum_{i=1}^{n} \mathcal{K}_{is}\mathcal{K}_{is'}$ où

— $m = m_1 + m_2 + ... + m_p$ est nombre total des modalités des variables.

— si $s \neq s', B_{ss'}$ est la table de contingence des variables X_j et $X_{j'}$,

— si $s = s', B_{ss'}$ est une matrice diagonale contenant les effectifs marginaux de K dans la diagonale, notés $n_1, n_2, ..., n_s, ...n_m$.

Propriété 1.3.2. - *B est symétrique.*

- La somme des lignes (resp. des colonnes) de B est pn_s .

- La somme des éléments de B est p^2n.

- si on considère les données du tableau disjonctif $\mathcal{K}$ comme des observations de variables qualitatives, alors le tableau de Burt représente la variance de $\mathcal{K}$ à un facteur multiplicatif près.

Exemple 1.3.4. *Le tableau ci-dessous résume les données de six personnes avec trois variables (* ***Sexe*** *: H(Homme) ,F(Femme)) ; (* ***Nationalité*** *(A (Algerienne) ,E(Etranger)) ; (****Couleur des yeux*** *: B(bleu), M(marron),N(Noir)) :*

Sexe	*H*	*F*	*F*	*H*	*F*	*H*
Nationnalité	*A*	*E*	*E*	*E*	*A*	*A*
Couleur des yeux	*B*	*M*	*N*	*B*	*M*	*N*

— *Le tableau disjonctif complet* $\mathcal{K}$ *et le tableau de Burt* $B = \mathcal{K}^t \times \mathcal{K}$

$\mathcal{K}=$

	H	*F*	*A*	*E*	*B*	*M*	*N*	$\mathcal{K}_{i.}$
1	*1*	*0*	*1*	*0*	*1*	*0*	*0*	*3*
2	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*3*
3	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*3*
4	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*3*
5	*0*	*1*	*1*	*0*	*0*	*1*	*0*	*3*
6	*1*	*0*	*1*	*0*	*0*	*0*	*1*	*3*
$\mathcal{K}_{.s}$	***3***	***3***	***3***	***3***	***2***	***2***	***2***	*n.p=18*

, $B=$

	H	*F*	*A*	*E*	*B*	*M*	*N*
H	***3***	*0*	*2*	*1*	*2*	*0*	*1*
F	*0*	***3***	*1*	*2*	*0*	*2*	*1*
A	*2*	*1*	***3***	*0*	*1*	*1*	*1*
E	*1*	*2*	*0*	***3***	*1*	*1*	*1*
B	*2*	*0*	*1*	*1*	***2***	*0*	*0*
M	*0*	*2*	*1*	*1*	*0*	***2***	*0*
N	*1*	*1*	*1*	*1*	*0*	*0*	***2***

1.3.4 Distance du χ^2

En ACM on utilise la distance du χ^2 pour comparer deux individus d'écrits par deux points de $\mathbb{R}^m$ (deux profil-lignes) ou deux modalités d'écrites par deux points de $\mathbb{R}^n$ ((deux profil-colonnes). En ACP, les individus et les variables étaient les lignes et les colonnes d'une même matrice (la matrice des données centrées-réduites). En ACM, les individus et les modalités sont les lignes et les colonnes de deux matrices différentes (resp. la matrice des profil-lignes et la matrice des profil-colonnes). Pour comparer deux individus ou deux modalités, on utilise en ACM la distance du χ^2.

— Distance du χ^2 entre deux individus : métrique D_c^{-1} :

$$d^2(i,i') = \sum_{s=1}^{m} \frac{1}{f_{.s}} \left(\frac{\mathcal{K}_{is} - \mathcal{K}_{i's}}{p} \right)^2 = \frac{n}{p} \sum_{s=1}^{m} \frac{1}{\mathcal{K}_{.s}} (\mathcal{K}_{is} - \mathcal{K}_{i's})^2 = \frac{n}{p} \sum_{s=1}^{m} \frac{1}{n_s} (\mathcal{K}_{is} - \mathcal{K}_{i's})^2$$

Donc deux individus sont proches s'ils possèdent les mêmes modalités, sachant que l'on donne plus de "poids" dans cette distance au fait que ces deux individus ont en commun une modalité rare (n_s petit).

— Distance du χ^2 entre deux modalités : métrique D_r^{-1} :

$$d^2(s,s') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{\mathcal{K}_{is}}{n_s} - \frac{\mathcal{K}_{i's}}{n_{s'}} \right)^2 = n \sum_{i=1}^{n} \left(\frac{\mathcal{K}_{is}}{\mathcal{K}_{.s}} - \frac{\mathcal{K}_{is'}}{\mathcal{K}_{.s'}} \right)^2 = n \sum_{i=1}^{n} \left(\frac{\mathcal{K}_{is}}{n_s} - \frac{\mathcal{K}_{is'}}{n_{s'}} \right)^2$$

Donc deux modalités sont proches si elles sont possédées par les mêmes individus.

1.4 Inertie totale

En AFC, on pouvait interpréter statistiquement l'inertie des nuages de points (profil-lignes et colonnes) en terme de χ^2/n mesurant l'indépendance entre les deux variables qualitatives. En AFCM ce n'est plus le cas puisque l'on a :

$$I(L) = I(C) = \frac{m}{p} - 1$$

On a donc l'inertie qui dépend de $\frac{m}{p}$, le nombre moyen de catégories par variable.

Exercice : Démontrer la relation de l'inertie.

1.5 AFC du tableau disjonctif complet

1.5.1 Le principe :

Effectuer une ACM consiste à appliquer l'AFC au TDC c'est à dire à effectuer une ACP pondérée des nuages des point-individus et des point-modalités (centrés). En reprenant les résultats de l'analyse des correspondances et les notations adoptées ici :
Matrice des fréquences F de terme générale $f_{is} = \frac{\mathcal{K}_{is}}{np}$.

$$D_1 = D_r = \begin{pmatrix} f_{i.} & & \\ & \ddots & \\ & & f_{n.} \end{pmatrix} \quad , \quad D_2 = D_c = \begin{pmatrix} f_{.1} & & \\ & \ddots & \\ & & f_{.np} \end{pmatrix}$$

c'est à dire :

- Nous désignerons par D la matrice diagonale, d'ordre (m,m) ayant les mêmes éléments diagonaux que B ; ces éléments sont les effectifs correspondant à chacune des modalités
  $\begin{cases} d_{ss} = b_{ss} = \mathcal{K}_{.s}. \\ d_{ss'} = 0 pour tout s \neq s' \end{cases}$
- $D_c = \frac{D}{np}\mathbb{I}$ de terme générale $f_{.j} = \delta_{ij}\dfrac{\mathcal{K}_{.s}}{np}$
- $D_r = \frac{1}{n}\mathbb{I}_n$ de terme générale $f_{i.} = \dfrac{\delta_{ij}}{n}$

$\mathbb{I}_n$ est la matrice identité d'ordre (n,n) et δ_{ij} est tel que : $\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ si $i \cong j$.
Ses principes sont donc ceux de l'analyse des correspondances à savoir :

- mêmes transformations du tableau de données en profils-lignes et en profils-colonnes ;
- même critère d'ajustement avec pondération des points par leurs profils marginaux ;
- même distance, celle du χ^2.

L'analyse des correspondances multiples présente cependant des propriétés particulières dues à la nature même du tableau disjonctif complet. Nous allons énoncer les principes de cette analyse à partir du tableau disjonctif complet puis nous montrerons l'équivalence avec l'analyse du tableau de Burt

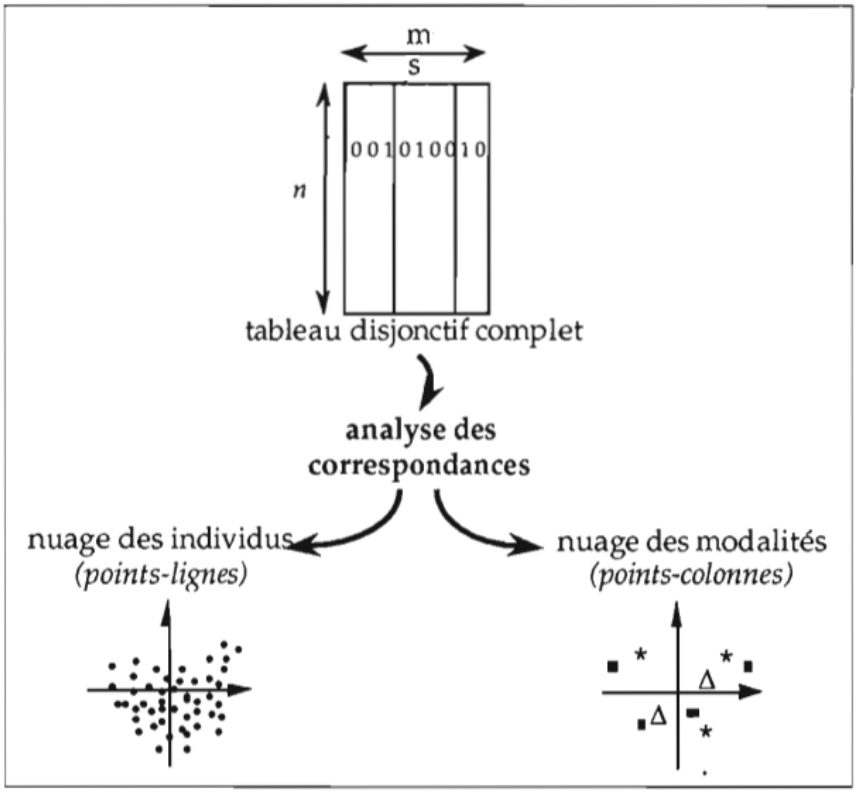

FIGURE 1.1 – Analyse des correspondances multiples

1.5.2 Axes factoriels et facteurs

Pour trouver les axes factoriels u_α on diagonalise la matrice :

$$W = F' D_r^{-1} F D_c^{-1} = \frac{1}{p} \mathcal{K}^t . \mathcal{K} . D^{-1}$$

de terme général

$$\omega_{ss'} = \frac{1}{p \mathcal{K}_{.s'}} \sum_{i=1}^{n} \mathcal{K}_{is} \mathcal{K}_{is'}$$

Dans $\mathbb{R}^m$, l'équation du α^{ime} axe factoriel u_α est :

$$\frac{1}{p} \mathcal{K}^t . \mathcal{K} . D^{-1} u_\alpha = \lambda_\alpha u_\alpha .$$

L'équation du α^{ime} facteur $\phi_\alpha = D^{-1} . u_\alpha$ s'écrit :

$$\frac{1}{p} D^{-1} \mathcal{K}^t . \mathcal{K} . \phi_\alpha = \lambda_\alpha \phi_\alpha .$$

De même, l'équation du α^{ime} facteur ψ_α dans $\mathbb{R}^n$ s'écrit :

$$\frac{1}{p} \mathcal{K} . D^{-1} . \mathcal{K}^t . \psi_\alpha = \lambda_\alpha \psi_\alpha .$$

Les facteurs ϕ_α et ψ_α (de norme λ_α) représentent les coordonnées des points-lignes et des points-colonnes sur l'axe factoriel α.

Les relations de transition entre les facteurs et ϕ_α et ψ_α sont :

$$\begin{cases} \phi_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} D^{-1}.\mathcal{K}^t\psi_\alpha, \\ \psi_\alpha = \dfrac{1}{p\sqrt{\lambda_\alpha}} \mathcal{K}\phi_\alpha, \end{cases}$$

La coordonnée factorielle de l'individu i sur l'axe α est donnée par :

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{s=1}^{m} \frac{\mathcal{K}_{is}}{\mathcal{K}_{.s}} \phi_{\alpha s}$$

c'est-à-dire :

$$\psi_{\alpha i} = \frac{1}{\sqrt{p\lambda_\alpha}} \sum_{s=1}^{m} \phi_{\alpha s}$$

De même, la coordonnée de la modalité s sur l'axe α est donnée par :

$$\phi_{\alpha s} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{\mathcal{K}_{is}}{\mathcal{K}_{.s}} \psi_{\alpha i}$$

Remarque 1.5.1.

1. *Dans le calcul des relations de transitions où quasi-barycentriques ci-dessus, les individus ne sont pas pondérés. Il s'agit de simples calculs de moyennes arithmétiques de coordonnées.*
2. *Si le tableau disjonctif n'est pas complet (c'est-à-dire si, pour au moins un individu, aucune modalité de réponse à une question n'a été choisie), les modalités d'une même variable ne sont plus centrées sur le centre de gravité du nuage global.*
3. *Le codage disjonctif complet permet de transformer une variable continue en une variable nominale dont les modalités sont des classes ordonnées. Il est alors utile de tracer la trajectoire qui relie les classes, trajectoire qui peut suggérer des liaisons non linéaires entre cette variable et les axes.*
4. *Dans l'analyse par rapport au centre de gravité G, on trouvera donc $m-p$ valeurs propres non nulles. En choisissant une base dans le support du nuage, on pourra se ramener à la recherche d'éléments propres d'une matrice d'ordre $m-p$.*

1.5.3 Représentation simultanée

La présentation de l'analyse des correspondances peut être formulée ici de façon particulière en raison du codage spécifique au tableau disjonctif complet.
Nous cherchons sur un même axe les coordonnées des n individus et des m modalités de façon que :

— la coordonnée d'un individu i soit la moyenne arithmétique des coordonnées des modalités qu'il a choisies .

— la coordonnée d'une modalité s soit la moyenne arithmétique des coordonnées des individus qui l'ont choisie .

Bien entendu, on obtient les relations dite quasi-barycentriques issues de l'analyse du tableau disjonctif complet $\mathcal{K}$ avec,

$$\begin{cases} \phi = \frac{1}{\sqrt{\lambda_\alpha}} D^{-1}.\mathcal{K}^t\psi, \\ \psi = \frac{1}{p\sqrt{\lambda_\alpha}} \mathcal{K}\phi, \end{cases}$$

La représentation simultanée des individus et des modalités est importante pour l'interprétation des résultats. Cependant elle n'est pratiquement pas utilisée, d'une part pour des raisons d'encombrement graphique (on dispose souvent de plusieurs centaines voire de plusieurs milliers d'individus) et d'autre part parce que les individus sont, dans la plupart des applications, anonymes. Ils ne présentent de l'intérêt que par l'intermédiaire de leurs caractéristiques. On peut cependant vouloir projeter les individus sur un plan factoriel afin d'apprécier leur répartition et les zones de densité.

1.6 Règles d'interprétation

L'analyse des correspondances multiples met alors en évidence des types d'individus ayant des profils semblables quant aux attributs choisis pour les décrire. Compte tenu des distances entre les éléments du tableau disjonctif complet et des relations barycentriques particulières.
Les individus et les modalités sont représentés sur ses plans de projection dont la lecture nécessite des règles d'interprétation.

1.6.1 Inertie expliquée

On a vu qu'en ACM, l'inertie totale du nuage des individus et du nuage des modalités vaut $\frac{m}{p} - 1$ et ne dépend donc que du nombre moyen de modalités par variable. De plus l'inertie totale est égale à $\lambda_1 + \lambda_2 + ... + \lambda_r$ où $r = \min(n-1; m-p)$ est le nombre de valeurs propre non nulles. Le pourcentage d'inertie expliquée par un axe α est donc :

$$\frac{\lambda_\alpha}{\lambda_1 + \lambda_2 + ... + \lambda_r} * 100$$

Remarque 1.6.1. *en ACM, les pourcentages d'inertie expliquées par les axes sont par construction "petits" et ne peuvent donc pas être interprétés comme en AFC ou en ACP. Le nombre d'axes retenus pour l'interprétation ou le recodage ne peut pas être choisi à partir de ces pourcentages d'inertie expliquées.*

1.6.2 Contributions

On reprend les formules de l'AFC et on trouve qu'en ACM la contribution d'un individu i et la contribution d'une modalité s à l'inertie de l'axe α s'écrivent :

$$\begin{cases} Ctr_\alpha(i) = \dfrac{1}{n}\dfrac{\psi_{\alpha i}^2}{\lambda_\alpha}, \\ Ctr_\alpha(s) = \dfrac{n_s}{np}\dfrac{\phi_{\alpha s}^2}{\lambda_\alpha}, \end{cases}$$

On en déduit qu'en pratique :

- Les individus les plus excentrés sur les plans factoriels sont ceux qui contribuent le plus.
- En revanche, les modalités les plus excentrées ne sont pas nécessairement celles qui contribuent le plus. En effet, leur contribution dépend de leur fréquence.

1.6.3 Cosinus carrés

On utilise les memes formules de $cos_\alpha^2(i)$ et $cos_\alpha^2(s)$ pour mesurer la qualité de la projection des individus et des modalités sur les axes factoriels.
En pratique, si deux individus **sont bien projetés** alors s'ils sont proches en projections, ils sont effectivement proches dans leur espace d'origine et on peut alors interpréter leur proximité :

- La proximité entre deux individus s'interprète en terme de distance (du χ^2) : deux individus se ressemblent s'ils ont choisis les mêmes modalités. C'est cohérent avec la relation barycentrique qui dit que les individus sont au barycentre des modalités qu'ils possèdent.
- La proximité entre deux modalités de **deux variables différentes** s'interprète en terme de distance (du χ^2) : deux modalités se ressemblent si elles sont possédées par les mêmes individus. C'est cohérent avec la relation barycentrique qui dit que les modalités sont au barycentre des individus qui les possèdent.

Afin de ne rien oublier pour l'interprétation des résultats, nous proposons de suivre le plan suivant :
- Définir le nombre de modalités des variables quantitatives, s'il y a des variables quantitatives intéressantes pour l'étude.
- Choisir le nombre d'axes de projection. Ce choix se fait toujours de la même façon que pour l'ACP ou l'AFC.
- Etudier les valeurs propres qui représentent l'inertie de chaque axe.
- Etudier la contribution des lignes et des modalités de la même façon que l'ACP.
- Etudier la contribution des variables en sommant les contributions des modalités d'une variable pour un facteur donné.
- Etudier les coordonnées des modalités et des individus actifs.
- Etudier les coordonnées des variables, des modalités et des individus supplémentaires s'il y en a.

1.7 Cas de deux variables

On se place dans le cas particulier où $p = 2$ c'est à dire le cas où l'on observe n individus sur deux variables qualitatives ayant respectivement m_1 et m_2 modalités.

On peut alors faire l'analyse factorielle d'un des deux tableaux suivant :

— le tableau disjonctif complet $\mathcal{K}$ de dimension $n \times (m_1 + m_2)$ comme en ACM.
— le tableau de contingence N de dimension $m_1 \times m_2$ croisant les deux variables qualitatives comme en AFC

On a la relation suivante entre les premières valeurs propres des deux analyses :

$$\mu_\alpha = (2\lambda_\alpha - 1)^2$$

où μ_α est la α^{me} valeur propre de l'analyse de N (l'AFC) et λ_α est la α^{me} valeur propre de l'analyse de $\mathcal{K}$ (l'ACM). On en déduit qu'à chaque valeur propre μ_α de l'AFC correspondent deux valeurs propres de l'ACM :

$$\begin{cases} \lambda_\alpha = \dfrac{1 + \sqrt{\mu_\alpha}}{2}, \\ \lambda_* = \dfrac{1 - \sqrt{\mu_\alpha}}{2}, \end{cases}$$

On a également la relation suivante entre les coordonnées factorielles des deux analyses :

$$\begin{pmatrix} \psi_\alpha \\ \phi_\alpha \end{pmatrix} \text{ correspondant à la valeur propre } \lambda_\alpha = \frac{1 + \sqrt{\mu_\alpha}}{2}.$$
$$\begin{pmatrix} \psi_\alpha \\ -\phi_\alpha \end{pmatrix} \text{ correspondant à la valeur propre } \lambda_* = \frac{1 - \sqrt{\mu_\alpha}}{2}$$

et où ψ_α et ϕ_α sont les composantes principales des profil-lignes et colonnes de de tableau de contingence N .

Les conséquences pratiques de ces résultats sont :

— Dans l'ACM de $p = 2$ variables, on ne retriendra que les valeurs propres $\lambda_\alpha > 1/2$.qui correspondent aux composantes $\begin{pmatrix} \psi_\alpha \\ \phi_\alpha \end{pmatrix}$ En effet, les composantes qui correspondent aux valeurs propres λ_* sont liées aux précédentes par la relation ci-dessus.
— Les pourcentages d'inertie expliqués par les axes en ACM sont souvent trés faibles et ne peuvent donc pas être interprétés comme en AFC et en ACP.

1.8 AFC du tableau de Burt

Le tableau **B** de correspondance multiple, obtenu à partir d'un tableau disjonctif complet, est un assemblage particulier des tableaux de contingence qui sont les faces de l'hypercube de contingence.

L'analyse des correspondances appliquée à un tableau disjonctif complet $\mathcal{K}$ est équivalente à l'analyse du tableau de Burt **B** et produit les mêmes facteurs.

L'analyse des correspondances du tableau de Burt B, tableau symétrique d'ordre (m,m), se ramène à l'analyse d'un nuage de m points-modalités dans $\mathbb{R}^m$. Les marges de ce tableau, en ligne comme en colonne, sont les éléments diagonaux de la matrice D.

l'analyse du tableau disjonctif complet $\mathcal{K}$, la matrice à diagonaliser est :

$$W = \frac{1}{p} D^{-1} \mathcal{K}^t . \mathcal{K} = \frac{1}{p} D^{-1} . B$$

Pour l'analyse du tableau de B associé à $\mathcal{K}$, le tableau des fréquences relatives F s'écrit :

$$F = \frac{1}{np^2} B$$

et

$$D_r = \frac{1}{np} D$$

On diagonalise la matrice :

$$\boxed{W^\star = \tfrac{1}{p^2} D^{-1} B . D^{-1} . B}$$

ce qui donne :

$$\boxed{W^\star = W^2}$$

On a dans l'ACM avec **TDC** L'équation du α^{ime} facteur s'écrit :

$$\frac{1}{p} D^{-1} \mathcal{K}^t . \mathcal{K} . \phi_\alpha = \lambda_\alpha \phi_\alpha .$$

En prémultipliant les deux membres de cette relation par $\frac{1}{p} D^{-1} B$ on trouve :

$$\frac{1}{p^2} D^{-1} B . D^{-1} B . \phi_\alpha = \lambda_\alpha^2 \phi_\alpha .$$

Les facteurs des deux analyses sont donc colinéaires dans $\mathbb{R}^m$ mais les valeurs propres associées diffèrent. Celles issues de l'analyse de B, notées λ_B, sont le carré de celles issues de l'analyse de TDC $\mathcal{K}$:

$$\lambda_B = \lambda_\alpha^2$$

Les facteurs ϕ_α issus de l'analyse de $\mathcal{K}$, représentant les coordonnées factorielles des modalités, ont pour norme λ, alors que le facteur correspondant de l'analyse de B, noté $\phi_{B\alpha}$, aura pour norme λ^2.

D'où la relation liant les deux systèmes de coordonnées factorielles :

$$\phi_{B\alpha} = \phi_\alpha \sqrt{\lambda_\alpha} .$$

Cela veut dire dire que les coordonnées factorielles des points-modalités de l'analyse du tableau de Burt sont le mêmes que celles de l'analyse du TDC à $\sqrt{\lambda_\alpha}$ prés. Cela veut dire également qu'on peut obtenir les valeurs propres de l'analyse du tableau de Burt en élevant au carré les premières valeurs propres de l'analyse du TDC.

1.9 Conclusion

L'ACM est donc une analyse factorielle qui permet l'étude de plusieurs variables qualitatives, de ce fait elle est une généralisation de l'AFC. Elle est donc applicable aux tableaux de variables qualitatives, mais aussi quantitatives après construction de classes à partir de celles-ci. Le fait de pouvoir interpréter l'ACM de plusieurs façons rend cette méthode très riche et d'emploi facile. Elle peut être très complémentaire de l'ACP et bien sûr des méthodes de classification que l'on traiter dans le chapitre suivant.

1.10 Exercices

Exercice 1.10.1. *Montrez que les coordonnées des individus sur un axe sont de moyenne nulle.* Indication : Commencez par montrer que 1 est valeur propre triviale associée au vecteur $\mathbb{1}_m$ (on pourra se limiter au cas de 2 variables $p = 2$ pour la démonstration).
Utilisez le fait que les vecteurs propres de l'ACP sont D^{-1} orthonormés, sans oublier le fait que l'on préfère représenter les profils plutôt que les catégories

Exercice 1.10.2. *Calculer l'inertie de l'AFCM.*

Exercice 1.10.3. *Une enquête réalisée sur un échantillon comprenant 6 individus qui ont réponds à trois questions :*
- Q1 : Etes vous un homme ou une femme (H si Homme et F si une femme).
- Q2 : CSP (C si Cadre, O si Ouvrier et P si Profession libérale).
- Q3 : Type d'habitat (Pr si Propriétaire, L si Locataire et S si Sans logement).
Le résultat de cette enquête est donnée dans le tableau suivant :

	1	*2*	*3*	*4*	*5*	*6*
Lieu de résidence	*F*	*H*	*H*	*H*	*F*	*F*
CSP	*P*	*P*	*O*	*C*	*O*	*C*
Type d'habitat	*Pr*	*Pr*	*L*	*L*	*Pr*	*S*

1. *Quelle méthode utilisé pour faire l'analyse ? justifier ?*
2. *Transformer le tableau de données en un tableau logique.*
3. *Determiner le le tableau disjonctif complet.*
4. *Determiner le tableau de Burt.*

Exercice 1.10.4. *Soit le tableau de donnéees suivant où cinq individus sont décrits à l'aide de deux variables qualitatives X et Y possédant respectivement deux (x_1 et x_2) et trois (y_1, y_2 et y_3) modalités.*

	1	*2*	*3*	*4*	*5*
X	x_1	x_2	x_2	x_1	x_1
Y	y_2	y_3	y_1	y_3	y_2

.

Questions

1. *Rappeler la définition d'un tableau de contingence ?*
2. *Quelle méthode utilisé pour faire l'analyse ? justifier ?*
3. *Construire le tableau disjonctif complet associé au tableau .*
4. *En déduire le tableau des profils des individus et celui des profils des modalités.*
5. *À partir de tableau des profils lignes, calculer la distance du χ^2 entre les individus 3 et 4.*

Nous nous intéressons à un jeu de données fictif. Il comporte les réponses de 10 personnes aux trois questions suivantes :
a- Êtes-vous un homme ou une femme ?
b- Quel est votre niveau de revenus : moyen ou élevé ?
c- Choisissez le dessert que vous préférez parmi les trois suivants : un fruit (A), une crème glacée (B), du chocolat (C) ?

Les résultats dans le tableau suivant :

	1	2	3	4	5	6	7	8	9	10
Sexe	F	F	F	F	F	M	M	M	M	M
Revenu	M	M	E	E	E	E	E	M	M	M
préf	A	A	B	C	C	C	B	B	B	A

1. Quelle méthode utiliser pour faire l'analyse ?
2. Obtenir la table de contingence, puis le tableau de Burt et ensuite le tableau disjonctif complet ?
3. Réaliser l'AFCM du tableau de données des deux manières élémentaires suivantes :
   a L'analyse factorielle des correspondances du tableau de Burt.
   b L'analyse factorielle des correspondances du tableau de disjonctif complet.
4.) Quels commentaires pouvez-vous formuler sur les liens entre les résultats des deux analyses précédentes.

```
     eigenvalue percentage cumulative
dim1       0.36      58.27       58.27
dim2       0.22      34.48       92.74
dim3       0.04       6.60       99.35
dim4       0.00       0.65      100.00
dim5       0.00       0.00      100.00
dim6       0.00       0.00      100.00
```

```
     eigenvalue  percentage  cumulative
dim1       0.60       45.21       45.21
dim2       0.46       34.78       79.99
dim3       0.20       15.22       95.21
dim4       0.06       4.79       100.00
dim5       0.00       0.00       100.00
dim6       0.00       0.00       100.00
dim7       0.00       0.00       100.00
```

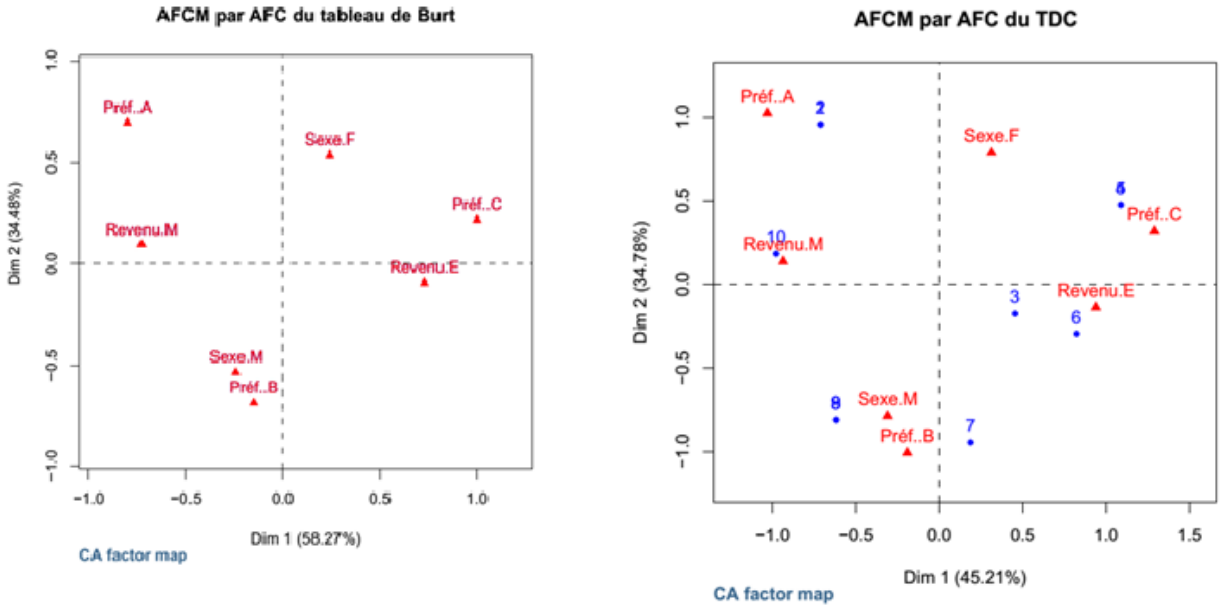

FIGURE 1.2 – Illustration des tableaux TDC et Burt