L'Analyse Factorielle des Correspondances

L'analyse factorielle des correspondances, notée AFC, est une analyse destinée au traitement des tableaux de données où les valeurs sont positives et homogènes comme les tableaux de contingence (qui constituent la majeure partie des tableaux traités par cette méthode).

L'AFC a été introduite de façon complète dans les années 60 par JP BENZECRI. L'AFC est une ACP. Les composantes principales sont toujours obtenues à partir de la distance entre les différents points des nuages multidimmensionnels, mais les points ont des coordonnées qui ont subis une transformation préalable permettant de conserver une métrique identique à celle de l'ACP pour calculer ces distances.

Le but principal de l'AFC reste donc le même ; lire l'information contenue dans un espace multidimensionnel par une réduction de la dimension de cet espace tout en conservant un maximum de l'information contenu dans l'espace de départ.

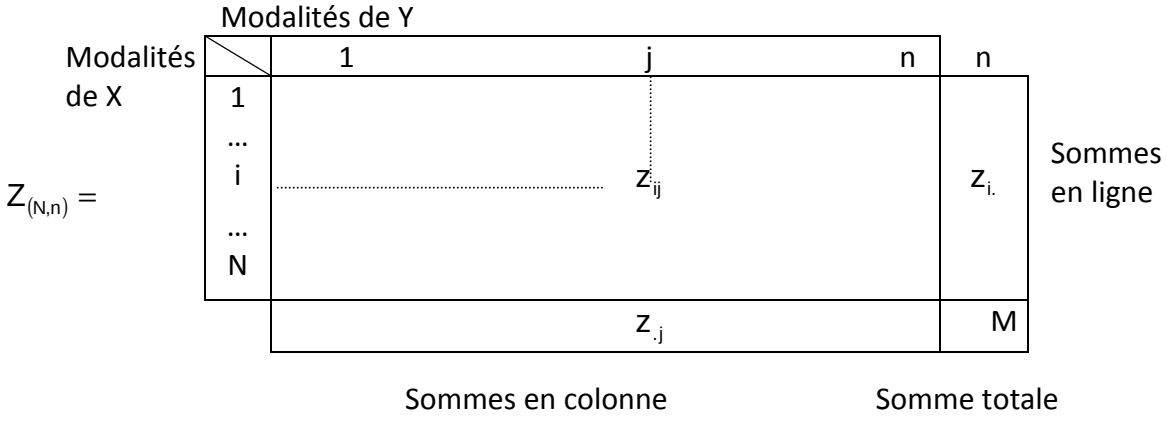
1 Le Tableau de données

L'AFC s'applique essentiellement à des tableaux de contingence. C'est un tableau d'effectifs qui contient à l'intersection de la ligne i et de la colonne j des zij individus. Il s'agit de la ventilation d'une population totale M selon deux caractères quelconques X en ligne et Y en colonne. Ce sont donc des caractères qualitatifs nominaux et/ou ordinaux.

L'étude traditionnelle d'un tel tableau se concentre le plus souvent sur la dépendance ou l'indépendance entre les deux caractères. Elle s'effectue généralement en utilisant le test du χ^2 et plus particulièrement par l'analyse de la variance (le rapport de corrélation) et la régression lorsque les deux caractères sont qualitatifs ordinaux en classes. (Cf module 1)

Dans un tableau de contingence, les modalités des caractères sont exclusives les unes par rapport aux autres et exhaustives. Il en résulte que les sommes en ligne et en colonne du tableau ont un sens.

- le tableau des données Z(N,n) se présente alors de la façon suivante :



Avec : $z_{i.} = \sum_{j=1}^{n} z_{ij}$ $z_{.j} = \sum_{i=1}^{N} z_{ij}$ $M = \sum_{i=1}^{N}\sum_{j=1}^{n} z_{ij} = \sum_{i=1}^{N} z_{i.} = \sum_{j=1}^{n} z_{.j}$

Exemple : au cours d'une enquête sur les vacances on a demandé à un échantillon de 100 indvidus d'indiquer leur Catégorie Socio professionnelle (caractère X) ainsi que le mode d'hébergement utilisé lors de leurs dernières vacances (Caractère Y)

Le tableau de données initial est donc

Individus	*CSP*	*Mode d'hébergement*
1	*Chef d'entreprise*	*Hôtel*
2	*Ouvrier*	*Camping*
3	*Cadre moyen*	*Famille , amis*
4	*Ouvrier*	*Camping*
5	*Profession intermédiaire*	*Location , gîte*
6	*Agriculteur*	*Famille , amis*
7	*Profession intermédiaire*	*Location , gîte*
8	*Cadre moyen*	*Camping*
…	…	…
100	*Employé*	*Camping*

Le tableau de contingence croisant les caractères X et Y est alors :

$Z_{(7,4)}$=

CSP\Mode d'hébergement	***Camping***	***Hôtel***	***Famille , amis***	***Location , gîte***	***Total(1)***
Agriculteur	*2*		*8*	*2*	*12*
Cadre moyen	*4*	*2*	*1*	*5*	*12*
Chef d'entreprise	*1*	*5*	*1*	*3*	*10*
Employé	*8*	*1*	*3*	*3*	*15*
Ouvrier	*9*		*3*	*2*	*14*
Profession intermédiaire	*3*	*1*	*2*	*13*	*19*
Retraité	*5*	*2*	*9*	*2*	*18*
Total (2)	***32***	***11***	***27***	***30***	***100***

Dans ce tableau on a :

N=7= nombre de modalités du caractère X : CSP

n=4= nombre de modalités du caractère Y : mode d'hébergement

M=100=nombre total d'individus

Les modalités des caractères sont exclusives : un individu n'a qu'une CSP et un seul mode d'hébergement (il s'agit du mode des dernières vacances)

Les modalités des caractères sont exhaustives (tous les individus sont renseignés).

Dans la matrice Z on voit, par exemple, que 2 agriculteur ont passé leur dernières vacances au camping.

Codification disjonctive du tableau de données

La codification disjonctive consiste à mettre 1 à la modalité que possède l'individu. Ainsi, sur le tableau de données de l'exemple précédent, on obtient :

	CSP=X							Mode d'hébergement=Y			
Individus	Agriculteur	Cadre moyen	Chef d'entreprise	Employé	Ouvrier	Profession intermédiaire	Retraité	Camping	Hôtel	Famille , amis	Location , gîte
1	0	0	1	0	0	0	0	0	1	0	0
2	0	0	0	0	1	0	0	1	0	0	0
3	0	1	0	0	0	0	0	0	0	1	0
4	0	0	0	0	1	0	0	1	0	0	0
5	0	0	0	0	0	1	0	0	0	0	1
6	1	0	0	0	0	0	0	0	0	1	0
7	0	0	0	0	0	1	0	0	0	0	1
8	0	1	0	0	0	0	0	1	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…
100	0	0	0	1	0	0	0	1	0	0	0

On peut vérifier avec cet exemple que :

$Z_{(7,4)} = X'_{(7,100)}\, Y_{(100,4)}$ *où X' est la matrice transposée de* $X_{(100,7))}$ *(Transposé de la Variable CSP du tableau précédent) et* $Y_{(100,4)}$ *la matrice Mode d'hébergement*

➢ L'AFC s'intéresse plus particulièrement aux effectifs marginaux des tableaux que l'on appelle **profils**. Le tableau Z peut être alors transformé selon deux autres tableaux appelés tableaux de profils.

Ainsi, de $Z_{(N,n)}$ on peut déduire deux matrices $X_{(N,n)}$ et $Y_{(N,n)}$:

- $X_{(N,n)}(I) = \left\{\frac{z_{ij}}{z_{i.}}\right\} = \left\{\frac{p_{ij}}{p_{i.}}\right\}$

Fréquences relatives des effectifs lignes

Tableau des pourcentages lignes

Avec $p_{ij} = \frac{z_{ij}}{M}$ et $p_{i.} = \frac{z_{i.}}{M}$

p_{ij} sont les fréquences relatives du tableau (les pourcentages)

On peut alors vérifier que : $\frac{z_{ij}}{z_{i.}} = \frac{z_{ij}/M}{z_{i.}/M} = \frac{p_{ij}}{p_{i.}}$

- $Y_{(N,n)}(J) = \left\{\frac{z_{ij}}{z_{.j}}\right\} = \left\{\frac{p_{ij}}{p_{.j}}\right\}$

Fréquences relatives des effectifs colonnes

Tableau des pourcentages colonnes

Avec $p_{.j} = \frac{z_{.j}}{M}$

Selon l'étude, il faudra choisir le tableau des profils adaptés car ils n'ont pas le même sens économique.

Exemple : à partir du tableau de l'exemple précédent, on peut calculer le tableau de p_{ij} ainsi que les deux matrices X et Y :

CSP\Mode d'hébergement	Camping	Hôtel	Famille , amis	Location , gîte	Total (1)
Agriculteur	2		8	2	12
Cadre moyen	4	2	1	5	12
Chef d'entreprise	1	5	1	3	10
Employé	8	1	3	3	15
Ouvrier	9		3	2	14
Profession intermédiaire	3	1	2	13	19
Retraité	5	2	9	2	18
Total (2)	**32**	**11**	**27**	**30**	**100**

Tableau de pij

CSP\Mode d'hébergement	Camping	Hôtel	Famille , amis	Location , gîte	Somme
Agriculteur	0,02	0	0,08	0,02	0,12
Cadre moyen	0,04	0,02	0,01	0,05	0,12
Chef d'entreprise	0,01	0,05	0,01	0,03	0,1
Employé	0,08	0,01	0,03	0,03	0,15
Ouvrier	0,09	0	0,03	0,02	0,14
Profession intermédiaire	0,03	0,01	0,02	0,13	0,19
Retraité	0,05	0,02	0,09	0,02	0,18
Somme	**0,32**	**0,11**	**0,27**	**0,3**	**1**

Matrice X : profils lignes

CSP\Mode d'hébergement	Camping	Hôtel	Famille , amis	Location , gîte	Somme
Agriculteur	0,1667	0,0000	0,6667	0,1667	1
Cadre moyen	0,3333	0,1667	0,0833	0,4167	1
Chef d'entreprise	0,1000	0,5000	0,1000	0,3000	1
Employé	0,5333	0,0667	0,2000	0,2000	1
Ouvrier	0,6429	0,0000	0,2143	0,1429	1
Profession intermédiaire	0,1579	0,0526	0,1053	0,6842	1
Retraité	0,2778	0,1111	0,5000	0,1111	1
					7

Matrice Y : profils colonnes

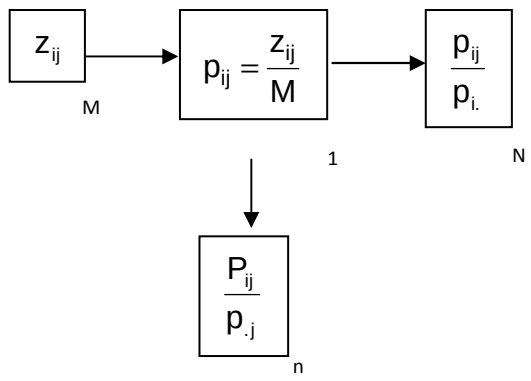
CSP\Mode d'hébergement	Camping	Hôtel	Famille , amis	Location , gîte	
Agriculteur	0,0625	0,0000	0,2963	0,0667	
Cadre moyen	0,1250	0,1818	0,0370	0,1667	
Chef d'entreprise	0,0313	0,4545	0,0370	0,1000	
Employé	0,2500	0,0909	0,1111	0,1000	
Ouvrier	0,2813	0,0000	0,1111	0,0667	
Profession intermédiaire	0,0938	0,0909	0,0741	0,4333	
Retraité	0,1563	0,1818	0,3333	0,0667	
Somme	**1**	**1**	**1**	**1**	4

Le sens économique des matrice X et Y est différent :

En effet, on peut dire à partir de X que 16,67% des agriculteurs vont au camping, et à partir de Y on affirme que 6,25% des personnes allant au camping sont des agriculteurs.

En résumé :

La transformation du tableau $Z_{(N,n)}$ en un tableau $P_{(N,n)}$ est la première étape de l'AFC. On peut alors disposer de deux autres tableaux de sens différents : le tableau des profils lignes $\frac{p_{ij}}{p_{i.}}$ et le tableau des profils colonnes $\frac{p_{ij}}{p_{.j}}$ ce que l'on représente schématiquement par :



2 La transformation initiale des données

Dans un tableau de contingence, les mots individus et variables n'ont pas la même signification que dans le tableau de l'ACP. En effet, dans le tableau de contingence les lignes et les colonnes représentent les modalités de deux caractères. Pour conserver une homogénéïté dans la présentation des deux analyses, on pose par convention que les N modalités du caractère X en lignes portent le nom d'individus et que les n modalités du caractère Y en colonnes portent le nom de variable.

Les individus zij du tableau de contingence sont appelés valeurs pour éviter les confusions.

On observe également que le tableau de départ Z peut être écrit indifféremment avec la modalité X en ligne ou en colonne (même chose pour Y) sans que la nature du tableau soit

modifiée. Par contre, dans ce cas, les tableaux de profil ligne et colonne n'ont plus le même sens (Cf exemple précédent).

L'AFC est une ACP et donc par analogie, à partir de la matrice Z ou de ses transformées en matrices de profils, on peut considérer que l'information contenue dans le tableau peut être analysée à partir de deux espaces :

- L'espace R^n des « variables » (modalités colonnes) dans lequel on peut représenter le nuage des N points « individus » (modalité ligne). Chaque individu a pour coordonnée $x_{ij} = \frac{p_{ij}}{p_{i.}}$ et dans cet espace on utilise le tableau des profils lignes.

Dans R^n, on s'intéresse aux proximités relatives des points individus, c'est-à-dire aux profils lignes, d'où le choix de cette matrice.

	1	…	j	…	n
1					
…					
i			$p_{ij}/p_{i.}$		
…					
N					

Coordonnées du point i dans R^n

- L'espace R^N des « individus » dans lequel on peut représenter les n points variables. Chaque variable a pour coordonnées $y_{ij} = \frac{p_{ij}}{p_{.j}}$. Dans cet espace on utilise le tableau des profils colonne.

Dans R^N, on s'intéresse aux proximités relatives des variables, c'est-à-dire aux profils colonnes :

	1	…	j	…	n
1					
…					
i			$p_{ij}/p_{.j}$		
…					
N					

Coordonnées de j dans R^N

Par analogie avec l'ACP, l'information est donnée par la distance Euclidienne entre les points des nuages des deux espaces R^n et R^N.

<u>Plaçons nous par exemple dans R^n</u>

Calculons la distance euclidienne entre deux points quelconques : x(i) et x(i') de cet espace.

$$d^2(x(i),x(i')) = \sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}}\right)^2$$

En AFC, et contrairement à l'ACP, on n'utilise pas cette distance euclidienne. Plus précisément, on l'utilise mais après avoir effectué une transformation préalable des coordonnées des points du nuages. Dans l'espace R^n cette transformation s'écrit :

$$x_{ij} = \frac{1}{\sqrt{p_{.j}}}\frac{p_{ij}}{p_{i.}}$$

En définitive, dans l'espace R^n on calcule la distance entre deux points x(i) et x(i') par la formule :

$$d^2(x(i),x(i')) = \sum_{j=1}^{n}\left(\frac{1}{\sqrt{p_{.j}}}\frac{p_{ij}}{p_{i.}} - \frac{1}{\sqrt{p_{.j}}}\frac{p_{i'j}}{p_{i'.}}\right)^2 = \sum_{j=1}^{n}\frac{1}{p_{.j}}\left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}}\right)^2$$

On procède de façon équivalente pour l'espace R^N

Considérons dans cet espace deux points du nuage y(j) et y(j')

La transformation : $y_{ij} = \frac{1}{\sqrt{p_{i.}}}\frac{p_{ij}}{p_{.j}}$ conduit à la distance :

$$d^2(y(j),y(j')) = \sum_{i=1}^{N}\left(\frac{1}{\sqrt{p_{i.}}}\frac{p_{ij}}{p_{.j}} - \frac{1}{\sqrt{p_{i.}}}\frac{p_{ij'}}{p_{.j'}}\right)^2 = \sum_{i=1}^{N}\frac{1}{p_{i.}}\left(\frac{p_{ij}}{p_{.j}} - \frac{p_{ij'}}{p_{.j'}}\right)^2$$

En utilisant cette transformation préalable des coordonnées, on applique la distance euclidienne ce qui en d'autres termes revient à écrire que l'on peut effectuer une ACP sur les tableaux aux coordonnées transformées.

On peut vérifier que l'application de la distance euclidienne sur les données transformée est équivalente à l'application de la métrique du χ^2 sur les données non transformées.

Avec l'AFC on peut utiliser le principe **d'équivalence distributionnelle** :

Si on se place par exemple dans R^n et qu'on considère deux points : x(i) et x(i') confondus, on peut les remplacer par un point x(i'') qui aura pour fréquence la somme des fréquences relatives à ces deux points.

On démontre alors que cette substitution des deux points par x(i'') ne modifie pas l'information c'est-à-dire les distances entre les paires de points dans R^n.

Ce principe vaut aussi pour l'espace R^N

3 Détermination des composantes principales dans R^n

3.1 Caractéristiques des « variables » et construction de la matrice d'information

Comme pour l'ACP on se place dans l'espace des variables, on utilise donc la matrice des profils lignes. On vient de voir que dans cet espace les N points du nuage ont pour coordonnées :

$$x_{ij} = \frac{1}{\sqrt{p_{.j}}} \frac{p_{ij}}{p_{i.}}$$

On peut calculer les caractéristiques de ces variables (notées x_j pour j=1 à n) c'est-à-dire la moyenne et la covariance.

La moyenne

$\bar{x}_j = \sum_i p_{i.} x_{ij}$ (moyenne arithmétique pondérée)

$$\bar{x}_j = \sum_i p_{i.} \frac{1}{\sqrt{p_{.j}}} \frac{p_{ij}}{p_{i.}} = \sum_i \frac{p_{ij}}{\sqrt{p_{.j}}} = \frac{1}{\sqrt{p_{.j}}} \sum_i p_{ij} = \frac{1}{\sqrt{p_{.j}}} p_{.j}$$

D'où : $\boxed{\bar{x}_j = \sqrt{p_{.j}}}$

La covariance[1]

La covariance entre deux variables x_j et $x_{j'}$ est :

$$\mathrm{cov}(x_j, x_{j'}) = V_{jj'} = \sum_i p_{i.} \left[\left(\frac{1}{\sqrt{p_{.j}}} \frac{p_{ij}}{p_{i.}} - \sqrt{p_{.j}} \right) \left(\frac{1}{\sqrt{p_{.j'}}} \frac{p_{ij'}}{p_{i.}} - \sqrt{p_{.j'}} \right) \right]$$

D'où : $\boxed{V_{jj'} = \sum_i \frac{p_{ij} p_{ij'}}{\sqrt{p_{.j}} \sqrt{p_{.j'}} \, p_{i.}} - \sqrt{p_{.j}} \sqrt{p_{.j'}}}$

La matrice d'information (matrice des variances covariances des variables)

En faisant varier j et j' de 1 à n on construit alors la matrice V(n,n) des variances/covariances des variables. C'est la matrice d'information des variables et par analogie avec l'ACP, l'étape suivante de l'AFC sera la diagonalisation de cette matrice.

---

[1] Rappelons la formule de la covariance entre deux variables X et Y :

$$\mathrm{COV}(X,Y) = \frac{1}{N} \sum_i \sum_j n_{ij} (x_i - \bar{x})(y_j - \bar{y}) = \sum_i \sum_j f_{ij} (x_i - \bar{x})(y_j - \bar{y}) = \sum_i f_{i.} x_i y_j - \overline{xy} = \sum_j f_{.j} x_i y_j - \overline{xy}$$

Si x=y alors $\mathrm{COV}(x,x) = V(x)$

3.2 Diagonalisation de la matrice des variances–covariances ou de la matrice d'inertie

La matrice V précédente permet de calculer par les valeurs propres et les vecteurs propres normés, la matrice du changement de base. En AFC on n'utilise pas toujours la matrice V mais une matrice plus simple appelée la matrice d'inertie.

Considérons la matrice V(n,n) de terme général $V_{jj'}$

On peut démontrer que le premier vecteur propre noté u_0 issu de cette matrice V a pour coordonnées :

$$u_0 = \begin{bmatrix} u_{01} \\ \dots \\ u_{0i} \\ \dots \\ u_{0n} \end{bmatrix} = \begin{bmatrix} \sqrt{p_{.1}} \\ \dots \\ \sqrt{p_{.i}} \\ \dots \\ \sqrt{p_{.n}} \end{bmatrix} = \bar{x}_j$$

Et qu'il est associé à la première valeur propre $\lambda_0 = 0$ de V

On sait que tous les vecteurs propres sont orthogonnaux deux à deux. Donc le produit scalaire d'un vecteur propre up quelconque de V avec u0 est égal à zéro, ce qui s'écrit :

$$u'_p \times u_0 = 0 \Leftrightarrow$$

$$\begin{bmatrix} u_{p1} & \dots & u_{pj} & \dots & u_{pn} \end{bmatrix} \begin{bmatrix} u_{01} \\ \dots \\ u_{0j} \\ \dots \\ u_{0n} \end{bmatrix} = u_{p1} \times u_{01} + \dots + u_{pi} \times u_{0i} + \dots + u_{pn} \times u_{0n} = 0.$$

$$\Leftrightarrow \sum_{j=1}^{n} u_{pj}.u_{0j} = 0 \Leftrightarrow \sum_{j=1}^{n} u_{pj}\sqrt{p_{.j}} = 0 \quad \mapsto \text{(Relation 1)}$$

Ecrivons que u_p est le vecteur propre associé à la valeur propre λ_p de la matrice V :

$Vu_p = \lambda_p u_p$ (qui est l'écriture de la diagonalisation de la matrice V)

Soit encore :

$$\Leftrightarrow \begin{bmatrix} V_{jj'} \end{bmatrix} \begin{bmatrix} u_{p1} \\ u_{pj} \\ u_{pn} \end{bmatrix} = \lambda_p \begin{bmatrix} u_{p1} \\ u_{pj} \\ u_{pn} \end{bmatrix}$$

Pour le $j^{ième}$ terme :

$$V_{j1}u_{p1} + \dots + V_{jj'}u_{pj'} + \dots + V_{jn}u_{pn} = \lambda_p u_{pj.}$$

$$\Leftrightarrow \sum_{j'=1}^{n} V_{jj'}u_{pj'} = \lambda_p u_{pj.}$$

Remplaçons $V_{jj'}$ par sa valeur :

$$\lambda_p u_{pj} = \sum_{j'=1}^{n}\left(\sum_i \frac{p_{ij}p_{ij'}}{p_{i.}\sqrt{p_{.j}}\sqrt{p_{.j'}}} - \sqrt{p_{.j}}\sqrt{p_{.j'}}\right)u_{pj'.}$$

$$= \sum_{j'}\sum_i \frac{p_{ij}p_{ij'}}{p_{i.}\sqrt{p_{.j}}\sqrt{p_{.j'}}}u_{pj'} - \sum_{j'}\sqrt{p_{.j}}\sqrt{p_{.j'}}u_{pj'}$$

$$= \sum_{j'}\sum_i \frac{p_{ij}p_{ij'}}{p_{i.}\sqrt{p_{.j}}\sqrt{p_{.j'}}}u_{pj'} - \sqrt{p_{.j}}\sum_{j'}\sqrt{p_{.j'}}u_{pj'}$$

Or d'après la relation 1 : $\sum_{j'}\sqrt{p_{.j'}}u_{pj'} = 0$

D'où :

$$\lambda_p u_{pj} = \sum_{j'=1}^{n}\left(\sum_i \frac{p_{ij}p_{ij'}}{p_{i.}\sqrt{p_{.j}}\sqrt{p_{.j'}}}u_{pj'}\right) \quad \text{(Relation 2)}$$

On pose $\sum_i \frac{p_{ij}p_{ij'}}{p_{i.}\sqrt{p_{.j}}\sqrt{p_{.j'}}} = s_{jj'}$ et la relation 2 s'écrit : $\lambda_p u_{pj} = \sum_{j'} s_{jj'}u_{pj'} = \sum_{j'} v_{jj'}u_{pj'}$

Appelons S la matrice de terme $s_{jj'}$ on a alors

$$\lambda_p u_p = Vu_p = Su_p$$

La matrice S porte le nom de matrice d'inertie. On peut constater que les vecteurs propres de la matrice V sont identiques à ceux de S. Il est donc indifférent de diagonaliser la matrice S ou V.

Dans de nombreux logiciels informatique on utilise S qui a une expression plus simple que celle de V. Il existe cependant une différence entre ces deux matrices : en effet, on peut démontrer que la première valeur propre de S est 1 alors que celle de V est 0.

Le premier vecteur propre associé à cette première valeur propre définit un axe principal pour lequel les projections des individus et des variables possèdent une variance (dispersion) nulle. Ce qui signifie que toutes les projections possèdent les mêmes coordonnées.
L'axe factoriel correspondant à cette valeur propre est donc exclu de l'analyse.

3.3 Le choix du nombre de composantes principales

On appelle, en AFC, moment total d'inertie (Mt) du nuage des N individus dans l'espace R^n la somme pondérée des carrés des distances des points du nuage à leur centre de gravité :

$$Mt = \sum_{i=1}^{n} p_{i.}\, d^2\big(X(i), G\big)$$

avec G le centre de gravité du nuage de points et

$$d^2(X(i),G)=\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}}-\sqrt{p_{.j}}\right)^2=\sum_j\left(x_{ij}-\bar{x}_j\right)^2$$

la distance au carrée entre les points du nuage et le centre de gravité.

On constate que ce moment est la variance multidimensionnelle dont on sait (Cf ACP) qu'elle est aussi donnée par la trace de la matrice d'information S ou V.

En définitive on peut écrire que $Mt=\sum_j s_{jj}=tr[S]=\sum_j V_{jj}$

Or la trace de S est égale à la somme des ses valeurs propres λ_j

$$tr[S]=\sum_j \lambda_j$$

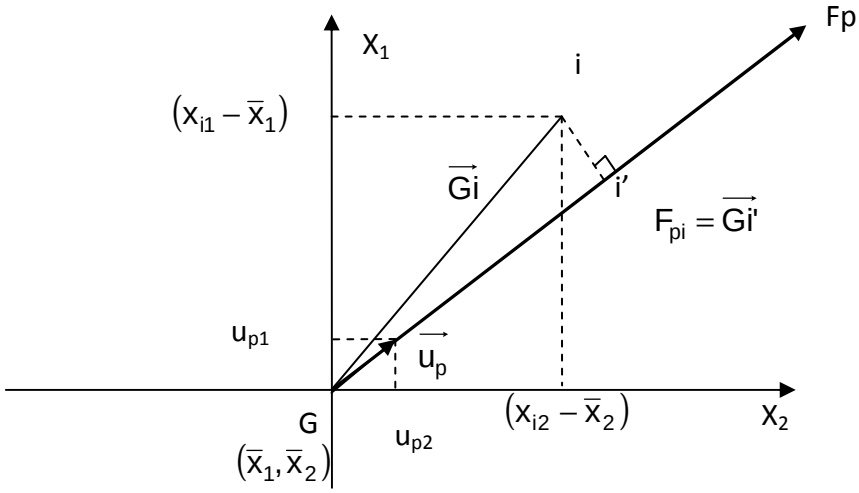
On peut donc, comme en ACP, calculer la part de variance totale expliquée par la j[ième] composante principale.

$$\frac{\lambda_j}{Mt-1}\times 100=\frac{\lambda_j}{tr[S]-1}\times 100=\frac{\lambda_j}{\sum\lambda_j-1}$$ (-1 car la première valeur propre=0 cf supra)

Comme en ACP, l'AFC est réalisable si avec 1, 2 ou au maximum 3 axes principaux on explique près de 70% de la variance totale.

4 Les coordonnées des projections des individus et des variables sur les axes principaux

Contrairement à l'ACP, en AFC les projections sur les axes principaux du nuage des individus et des variables s'effectuent sur un même graphique. On parle de projection simultanée. Pour rappeler comment on réalise la projection d'un point sur un axe factoriel (Cf ACP) considérons, par exemple, un individu i et un axe principal noté F_p et plaçons nous dans un espace à deux dimensions de deux variables x_1 et x_2.



La projection orthogonale du point i sur l'axe Fp est donnée par le produit scalaire :

$$\overrightarrow{Gi'} = \overrightarrow{Gi} \times \overrightarrow{u_p} = \overrightarrow{u_{p'}} \times \overrightarrow{Gi}$$

or ces vecteurs ont les coordonnées suivantes :

$$\overrightarrow{u_p} = \begin{pmatrix} u_{p1} \\ u_{p2} \end{pmatrix} \text{ et } \overrightarrow{Gi} = \begin{pmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_2 \end{pmatrix} \text{ donc}$$

$$\overrightarrow{Gi'} = \overrightarrow{u_{p'}} \times \overrightarrow{Gi} = [u_{p1}, u_{p2}] \times \begin{bmatrix} (x_{i1} - \bar{x}_1) \\ (x_{i2} - \bar{x}_2) \end{bmatrix} = u_{p1}(x_{i1} - \bar{x}_1) + u_{p2}(x_{i2} - \bar{x}_2)$$

D'où $\overrightarrow{Gi'} = F_{pi} = \sum_{j=1}^{2} u_{pj}(x_{ij} - \bar{x}_j)$

En généralisant ce résultat à l'espace complet R^n on a :

$$F_{pi} = \sum_{j=1}^{n} u_{pj}\left(\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} - \sqrt{p_{.j}}\right) = \sum_{j} u_{pj}\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} - \underbrace{\sum_{j} u_{pj}\sqrt{p_{.j}}}_{0}$$ D'après la relation 1

Posons $a_{pj} = \dfrac{u_{pj}}{\sqrt{p_{.j}}}$:

$$\boxed{F_{pi} = \sum_{j} u_{pj}\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} = \sum_{j}\frac{p_{ij}}{p_{i.}}a_{pj}} \quad (A)$$

Cette relation permet de vérifier que les coordonnées de tous les individus sur l'axe principal qui a pour vecteur unitaire u_{0j} $\left(=\sqrt{p_{.j}}\right)$ sont égales à 1. En effet, pour p=0 on a :

$$F_{0i} = \sum_{j} u_{0j} \times \frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} = \sum_{j}\sqrt{p_{.j}}\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} = \frac{1}{p_{i.}}\sum_{j} p_{ij} = \frac{p_{i.}}{p_{i.}} = 1$$

Donc $F_{0i} = 1$ quel que soit i

Dans la formule A, on appelle a_{pj} le rapport $\dfrac{u_{pj}}{\sqrt{p_{.j}}}$

On pourrait vérifier avec les formultes de transition que cette quantité a_{pj} correspond à la projetion orthogonale de la variable j sur l'axe principal p noté dans cet espace a_p

De ce fait, la formule A :

$F_{pi} = \sum_j \frac{p_{ij}}{p_{i.}} a_{pj}$ n'est autre que le calcul du centre de gravité (de la moyenne pondérée) des coordonnées des projections des j variables

D'où la propriété barycentrique de l'AFC :

Les coordonnées des projections orthogonales de chaque point i sont le barycentre (moyenne pondérée) des coordonnées des projections des points j. Et réciproquement, les coordonnées des projections de chaque point j sont le barycentre des projections des points i.

Cette propriété découle des Formules de Transition entre le tableau des profils lignes de l'espace R^n et celui des profils colonnes de l'espace R^N.

Cette propriété barycentrique permet donc d'écrire que :

$\boxed{a_{pj} = \sum_{i=1}^{N} \frac{p_{ij}}{p_{.j}} F_{pi}}$ (B) ou encore $\boxed{F_{pi} = \sum_{j=1}^{n} \frac{p_{ij}}{p_{i.}} a_{pj}}$ (A)

Les Formules de transition montrent alors que la réalisation simultanée des écritures A et B n'est pas possible. Pour que cela le soit, il faut introduire dans les formules précédentes le paramètre $\frac{1}{\sqrt{\lambda_p}}$

Les formules des projections simultanées des variables et des individus s'écrivent alors :

$\hat{F}_{pi} = F_{pi} = \sum_j u_{pj} \frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}}$ ou encore $\hat{F}_{pi} = \frac{1}{\sqrt{\lambda_p}} \sum_{j=1}^{n} \frac{p_{ij}}{p_{i.}} \hat{a}_{pj}$

Et $\hat{a}_{pj} = \sqrt{\lambda_p}\, a_{pj}$ ou encore $\hat{a}_{pj} = \frac{1}{\sqrt{\lambda_p}} \sum_{i=1}^{N} \frac{p_{ij}}{p_{.j}} \hat{F}_{pi}$

Dans $\hat{F}$ et $\hat{a}$, le ^ signifie la valeur calculée.

Par exemple $\hat{F}_{pi}$ (le calcul de la projection de i sur F_p) peut être calculé en utilisant la formule F_{pi} de la relation A ou bien en utilisant la formule A où on connaît $\hat{a}_{pj}$ qui est donné par la deuxième ligne de l'encadré.

Au-delà de la formule, ce qui est important de retenir dans l'AFC, c'est la propriété barycentrique qu'elle révèle. Il s'agit d'une spécificité qui n'existe pas en ACP.

Ainsi on pourra dire par exemple que sur un plan factoriel choisi, un point i est d'autant plus proche d'un point j que le point j (variable) contribue le plus fortement possible au profil de l'individu i.

5 Les aides à l'interprétation

En AFC elles sont identiques à celles de l'ACP.

Les points supplémentaires ou points inactifs

Comme en ACP, il peut arriver qu'un (ou plusieurs) points individus et/ou variables se situent en dehors ou éloigné des autres points. Cela signifie qu'il possède dans le tableau de départ un profil tout à fait spécifique.

Sa (ou ses positions) dans le plan factoriel étant isolée, elle empêche une étude précise des proximités des autres points projetés.

Il est recommandé dans ce cas de rendre ce (ou ces) points inactifs (on le met en supplémentaire), ce qui revient à réaliser l'AFC du tableau de départ en éliminant la ligne ou la colonne qui représente cet individu (ou cette variable).

Ce point possède cependant dans l'espace, des coordonnées, et même s'il ne participe pas à l'AFC, il est alors possible de calculer ses nouvelles coordonnées dans l'espace. On peut donc représenter sur un plan factoriel, ce ou ces points rendus inactifs.

caractéristiques des projections des variables et des individus

	Individus	Variables	
Moyenne	$\overline{F_p} = 0$	$\overline{a_p} = 0$	Donc même centre de gravité, et même origine sur le graphique
variance	$V[F_p] = \lambda p$	$V[a_p] = \lambda p$	
Part de la variance	$\frac{\lambda_p}{tr[S]-1}$	$\frac{\lambda_p}{tr[S]-1}$	Le premier axe factoriel n'est pas utilisé

On peut démontrer facilement les résultats concernant les moyennes (les autres caractéristiques l'ont été dans le cours)

$$\overline{F_p} = \sum_i p_{i.}\left(\sum_j \frac{p_{ij}}{p_{i.}} a_{pj}\right) = \sum_j \sum_i p_{ij} \frac{u_{pj}}{\sqrt{p_{.j}}} = \sum_j \left(\frac{u_{pj}}{\sqrt{p_{.j}}} \sum_i pij\right)$$

$$= \sum_j \frac{u_{pj}}{\sqrt{p_{.j}}} p_{.j} = \underbrace{\sum_j u_{pj}\sqrt{p_{.j}}}_{=0} \text{ et donc } \overline{F_p} = 0$$

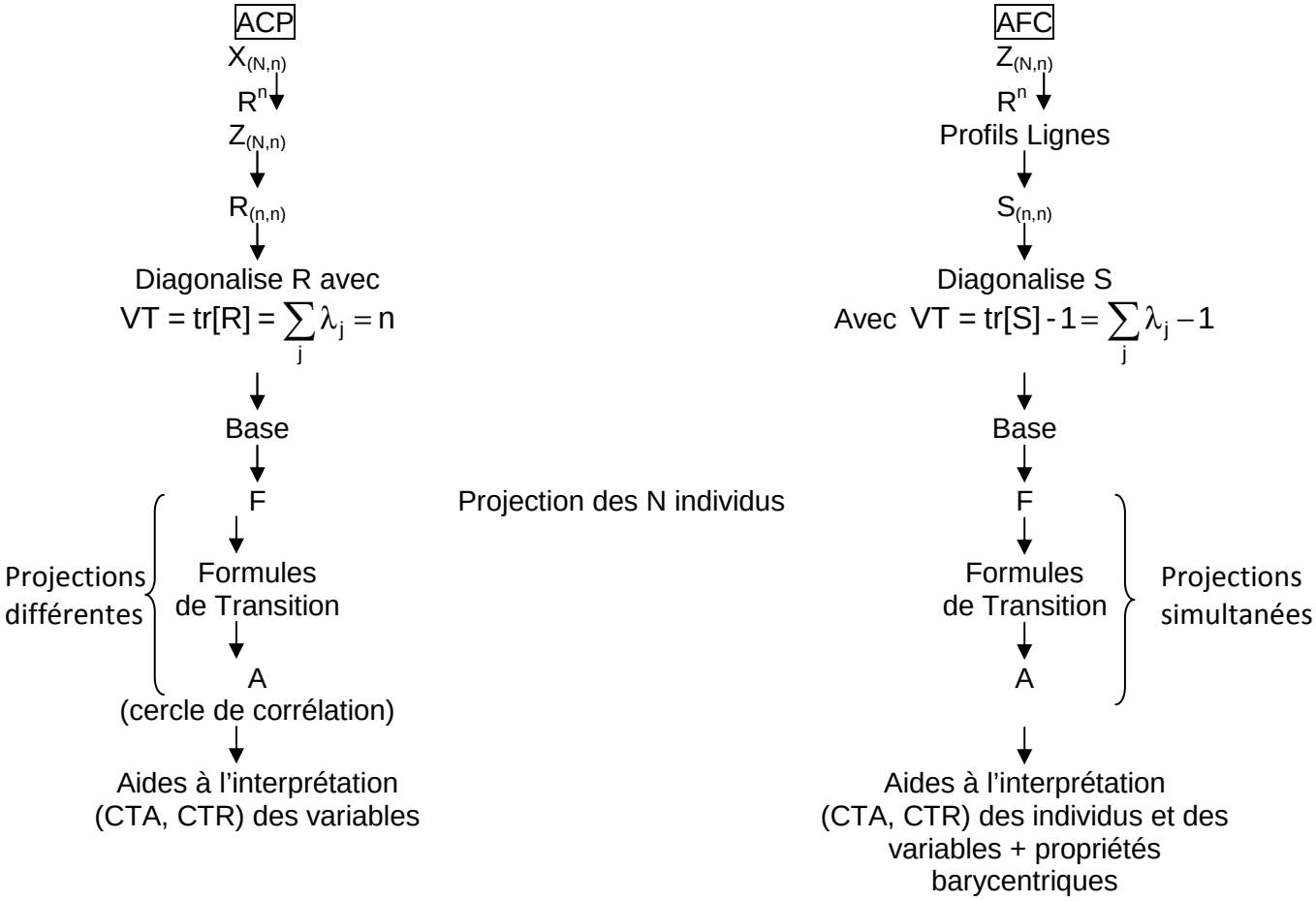
On peut appliquer la même démonstration pour $\overline{a_p} = 0$

Les CTA et CTR

Elles ont les mêmes définitions qu'en ACP mais il faut en l'occurrence les calculer pour les individus et pour les variables.

	CTA	CTR
Individus	$\frac{p_{i.} F_{pi}^2}{\lambda p}$ avec $\lambda p = V[Fp] = \sum_i p_{i.} F_{pi}^2$	$\frac{F_{pi}^2}{d^2(i;G)}$
variables	$\frac{p_{.j} \hat{a}_{pj}^2}{\lambda p} = p_{.j} a_{pj}^2$ avec $V[\hat{a}_p] = \sum_i p_{.j} \hat{a}_{pj}^2 = \lambda p$	$\frac{\hat{a}_{pj}^2}{d^2(j;G)}$
	$\sum CTA = 1$	$\sum CTR = 1$

L'analogie entre les deux méthodes peut être schématisée de la façon suivante :



6 - Introduction à l'AFCM

L'AFC est une méthode factorielle qui ne concerne que deux caractères (2 questions) d'une population de M individus.

Or il arrive fréquement que la population soit caractérisée par plusieurs caractères. Dans ce cas on utilise une extension de l'AFC que l'on appelle l'AFCM (Analyse Factorielle des Correspondances Multiples). Le mot multiple signifiant que l'on dispose de plusieurs caractéristiques sur la population au lieu de 2 pour l'AFC.

Comme il s'agit d'une extension, les concepts utilisés dans l'AFC (comme ceux de l'ACP) sont repris par l'AFCM ; (transformation des données, diagonalisation de la matrice d'information, calcul des composantes principales, calcul des CTA et CTR, Formules de Transition des composantes principales des variables et projections simultanées).

Le tableau de départ est souvent le tableau d'une enquête ou d'un sondage. Il se présente avec en lignes N individus enquêtés et en colonnes n questions posées à ces individus. Chacune de ces questions possède plusieurs modalités de réponses. Le nombre total de modalités est noté M.

6.1 Le Tableau disjonctif complet

Ce tableau d'enquête est écrit sous une forme disjonctive : on affecte le chiffre 1 lorsque l'individu possède la modalité d'une question, 0 sinon. Les modalités de chaque question sont exclusives (un seul 1 par question) et exhaustives (la somme des modalités d'une question=1)

De ce fait, la somme d'une ligne est toujours égale au nombre de questions n. Le tableau disjonctif porte alors le nom de tableau disjonctif complet (TCD) et s'écrit :

$TCD_{(N,M)} =$

	Question J_1			…	Question J_j			…	Question J_n			Marge
	1	…	m_1		1	…m…	m_j		1	…	m_R	$K_{i.}$
1												n
…												n
i						$K_{ijm} = \begin{cases} 1 \\ 0 \end{cases}$						n
…												n
N												n
Marge $K_{.jm}$						$K_{.jm}$						nxN

Ce tableau est tel que :

$K_{ijm} = 1$ si l'individu i possède la modalité m de J_j

$K_{ijm} = 0$ sinon (l'individu i ne possède pas la modalité m de J_j)

$K_{i.} = \sum_{jm} K_{ijm} = n$ par construction

$K_{.jm} = \sum_{i} K_{ijm} =$ le nombre d'individus qui possède la modalité jm de la variable J_j

$n \times N$ représente dans ce tableau l'effectif total

Exemple de passage du tableau d'enquête au tableau disjonctif :

		Sexe	Nationnalité	Couleur Yeux
Individus	1	homme	Français	Bleu
	2	femme	Etranger	Marron
	3	femme	Etranger	Noir
	4	homme	Etranger	Bleu
	5	femme	Français	Marron
	6	homme	Français	Noir
	N	femme	Français	Bleu

Codification (pour la saisie des réponses)		
Sexe	Nationnalité	Couleur Yeux
1	1	1
2	2	2
2	2	3
1	2	1
2	1	2
1	1	3
2	2	1

Tableau disjonctif						
Homme	Femme	Français	Etranger	Yeux bleus	Marron	Noir
1	0	1	0	1	0	0
0	1	0	1	0	1	0
0	1	0	1	0	0	1
1	0	0	1	1	0	0
0	1	1	0	0	1	0
1	0	1	0	0	0	1
0	1	1	0	1	0	0

6.2 Le tableau de Burt

A partir du tableau TDC on peut construire le tableau de Burt :

$BURT_{(M,M)} = TDC'_{(M,N)} \times TDC_{(N,M)}$

Le tableau de Burt est donc le produit matriciel entre la transposée du tableau disjonctif complet et lui même.

Le tableau de Burt est donc une matrice carrée et symétrique qui croise les questions entre elles. Sur sa diagonale principale on trouve le croisement des questions entre elles (le tris à plat) et de part et d'autre de la diagonale principale les croisements entre deux questions distinctes (tris croisés).

Exemple : cet exemple ne concerne que les 6 premiers individus du tableau précédent

Tableau de contingence complet

Homme	Femme	Français	Etrangers	Bleu	Marron	Noir
1	0	1	0	1	0	0
0	1	0	1	0	1	0
0	1	0	1	0	0	1
1	0	0	1	1	0	0
0	1	1	0	0	1	0
1	0	1	0	0	0	1

TDC' (transposé du tableau de contingence)

Homme	1	0	0	1	0	1
Femme	0	1	1	0	1	0
Français	1	0	0	0	1	1
Etrangers	0	1	1	1	0	0
Yeux bleux	1	0	0	1	0	0
Marron	0	1	0	0	1	0
Noir	0	0	1	0	0	1

Tableau de Burt

Homme	Femme	Français	Etrangers	Bleu	Marron	Noir
3	0	2	1	2	0	1
0	3	1	2	0	2	1
2	1	3	0	1	1	1
1	2	0	3	1	1	1
2	0	1	1	2	0	0
0	2	1	1	0	2	0
1	1	1	1	0	0	2

Tris à plat

Tris croisé

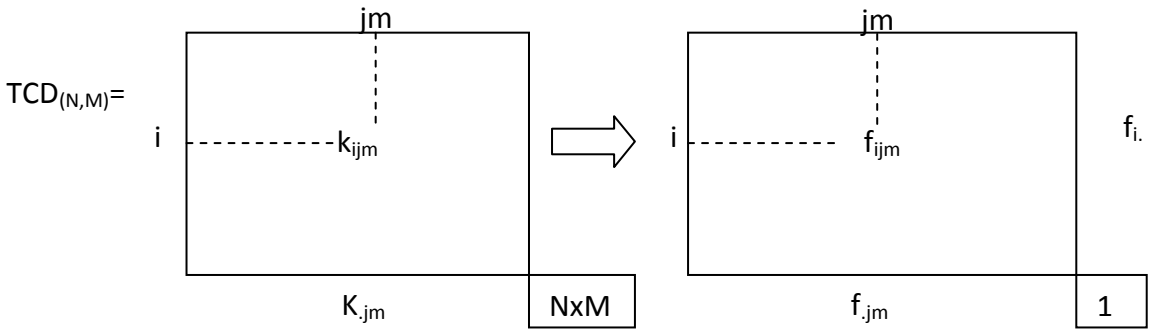
On peut ainsi voir sur l'exemple que :
Tris à plat : Nombre d'hommes=3 ; nombre de femmes=3
Etrangers=3 ; Français=3
Tris croisés : Parmi les hommes, il y a 2 français et 1 étranger
Parmis les femmes il y a 1 Français et 2 étrangers

6.3 Résumé de l'AFCM

L'AFCM, c'est l'AFC du tableau disjonctif complet (TDC). Pour appliquer l'AFC à ce tableau on pose par convention que les lignes du tableau sont les N modalités d'un caractère Y et que les M colonnes sont les modalités d'un autre caractère X.

La démarche suivie par l'AFCM est donc celle de l'AFC en tenant compte des particularités du TDC.

- La première transformation consiste à calculer le tableau des fréquences relatives.



Avec $f_{ijm} = \frac{k_{ijm}}{n \times N}$

Ce tableau présente une particularité par rapport à l'AFC. Les $f_{i.}$ (profils lignes ou distributions marginales) sont tels que :

$$f_{i.} = \sum_{jm} \frac{k_{ijm}}{n \times N} = \frac{1}{n \times N} \sum_{jm} k_{ijn} = \frac{1}{n \times N} k_{i.} = \frac{n}{n \times N} = \frac{1}{N}$$

Donc $\boxed{f_{i.} = \frac{1}{N}}$: la distribution marginale ligne est une constante $= \frac{1}{N}$

De ce fait, le tableau des profils lignes obtenus en divisant par $= \frac{1}{N}$ ne change pas l'information contenue dans le tableau de départ, contrairement à l'AFC.

- Comme pour l'AFC on calcule alors la matrice d'inertie $S_{L(M,M)}$ (matrice carrée de dimension M) et on démontre alors que :

$tr[S_L] = \frac{M}{n}$: c'est-à-dire le nombre moyen de modalités par question

- Comme pour l'AFC on peut alors calculer la part de la variance expliquée par une composante principale :

$$\frac{\lambda_p}{tr[S_L] - 1} \times 100$$

En AFCM les pourcentages de variance totale expliqués sont souvent très faibles. C'est la raison pour laquelle on ne retient arbitrairement que deux, trois ou quatres axes sans trop se préoccuper du pourcentage de variance expliquée.

- On calcule alors comme en AFC les composantes principales retenues (produit scalaire). On calcule aussi les CTA et CTR qui permettent de sélectionner sur le graphique les individus qui participent le plus aux variances des composantes principales.
- Comme en AFC les formules de transition permettent de calculer directement les composantes des modalités puis leur CTA et CTR.

Ces composantes sont assujetties comme dans l'AFC à la propriété barycentrique. Le graphique final rassemble toutes ces projections.

- En AFCM les lignes du tableau disjonctifs sont souvent très nombreuses (beaucoup d'individus) ainsi que les colonnes. Il est donc recommandé d'utiliser des logiciels de classement pour sélectionner les CTA.
- L'AFCM se commente exactement comme une AFC, en tenant compte cependant de la spécificité du tableau de départ qui ne contient que des 0 et des 1 et qui permet de caractériser de façon différente les distances entre les points des nuages.
- L'AFCM possède cependant une particularité : les faisceaux. Lorsque dans le tableau de départ, une question est données avec des modalités qui possèdent un ordre (caractère ordinal), on peut joindre sur le graphique ces nouvelles modalités par une ligne brisée dans l'ordre du tableau. On obtient ainsi un ensemble de lignes briseés que l'on peut insérer dans un faisceau. La forme du faisceau peut alors indiquer s'il existe une relation linéaire ou non linéaire entre les différentes questions du faisceau. La direction des lignes brisées permet aussi d'indiquer le sens de la relation entre les questions.