Ecole Supérieure en Sciences Appliquées de Tlemcen

Analyse de données

Chapitre 5: Analyse Factorielle des Correspondances (AFC) et des Correspondances Multiples (AFCM)

Présentée par:

Dr Imane NEDJAR

Introduction

Analyses de données

Les modèles statistiques

Sont utilisés pour nettoyer les données au début par l'élimination des valeurs aberrantes, et aussi de visualiser les données, afin de construire l'ensemble initial d'exemples.

Les modèles factorielles

Cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l’algèbre linéaire.

Les modèles classification

Analyse Factorielle des Correspondances (AFC)

•**L'analyse des correspondances** (**AFC** ou **ACM**) étudie les proximités entre individus décrits par deux ou plusieurs variables qualitatives ainsi que les proximités entre les modalités de ces variables.

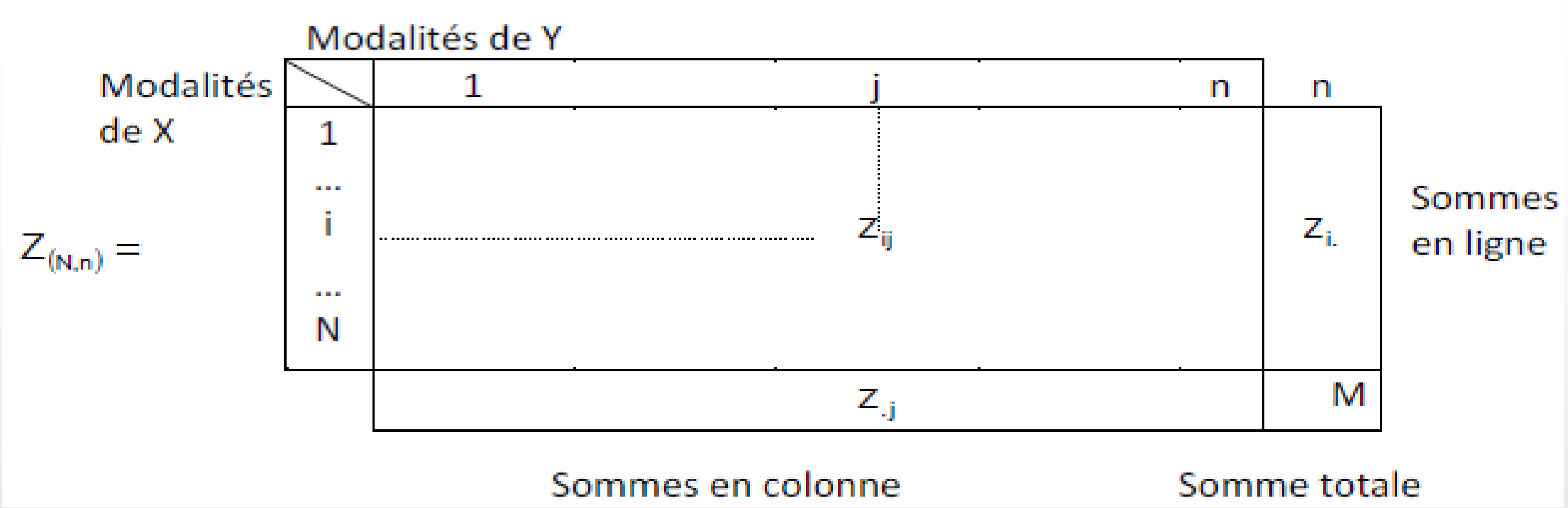
Analyse Factorielle des Correspondances (AFC)

- ACP = étude des liaisons contenues dans un tableau individus*variables, lorsque toutes les variables sont **quantitatives**.
- AFC (Analyse Factorielle des Correspondances) et l'ACM (Analyse des Correspondances Multiples) = étude des liaisons contenues dans un tableau individus*variables, lorsque toutes les variables sont **qualitatives**.
- **L'AFC** est l'étude des correspondances entre les modalités de **deux variables** qualitatives.
- **L'ACM** est une généralisation de l'AFC à **plus de deux variables** qualitatives.

Analyse Factorielle des Correspondances (AFC)

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.



Analyse Factorielle des Correspondances (AFC)

Exemple

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

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, gite
6	Agricuteur	Camping
7	Profession intermédiare	Location, gite
8	Cadre moyen	Campting
…….	…….	….
100	Employé	Hôtel

Analyse Factorielle des Correspondances (AFC)

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

CSP\Mode d'hébergement	Camping	Hôtel	Famille, amis	Location, gite	**Total(1)**
Agriculteur	2	0	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	0	3	2	**14**
Profession intermédiare	3	1	2	13	**19**
Retraité	5	2	9	2	**18**
Total(2)	**32**	**11**	**27**	**30**	**100**

On voit, par exemple, que 2 agriculteur ont passé leur dernières vacances au camping.

Analyse Factorielle des Correspondances (AFC)

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 $Y_{(N,n)}$:

$$Z_{.j} = \sum_{i=1}^{N} Z_{ij} \qquad Z_{i.} = \sum_{j=1}^{n} Z_{ij} \qquad M = \sum_{i=1}^{N} \sum_{j=1}^{n} Z_{ij}$$

$$P_{ij} = \frac{Z_{ij}}{M} \qquad P_{.j} = \frac{Z_{.j}}{M} \qquad P_{i.} = \frac{Z_{i.}}{M}$$

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

Analyse Factorielle des Correspondances (AFC)

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

Fréquences relatives des effectifs lignes

Tableau des pourcentages lignes

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

Fréquences relatives des effectifs colonnes

Tableau des pourcentages colonnes

Analyse Factorielle des Correspondances (AFC)

Exemple:

à partir du tableau de l'exemple précédent, on peut calculer le tableau de pourcentage

$$P_{ij} = \frac{Z_{ij}}{M}$$

CSP\Mode d'hébergement	Camping	Hôtel	Famille, amis	Location, gite	**Total(1)**
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	**0.10**
Employé	0.08	0.01	0.03	0.03	**0.15**
Ouvrier	0.09	0	0.03	0.02	**0.14**
Profession intermédiare	0.03	0.01	0.02	0.13	**0.19**
Retraité	0.05	0.02	0.09	0.02	**0.18**
Total(2)	**0.32**	**0.11**	**0.27**	**0.30**	**1**

Analyse Factorielle des Correspondances (AFC)

Exemple:

à partir du tableau de l'exemple précédent, on peut calculer le tableau de profils lignes (**Matrice X**)

CSP\Mode d'hébergement	Camping	Hôtel	Famille, amis	Location, gite	**Total(1)**
Agriculteur	0,16666667	0	0,66666667	0,16666667	**1**
Cadre moyen	0,33333333	0,16666667	0,08333333	0,41666667	**1**
Chef d'entreprise	0,1	0,5	0,1	0,3	**1**
Employé	0,53333333	0,06666667	0,2	0,2	**1**
Ouvrier	0,64285714	0	0,21428571	0,14285714	**1**
Profession intermédiare	0,15789474	0,05263158	0,10526316	0,68421053	**1**
Retraité	0,27777778	0,11111111	0,5	0,11111111	**1**
Total(2)	**0,32**	**0,11**	**0,27**	**0,3**	**1**

Analyse Factorielle des Correspondances (AFC)

Exemple:

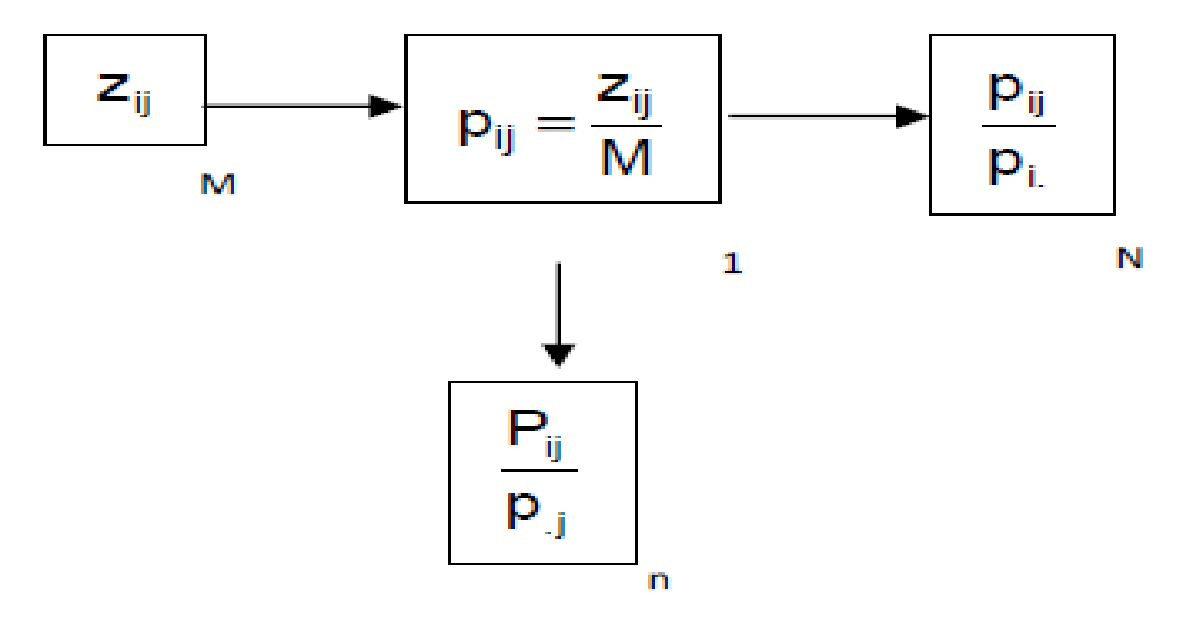
à partir du tableau de l'exemple précédent, on peut calculer le tableau de profils colonnes (**Matrice Y**)

CSP\Mode d'hébergement	Camping	Hôtel	Famille, amis	Location, gite	**Total(1)**
Agriculteur	0,0625	0	0,2962963	0,06666667	**0,12**
Cadre moyen	0,125	0,18181818	0,03703704	0,16666667	**0,12**
Chef d'entreprise	0,03125	0,45454545	0,03703704	0,1	**0,1**
Employé	0,25	0,09090909	0,11111111	0,1	**0,15**
Ouvrier	0,28125	0	0,11111111	0,06666667	**0,14**
Profession intermédiare	0,09375	0,09090909	0,07407407	0,43333333	**0,19**
Retraité	0,15625	0,18181818	0,33333333	0,06666667	**0,18**
Total(2)	**1**	**1**	**1**	**1**	

Analyse Factorielle des Correspondances (AFC)

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.



La transformation initiale des données

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 :

$$Y_{ij} = \frac{P_{ij}}{P_{.j}}$$

L'espace $\mathbb{R}^n$ des « variables » (modalités colonnes) dans lequel on peut représenter le nuage des N points « individus » (modalité ligne). Chaque individu a pour coordonnée

$$X_{ij} = \frac{P_{ij}}{P_{i.}}$$

dans cet espace on utilise le tableau des profils lignes.

Analyse Factorielle des Correspondances (AFC)

$$X_{ij} = \frac{P_{ij}}{P_{i.}}$$ dans cet espace on utilise le tableau des profils lignes.

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

Coordonnées du point i dans R^n

Analyse Factorielle des Correspondances (AFC)

L'espace R^N des «individus » (modalités lignes) dans lequel on peut représenter le nuage des n points «variables» (modalité colonne). Chaque variable a pour coordonnée

$$Y_{ij} = \frac{P_{ij}}{P_{.j}}$$ dans cet espace on utilise le tableau des profils colonnes.

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

Coordonnées de j dans R^N

La transformation initiale des données

L'information est donnée par la distance Euclidienne entre les points des nuages des deux espaces $\mathbf{R^n}$ et $\mathbf{R^N}$

Plaçons nous par exemple dans $\mathbf{R^n}$

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

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

Analyse Factorielle des Correspondances (AFC)

En AFC, 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 $\mathbf{R^n}$ cette transformation s'écrit :

$$X_{ij} = \frac{1}{\sqrt{P_{.j}}} \frac{P_{ij}}{P_{i.}}$$

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

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

Analyse Factorielle des Correspondances (AFC)

On procède de façon équivalente pour l'espace $\mathbf{R^N}$
Considérons dans cet espace deux points du nuage y(j) et y(j')

$$Y_{ij} = \frac{1}{\sqrt{P_{i.}}} \frac{P_{ij}}{P_{.j}}$$

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

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

Détermination des composantes principales dans R^n

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 calcule la moyenne et la covariance de ces variables (notées xj pour j=1 à n)

Détermination des composantes principales dans R^n

On calcule la moyenne et la covariance de ces variables (notées xj pour j=1 à n)

La moyenne

$$\overline{X}_j = \sum_i p_{i.} X_{ij}$$

Moyenne arithmétique pondérée

$$\overline{X}_j = \sum_i p_{i.} \frac{1}{\sqrt{p_{.j}}} \frac{p_{ij}}{p_{i.}} = \sum_i \frac{p_{ij}}{\sqrt{p_{.j}}} = \frac{1}{\sqrt{p_{.j}}} \sum_i p_{ij} = \frac{1}{\sqrt{p_{.j}}} p_{.j}$$

$$\overline{X}_j = \sqrt{p_{.j}}$$

La covariance

La covariance entre deux variables xj et xj' est :

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

$$v_{jj'} = \sum_i \frac{p_{ij}\,p_{ij'}}{\sqrt{p_{.j}}\sqrt{p_{.j'}}\,p_{i.}} - \sqrt{p_{.j}}\sqrt{p_{.j'}}$$

Analyse Factorielle des Correspondances Multiples (AFCM)

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

Or il arrive fréquemment 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, …).

AFCM ou **ACM** = Analyse des Correspondances Multiples

Analyse Factorielle des Correspondances Multiples (AFCM)

•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**.

Analyse Factorielle des Correspondances Multiples (AFCM)

$TCD_{(N,M)} =$

	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

Analyse Factorielle des Correspondances Multiples (AFCM)

•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 (une seul 1 par question) et

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.

Analyse Factorielle des Correspondances Multiples (AFCM)

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

$K_{ijm}=0$ sinon l'individu i ne possède la modalité m de Jm

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

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

n × N représente dans ce tableau l'effectif total

Analyse Factorielle des Correspondances Multiples (AFCM)

	Sexe	Nationalité	Couleur Yeux
1	Homme	Algérien	Bleu
2	Femme	Etranger	Marron
3	Femme	Etranger	Noir
4	Homme	Etranger	Bleu
5	Femme	Algérien	Marron
6	Homme	Algérien	Noir

Sexe	Nationalité	Couleur Yeux
1	1	1
2	2	2
2	2	3
1	2	1
2	1	2
1	1	3

Tableau disjonctif complet (TCD)

Homme	Femme	Algérien	Etranger	Yeux 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

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).

Analyse Factorielle des Correspondances Multiples (AFCM)

Le tableau de Burt

TCD'

Homme	1	0	0	1	0	1
Femme	0	1	1	0	1	0
Algérien	1	0	0	0	1	1
Etranger	0	1	1	1	0	0
Yeux bleu	1	0	0	1	0	0
Marron	0	1	0	0	1	0
Noir	0	0	1	0	0	1

TCD

Homme	Femme	Algérien	Etranger	Yeux 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

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

	Homme	Femme	Algérien	Etranger	Yeux bleu	Marron	Noir
Homme	3	0	2	1	2	0	1
Femme	0	3	1	2	0	2	1
Algérien	2	1	3	0	1	1	1
Etranger	1	2	0	3	1	1	1
Yeux bleu	2	0	1	1	2	0	0
Marron	0	2	1	1	0	2	0
Noir	1	1	1	1	0	0	2

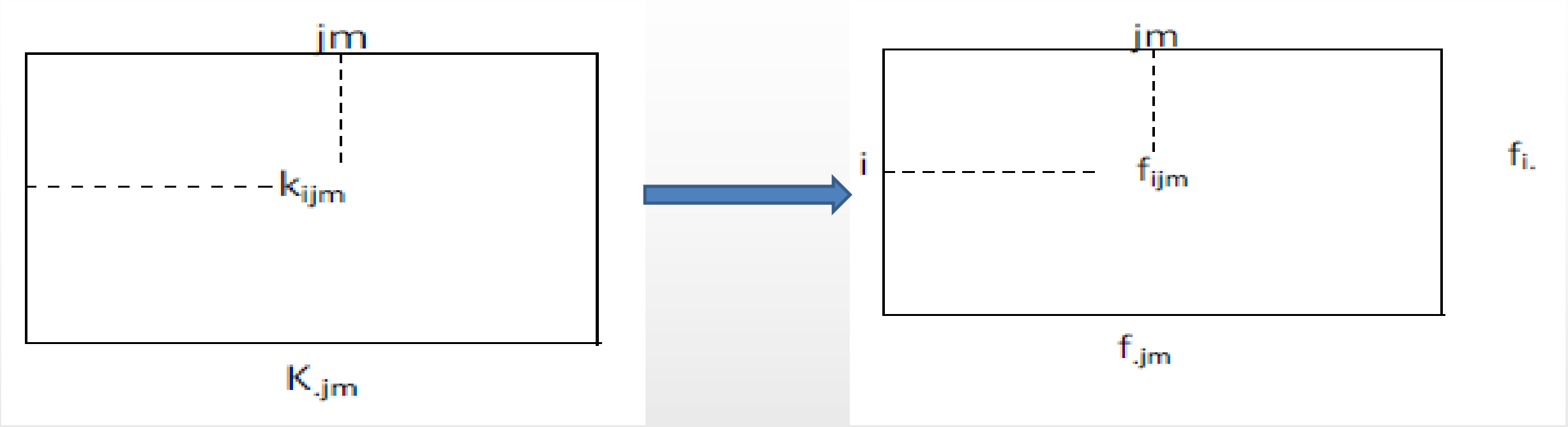
Tris à plat

Tris croisé

Passage de AFCM vers AFC

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



Passage de AFCM vers AFC

$$f_{ijm} = \frac{K_{ijm}}{n \times N}$$

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

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

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

Passage de AFCM vers AFC

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

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

➔Comme pour l'AFC on calcule la part de la variance expliquée par les composante principales