Ministère de l'enseignement supérieur et de la recherche scientifique
Université de JIJEL
Faculté des Sciences Exactes et Informatique
Département d'Informatique

AFC

Analyse Factorielle des Correspondances

M1 SIAD & ILM

Introduction

- L'Analyse Factorielle des Correspondances (A.F.C.) est une méthode factorielle de Statistique Descriptive Multidimensionnelle.
- Son objectif est d'analyser la liaison existant entre deux variables qualitatives (si on dispose de plus de deux variables qualitatives, on aura recours a l'Analyse des Correspondances Multiples.

Introduction

Elle permet d'analyser les informations contenues dans un tableau de contingence. Le but de cette méthode est la réduction de la dimension. L'AFC est une extension de l'analyse en composantes principales (ACP), basée sur la distance du khi-deux.

C’est quoi « les correspondances »?

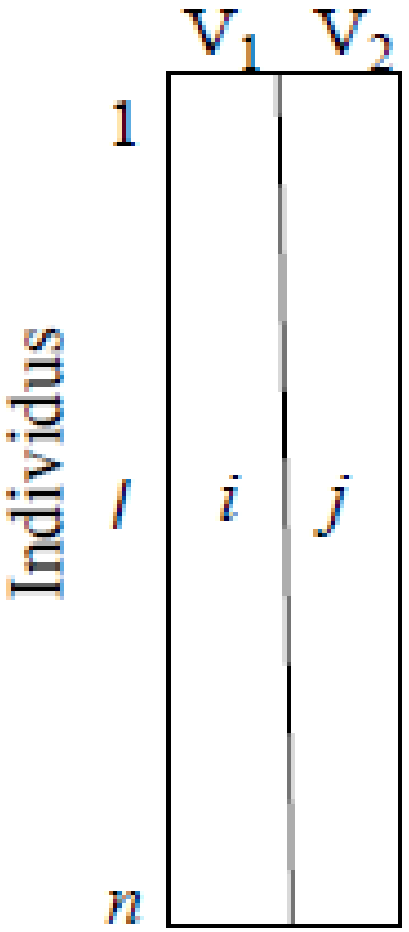
- Lorsque les variables sont **quantitatives, on fait une étude de corrélation.**
- Mais, lorsqu’on a aussi des variables **qualitatives, on doit faire une étude des correspondances.**

- Analyse Factorielle des Correspondances simple ⇒ **deux variables qualitatives**

Les données

- **Tableau de contingence**

  Soient V1 et V2 deux variables qualitatives ou bien catégorielles à p et q catégories (modalités), respectivement, décrivant un ensemble de n individus.



L'individu *l possède:* la modalité *i de V1, la modalité j de V2.*

Les données

- **Tableau de contingence**

  L'AF est basée sur le nuage de points, qu'on l'appelle tableau de contingence, on le note par N*. C'est la matrice des effectifs observés de p lignes et q colonnes.

Exemple : enquête n personnes interrogées deux questions à choix multiples

Les données

- **Tableau de contingence**

  On croisant les deux variables V1 et V2 on obtient:

$$N^* := \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pq} \end{pmatrix} \in \mathcal{M}(p \times q),$$

- x_{ij} : effectif observé: c'est le nombre d'individus possédant la modalité i de V1, la modalité j de V2.

Les données

Exemple:

	V1	V2
I1	A1	B1
I2	A2	B1
I3	A3	B3
I4	A1	B2

Les données

Exemple: le tableau de contingence

	B1	B2	B3
A1	1	1	0
A2	1	0	0
A3	0	0	1

Quelques notations standards

- **Effectif total**: noté par n, c'est la somme de tout les effectifs observés

$$n = \sum_{i=1}^{p} \sum_{j=1}^{q} x_{ij}.$$

- **Effectifs marginales des lignes** noté par Xi, avec

$$X_{i.} = \sum_{j=1}^{q} x_{ij}, \quad i = 1, \ldots, p,$$

- **Effectifs marginales des colonnes** Xj sont donnés par:

$$X_{.j} = \sum_{i=1}^{p} x_{ij}, \quad j = 1, \ldots, q.$$

Quelques notions standards

Exemple:

	B1	B2	B3	Effectifs marginales des lignes
A1	1	1	0	2
A2	1	0	0	1
A3	0	0	1	1
Effectifs marginales des colonnes	2	1	1	

Effectif total: n=4

Tableau des fréquences

Les fréquences sont calculées par la formule suivante:

$$f_{ij} = \frac{\text{Effectif de la Cellule (i, j)}}{\text{Effectif Total}} \quad f_{ij} = \frac{1}{n} x_{ij}$$

Tableau des fréquences

- La matrice des fréquences observées N est représentée comme suit:

$$N = \frac{1}{n} N^* = \begin{pmatrix} f_{11} & \cdots & f_{1q} \\ \vdots & \ddots & \vdots \\ f_{p1} & \cdots & f_{pq} \end{pmatrix} \in \mathcal{M}(p \times q)$$

Tableau des fréquences

Exemple: le tableau des fréquences

	B1	B2	B3
A1	1	1	0
A2	1	0	0
A3	0	0	1

Probabilités marginales

- La somme des fréquences d'une même ligne i; représente le pourcentage global de cette ligne, c'est la fréquence marginale de la modalité i.

$$f_{i.} = \sum_{j=1}^{q} f_{ij} = P(V_1 = i), \ pour\ i = 1, \ldots, p.$$

Probabilités marginales

- La somme des fréquences d'une même colonne j; représente le pourcentage global de cette colonne, c'est la fréquence marginale de la modalité j.

$$f_{.j} = \sum_{i=1}^{p} f_{ij} = P(V_2 = j), \ pour \ j = 1, \ldots, q.$$

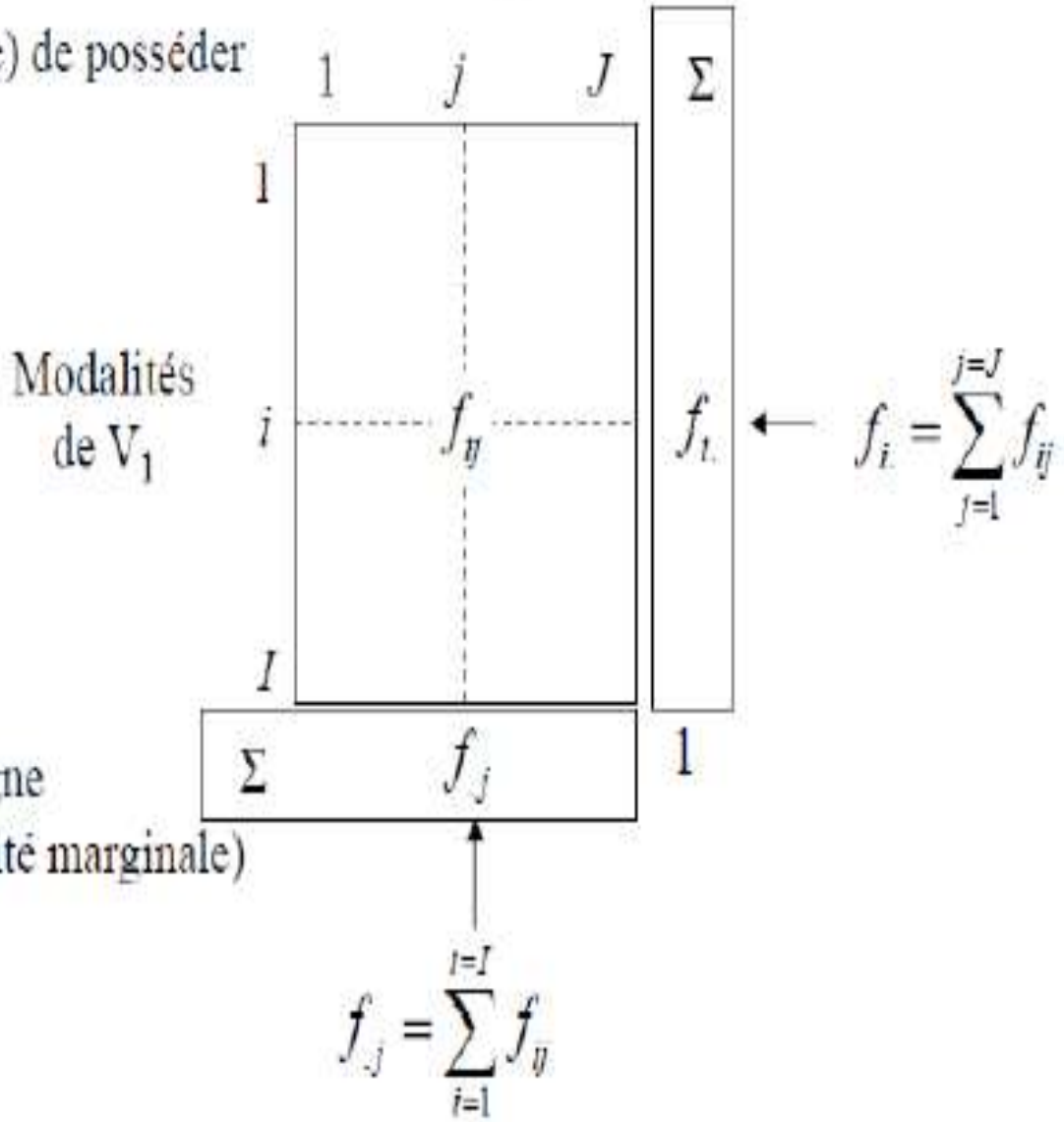
Probabilités marginales

- La somme des distributions marginales toujours égale à l'unité

$$\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = \sum_{i=1}^{p} f_{i\cdot} = \sum_{j=1}^{q} f_{\cdot j} = 1.$$

$$f_{ij} = \frac{x_{ij}}{n}$$

f_{ij} : probabilité (conjointe) de posséder
la modalité i de V_1,
la modalité j de V_2.



Probabilités marginales

Exemple:

	B1	B2	B3	fi.
A1	1/4	1/4	0	2/4
A2	1/4	0	0	1/4
A3	0	0	1/4	1/4
f.j	2/4	1/4	1/4	1

Distributions conditionnelles

- On définit les fréquences conditionnelles aux **profils-lignes $f_{i/j}$** (lire " fréquence de i sachant j "), par

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}}.$$

Distributions conditionnelles

- De même, les fréquences conditionnelles aux **profils-colonnes $f_{j/i}$** (lire"fréquencede j sachant i ")

$$f_{j/i} = \frac{f_{ij}}{f_{i\cdot}}$$

Distributions conditionnelles

- On définit la fréquence théorique $\overline{f}_{ij}$ par:

$$\overline{f}_{ij} = f_{i\cdot} f_{\cdot j}$$

Distributions conditionnelles

- On a aussi

$$\sum_{j=1}^{q} f_{j/i} = 1, \ \ pour\ i = 1, \ldots, p,$$

$$\sum_{i=1}^{p} f_{i/j} = 1, \ \ pour\ j = 1, \ldots, q.$$

Probabilités marginales

Exemple: Matrice des profils lignes

$P_L =$

	B1	B2	B3
A1	1/2	1/2	0
A2	1	0	0
A3	0	0	1

Matrice des profils colonnes

$P_C =$

	B1	B2	B3
A1	1/2	1	0
A2	1/2	0	0
A3	0	0	1

- Nuage des profils lignes

$$N_I := \{(f_{i1}/f_{i\bullet}, ..., f_{iJ}/f_{i\bullet}), i = 1, ..., I\} \subset \mathbb{R}^J.$$

On attribue à chaque ligne le poids $p_i = f_{i\bullet}$, point moyen : $G_I = (f_{\bullet 1}, ..., f_{\bullet J})$.

- Nuage des profils colonnes

$$N_J := \{(f_{1j}/f_{\bullet j}, ..., f_{Ij}/f_{\bullet j}), j = 1, ..., J\} \subset \mathbb{R}^I.$$

On attribue à chaque colonne le poids $p_j = f_{\bullet j}$, point moyen : $G_J = (f_{1\bullet}, ..., f_{I\bullet})$.

Exemple:

	B1	B2	B3	fi.
A1	1/4	1/4	0	2/4
A2	1/4	0	0	1/4
A3	0	0	1/4	1/4
f.j	2/4	1/4	1/4	

GI=(2/4, 1/4, 1/4)

GJ=(2/4, 1/4, 1/4)

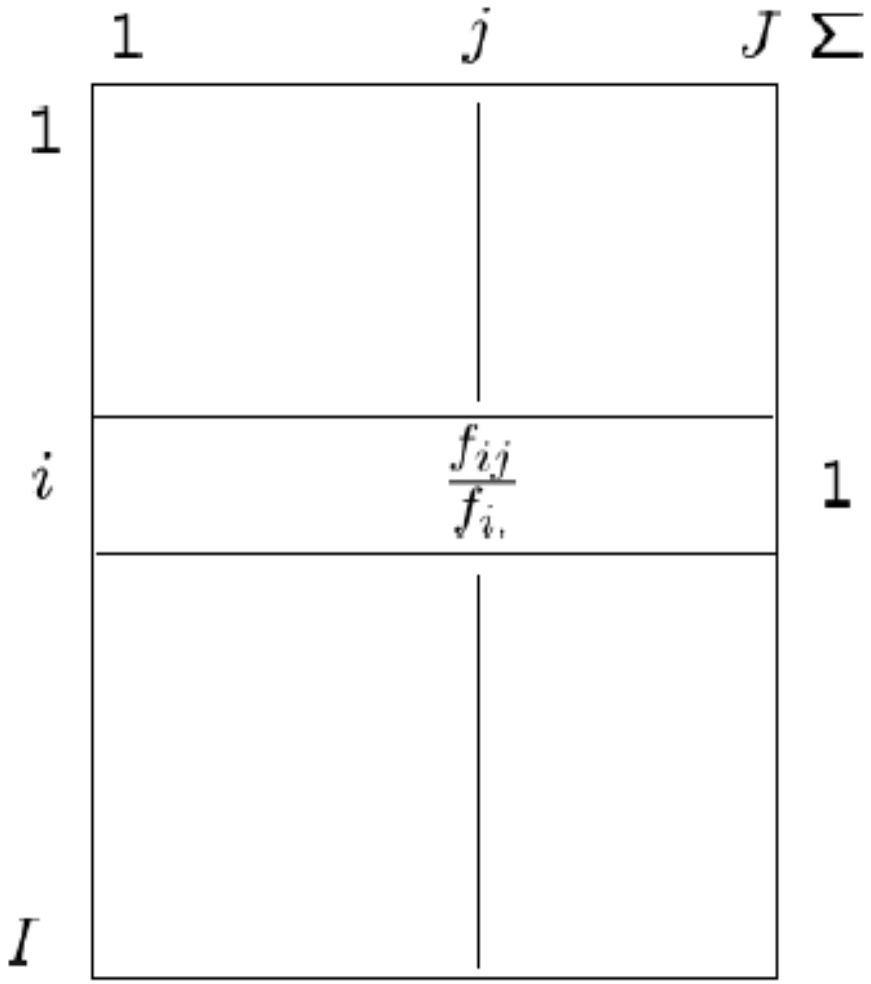
Liaison entre les variables

L'AF a pour but d'étudier la liaison entre les deux variables (V1 et V2), dite encore correspondance. Lorsque on étudie un tableau de contingence (une population de n individus, à travers de ces variables qualitatives), on s'intéresse à l'indépendance de ces deux variables.

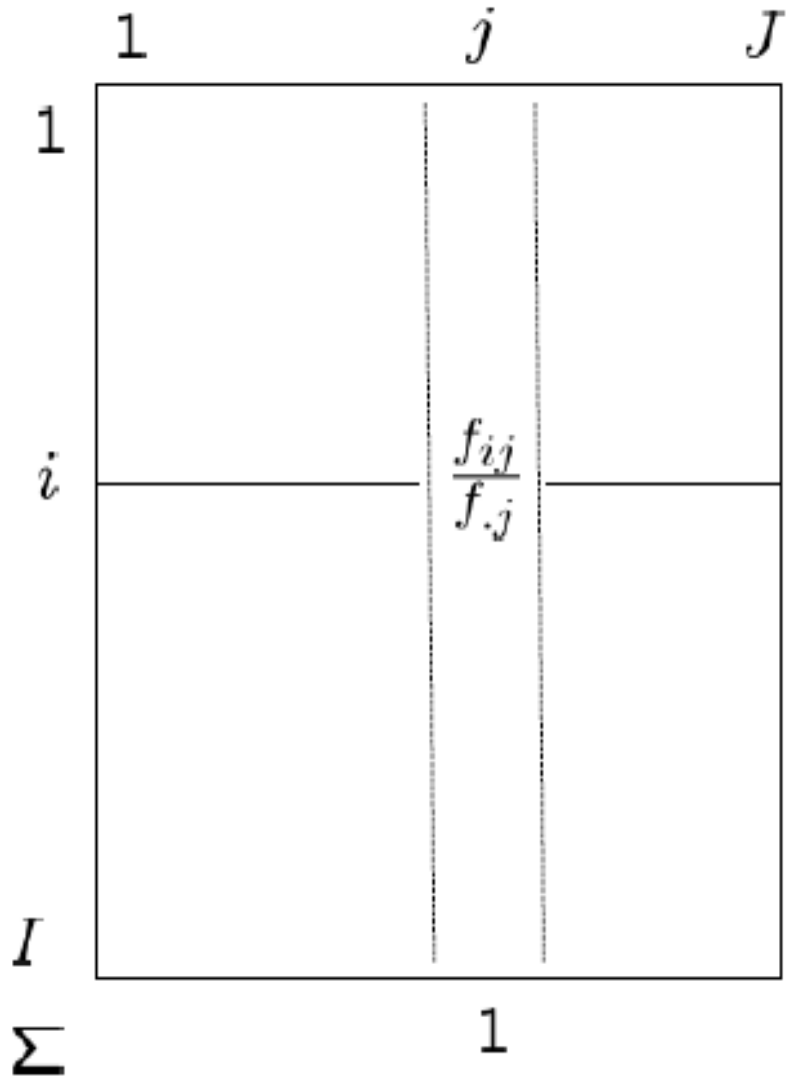
Liaison entre les variables

Si ces variables sont indépendantes alors, l'AFC n'a aucun sens, pour cela, il est classique d'étudier la significativité de la liaison entre les lignes et les colonnes.

Profils lignes et Profils colonnes



Profils lignes



Profils colonnes

Profils lignes

Distance entre les lignes i et l:

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

Métrique: $\left(\frac{1}{f_{.j}} \right)_{j=1,\ldots,J}$

Centre de gravité: $\sum_{i=1}^{I} f_{i.} \times \left(\frac{f_{ij}}{f_{i.}} \right) = (f_{.j})_{j=1,\ldots,J}$

Profils lignes

Distance entre la ligne i et le centre de gravité G:

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

Profils colonnes

Distance entre les colonnes j et k:

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

Centre de gravité:

$$\sum_{j=1}^{J} f_{.j} \times \left(\frac{f_{ij}}{f_{.j}} \right) = (f_{i.})_{i=1,\ldots,I}$$

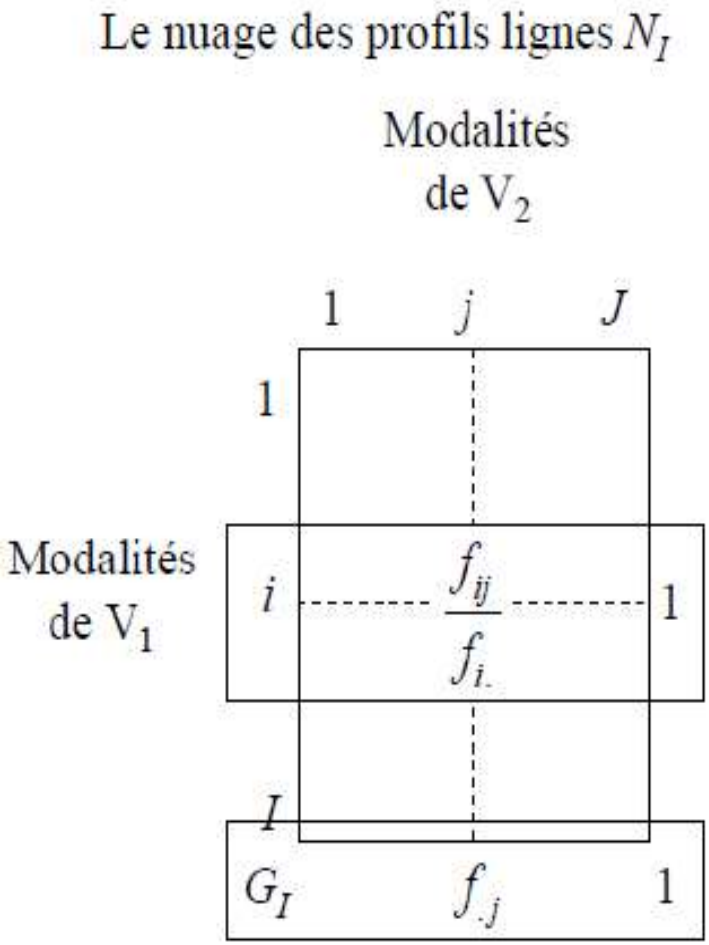
Profils colonnes

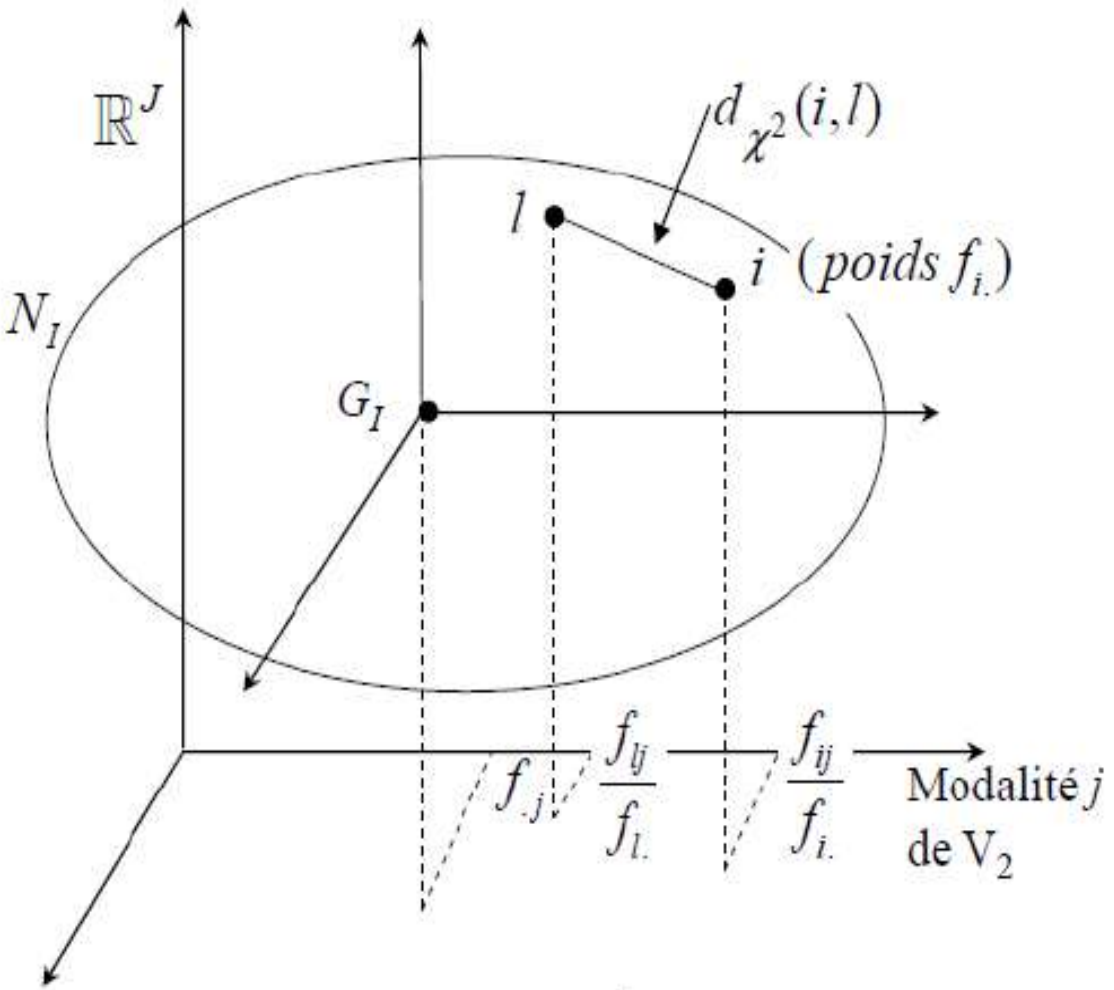
Distance entre la colonnes j et le centre de gravité G:

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

Nuage des Profils lignes

Le nuage des profils lignes N_I

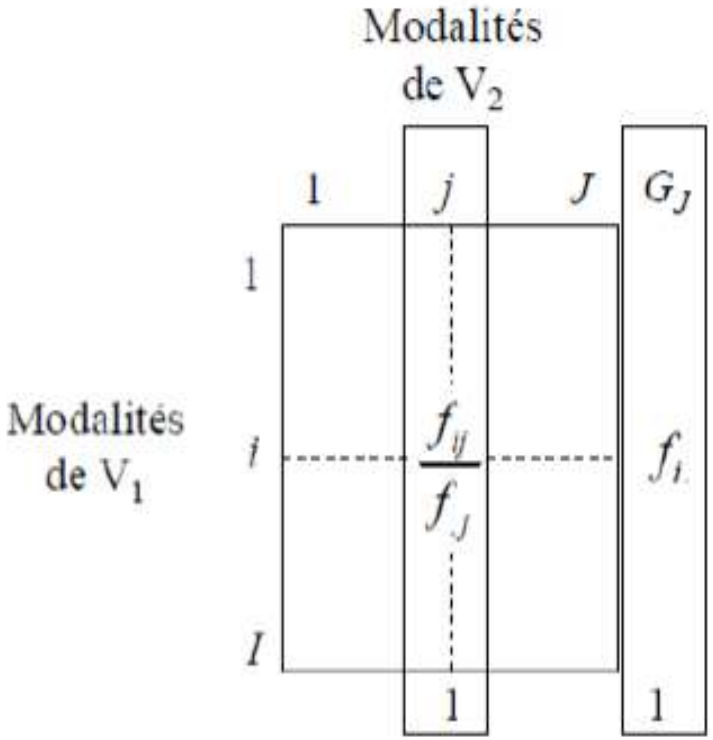


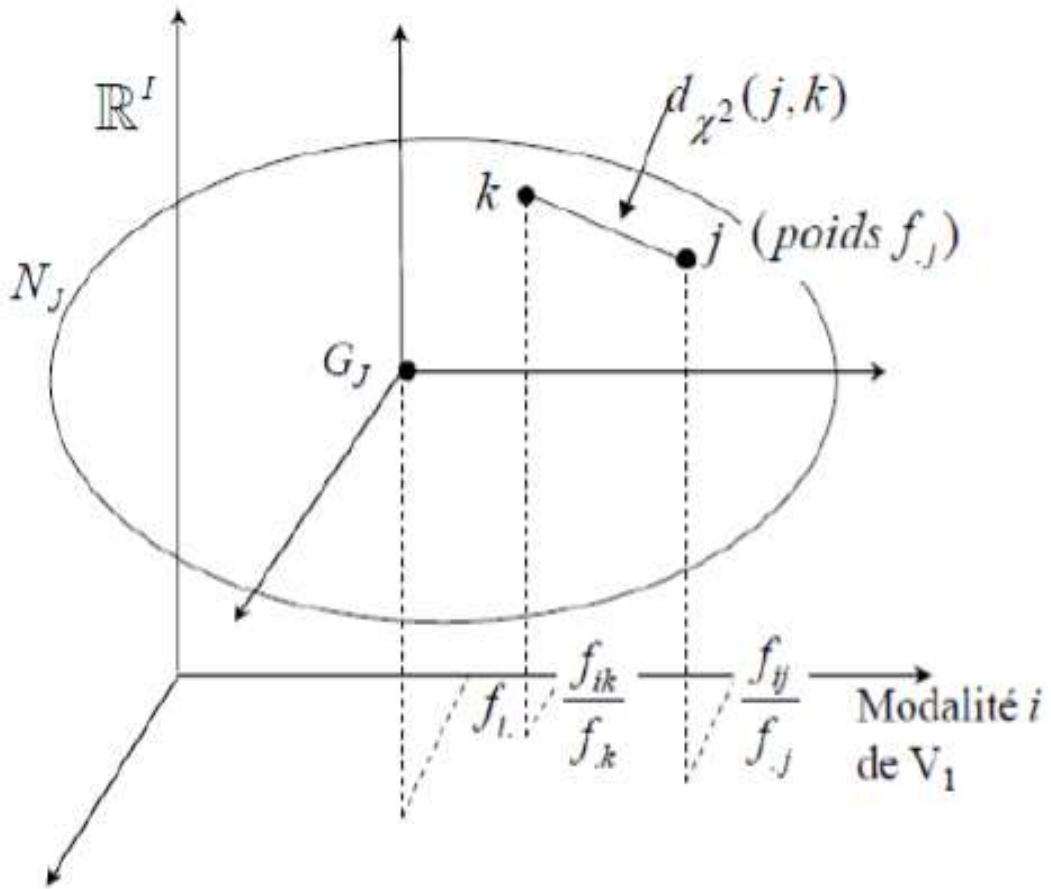


Distance du χ^2 : $d^2_{\chi^2}(i,l) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2$

Nuage des Profils colonnes

Le nuage des profils colonnes N_J





Distance du χ^2: $$d^2_{\chi^2}(j,k) = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2$$

L’Analyse Factorielle des Correspondances

- Méthode d’analyse de tables de contingences en termes de profils. (Lignes ou colonnes).
- **Objectif :Visualiser dans des plans factoriels, les nuages *N(I) et N(J)des points profils.***
- La notion de proximité est basée sur la distance du $\chi 2$.

L'Analyse Factorielle des Correspondances

- Effectuer deux ACP, celle du triplet **(X=Pl, M= D_J^{-1}, D = D_I)** puis celle du triplet **(X=^tPc, M = D_I^{-1}, D = D_J)** dont les composantes principales fourniront respectivement les représentations en projection de ***N(I)*** et ***N(J).***

Métriques associées à P

- L'espace R^I des colonnes de P est muni de la métrique diagonale des poids statistiques de la marge ligne,

$$\mathbf{D_I = diag(f_{1.}, \ldots, f_{I.})}$$

- L'espace R^J des lignes de P est muni de la métrique diagonale des poids statistiques de la marge colonne,

$$\mathbf{D_J = diag(f_{.1}, \ldots, f_{.J})}$$