Université Paul Sabatier
TOULOUSE III

BECNM – Analyses de Données Multivariées

Analyse Factorielle des Correspondances (AFC)

- *Contingency table analysis*
- *RQ-technique*
- *Reciprocal averaging*
- *Reciprocal ordering*
- *Correspondence analysis*

Gaël Grenouillet
gael.grenouillet@univ-tlse3.fr

Introduction

Quelques dates

On utilise seulement des variables positives.

- **A l'origine** :

  Conçue pour étudier des
  tableaux de contingence
  = tableaux d'effectifs (comptages)
  croisant les modalités de 2 variables

- **1940** – Guttman
  Fisher
- **1964** – Benzécri

- Application en **écologie** :
  ex : tableaux **espèces** x **échantillons**

- **1971** – Hatheway
- **1973** – Hill

- **Généralisée** à d'autres types de données
  (condition : valeurs positives)

- **1970-80** – Benzécri et al.

Domaine d'application

Tableau de contingence

Analyse les liens entre **variables qualitatives**

		Variable 1				
		Modalité 1	***Modalité 2***	***Modalité 3***	***…***	***Modalité p***
Variable 2	***Modalité 1***	eff_{11}	eff_{12}	eff_{13}		eff_{1p}
	Modalité 2	eff_{21}	eff_{22}			
	Modalité 3	eff_{31}	eff_{32}			
	…	…	…			
	Modalité n	eff_{n1}	eff_{n2}			eff_{np}

Exemple : couleur des yeux, profession, classe d'âge, …

Domaine d'application

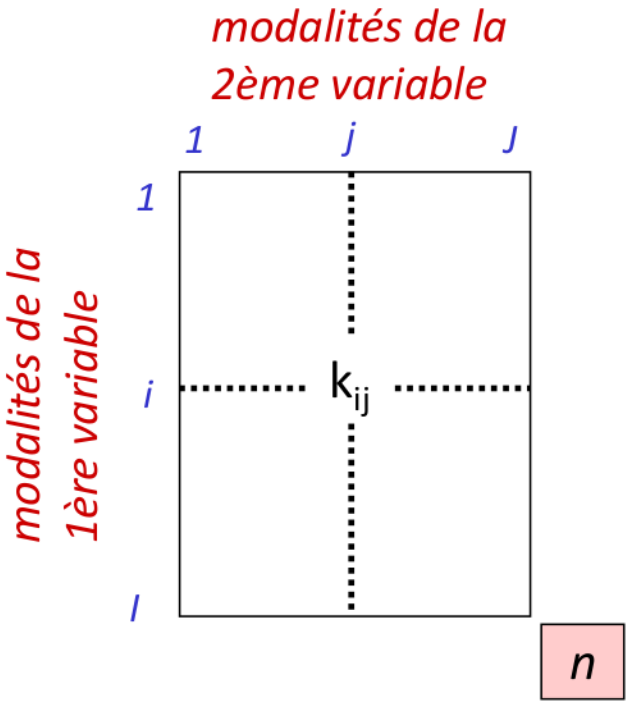
Tableau de contingence

Analyse les liens entre **variables qualitatives**

		Variable 1				
		Modalité 1	***Modalité 2***	***Modalité 3***	***…***	***Modalité p***
Variable 2	***Modalité 1***	eff_{11}	eff_{12}	eff_{13}		eff_{1p}
	Modalité 2	eff_{21}	eff_{22}			
	Modalité 3	eff_{31}	eff_{32}			
	…	…	…			
	Modalité n	eff_{n1}	eff_{n2}			eff_{np}

Exemple : couleur des yeux, profession, classe d'âge, …

Tableau de contingence



I : nombre de lignes

J : nombre de colonnes

k_{ij} : **nombre d'individus** possédant à la fois la modalité **i** de la 1ère variable et la modalité **j** de la 2ème variable

$$\sum_i \sum_j k_{ij} = n$$ (nombre total d'individus)

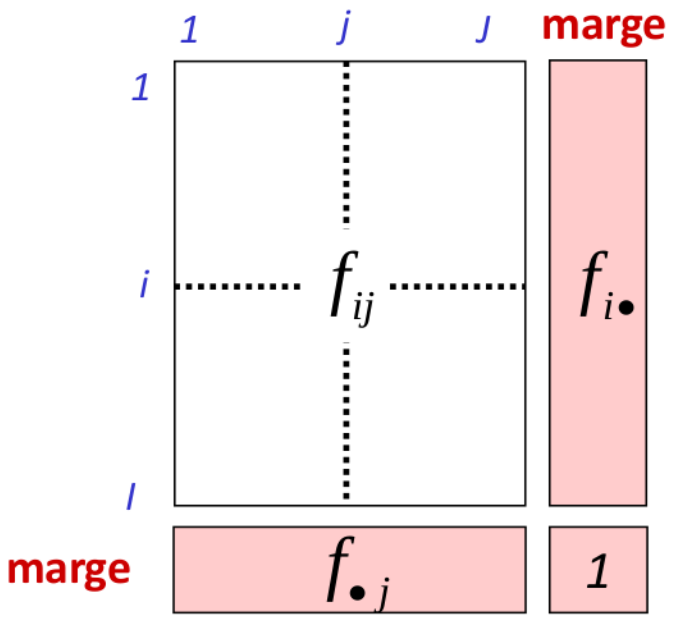
Analyse du tableau de contingence :

Ce ne sont pas les effectifs bruts qui nous intéressent mais les répartitions en % à l'intérieur d'une ligne ou d'une colonne

On parle de **profils-lignes** et de **profils-colonnes**

Tableau de contingence

Transformation → Tableau des fréquences relatives
définit une mesure de **probabilité**



$$f_{ij} = k_{ij} / n$$

$$f_{i\bullet} = \sum_j f_{ij}$$

$$f_{\bullet j} = \sum_i f_{ij}$$

$$\sum_i f_{i\bullet} = \sum_j f_{\bullet j} = \sum f_{ij} = 1$$

$f_{i\bullet}$ = profil-colonne moyen

$f_{\bullet j}$ = profil-ligne moyen

Tableau de contingence

- Il y a **indépendance** entre les deux variables si $f_{ij} = f_{i\bullet} f_{\bullet j}$

$\Leftrightarrow$

- toutes les **lignes** sont proportionnelles $\frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$
- toutes les **colonnes** sont proportionnelles $\frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$

- Il y a **liaison** entre les deux variables lorsque certaines cases f_{ij} diffèrent du produit $f_{i\bullet} f_{\bullet j}$

$f_{ij} > f_{i\bullet} f_{\bullet j}$ modalités i et j s'associent plus qu'elles ne le feraient sous l'hypothèse d'indépendance (H_0)
i et j s'attirent

$f_{ij} < f_{i\bullet} f_{\bullet j}$ modalités i et j s'associent moins que sous H_0
répulsion entre les deux modalités

Objectifs de l'AFC

Objectif fondamental

Etudier la liaison entre 2 variables

= étudier la proximité entre chaque profil et son profil moyen

= étudier l'écart du tableau à l'hypothèse d'indépendance

L'**AFC** cherche à obtenir une typologie des lignes, une typologie des colonnes, et relier ces deux typologies entre elles

L'**AFC** est une analyse factorielle → réduire la dimension des données en conservant le plus d'information possible

Originalité

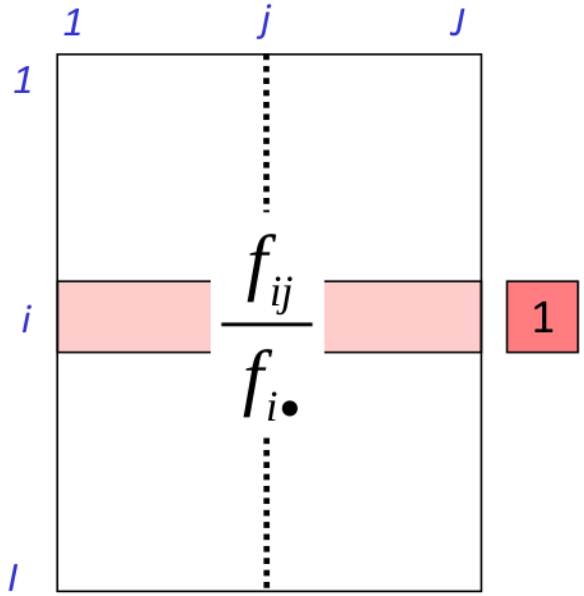
La notion de ressemblance entre 2 lignes ou entre 2 colonnes est différente de celle de l'ACP :

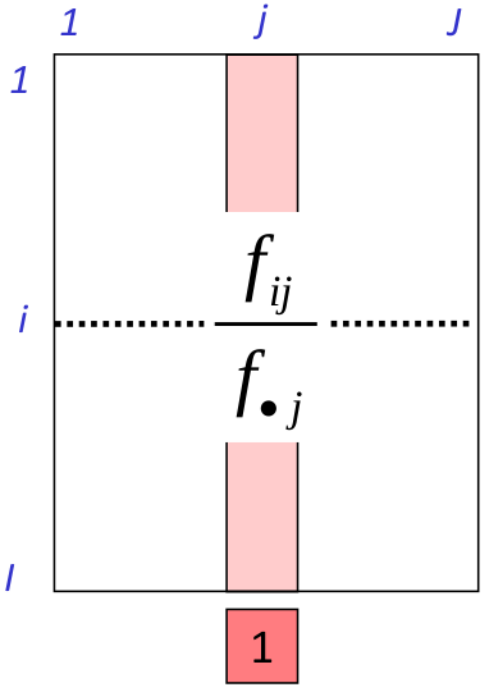
Les lignes et les colonnes jouent un rôle absolument symétrique

Principe

Transformations des données en profils

Profil-ligne





$\frac{f_{ij}}{f_{i\bullet}}$ = probabilité d'avoir la modalité j sachant que l'on a la modalité i (probabilité conditionnelle)

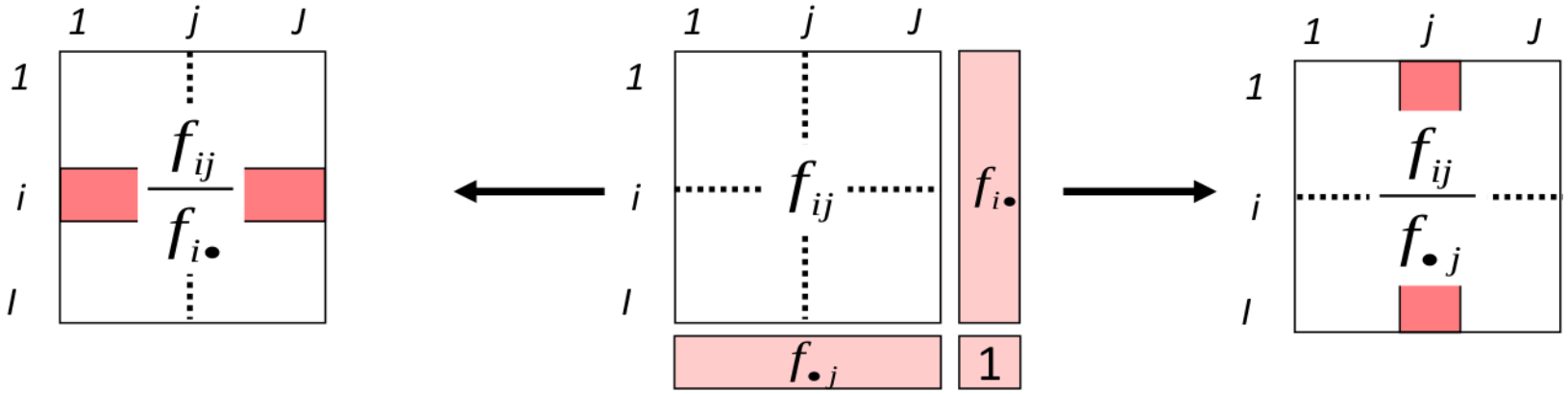
Profil-colonne

➡ Selon que l'on s'intéresse aux lignes ou aux colonnes, on ne considère pas le même tableau transformé

Schéma général

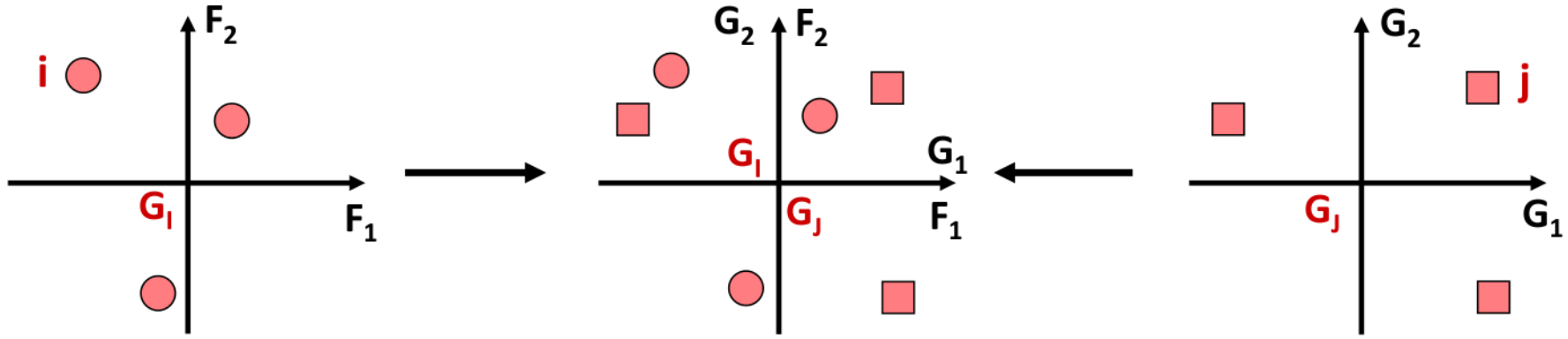
Tableau de contingence

Profils-lignes | **Fréquences relatives** | **Profils-colonnes**



I points dans R^J | **J points dans R^I**

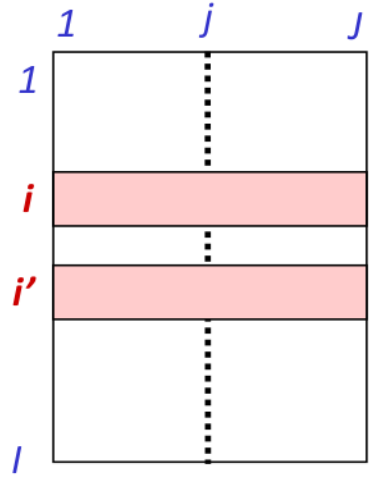
relations de transition

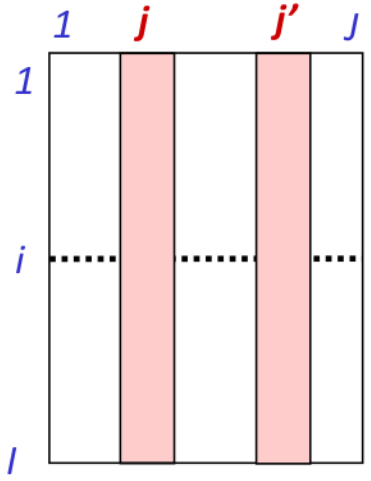


Principe

Ressemblance entre profils

La ressemblance entre deux lignes ou entre deux colonnes est définie par une **distance** entre leurs **profils** : **la distance du χ^2**





distance entre 2 profils-lignes

$$d^2(i,i') = \sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2$$

distance entre 2 profils-colonnes

$$d^2(j,j') = \sum_i \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2$$

Principe

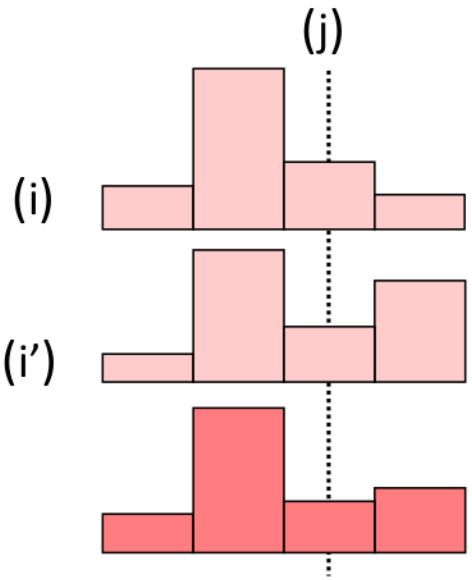
Ressemblance entre profils

La **distance du χ^2** est une distance **pondérée**

La **pondération** $\frac{1}{f_{\bullet j}}$ équilibre l'influence des colonnes sur la distance entre les lignes

$\frac{1}{f_{i\bullet}}$ équilibre l'influence des lignes sur la distance entre les colonnes

Ex de pondération



$f_{\bullet j}$ **poids** des colonnes

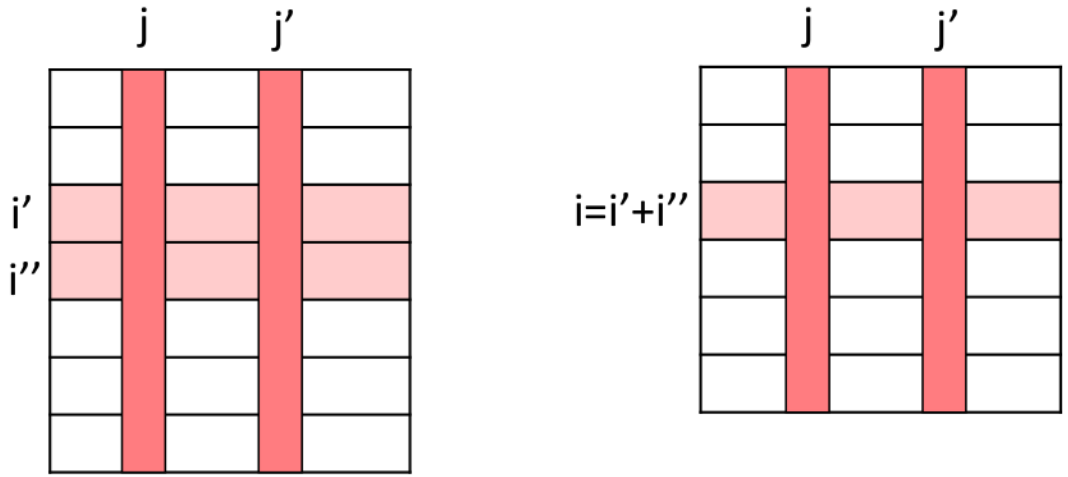
Principe

Ressemblance entre profils

Propriété fondamentale de la distance du χ^2

équivalence distributionnelle

Si deux modalités d'une variable présentent des profils identiques, il est possible de les agréger en une nouvelle modalité (affectée de la somme de leurs poids) sans modifier les distances entre les modalités de l'autre variable



$$d^2(j, j') \longleftrightarrow d^2(j, j')$$

Intérêt : assure la robustesse des résultats vis à vis de l'arbitraire du découpage en modalités des variables qualitatives

Principe

Géométrie des nuages

N_I

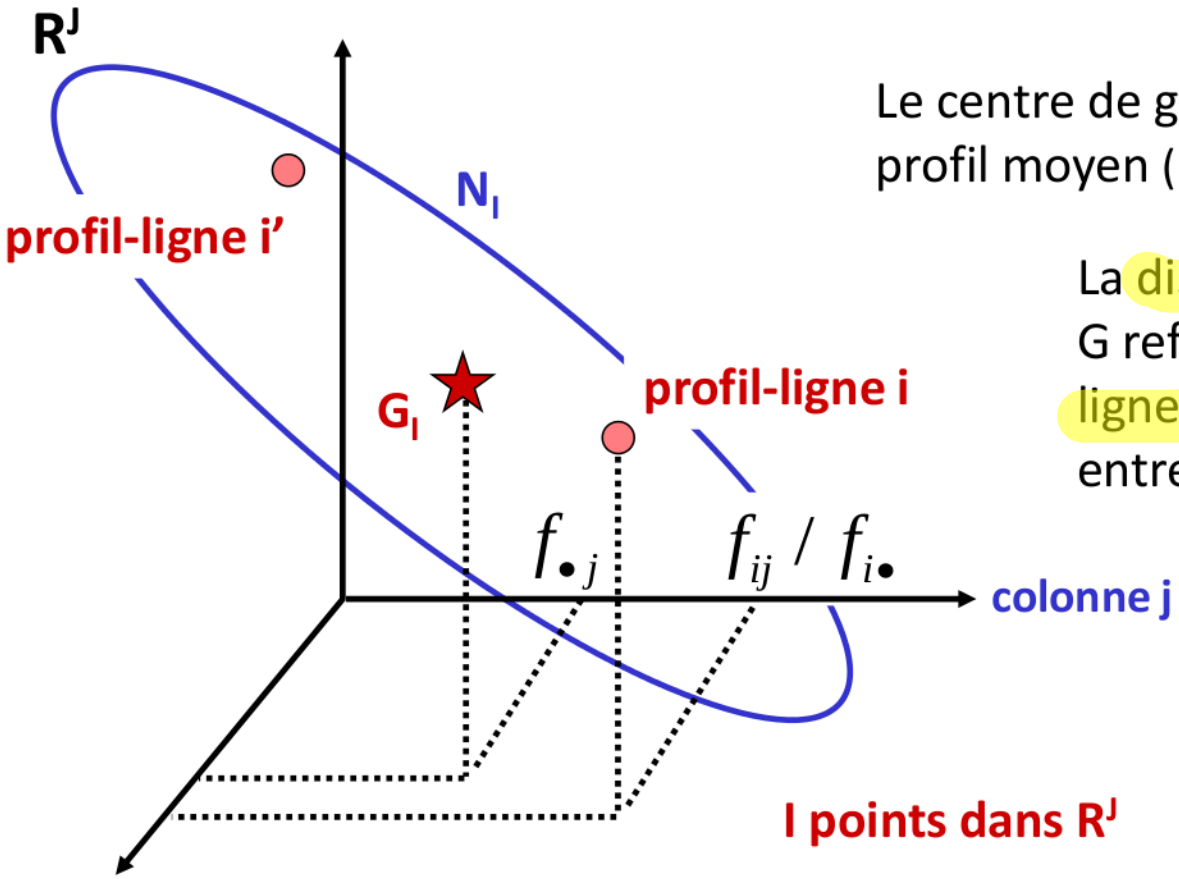
Nuage des profils-lignes

Les poids sont imposés
Le poids du point (profil-ligne) i = $f_{i\bullet}$

(*proportionnel à l'effectif de la modalité*)

Le centre de gravité **G** correspond au profil moyen (référence)

La dispersion du nuage autour de G reflète l'écart entre les profils-lignes et la marge, donc la liaison entre les 2 variables



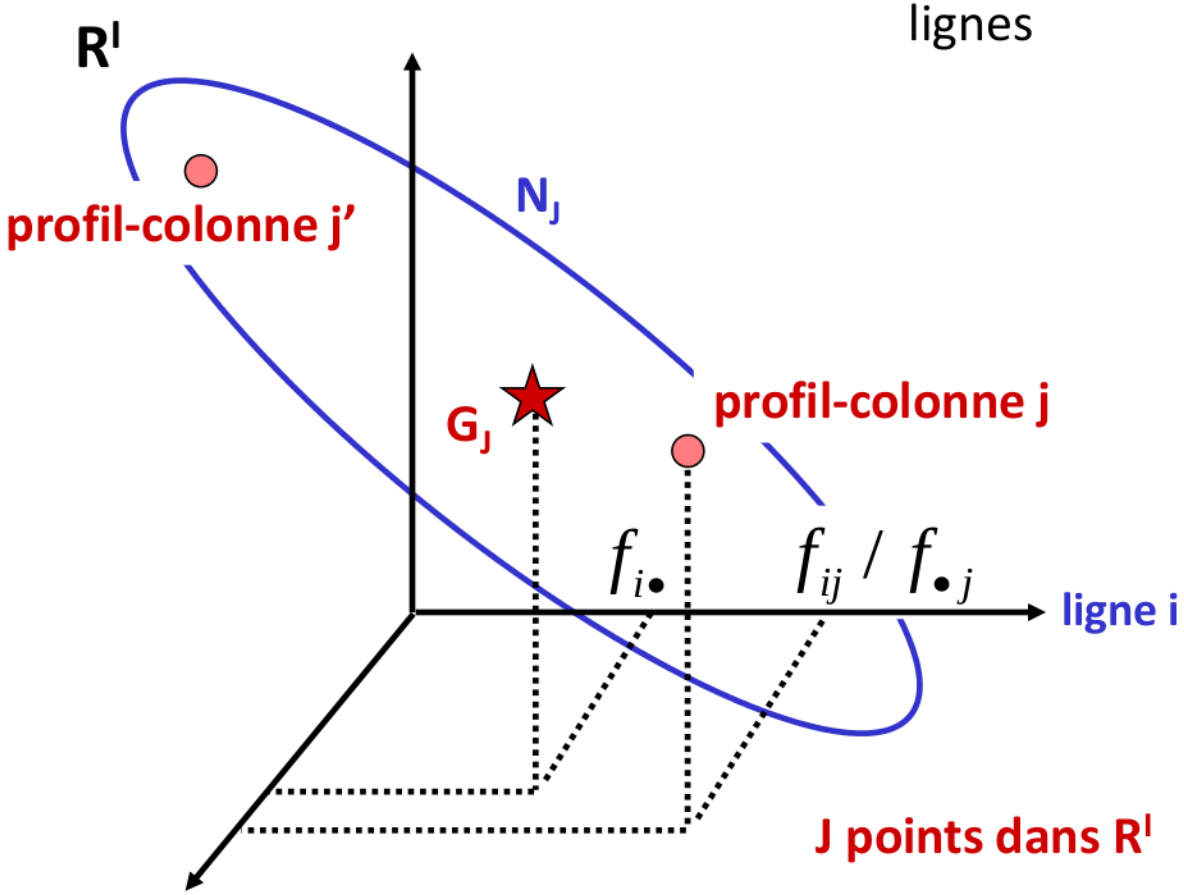
Principe

Géométrie des nuages

N_J

Nuage des profils-colonnes

La construction du nuage des profils-colonnes s'effectue selon une démarche strictement identique à celle des profils-lignes



Ecart entre les profils colonnes et la marge.
Laison entrer 2 variables.

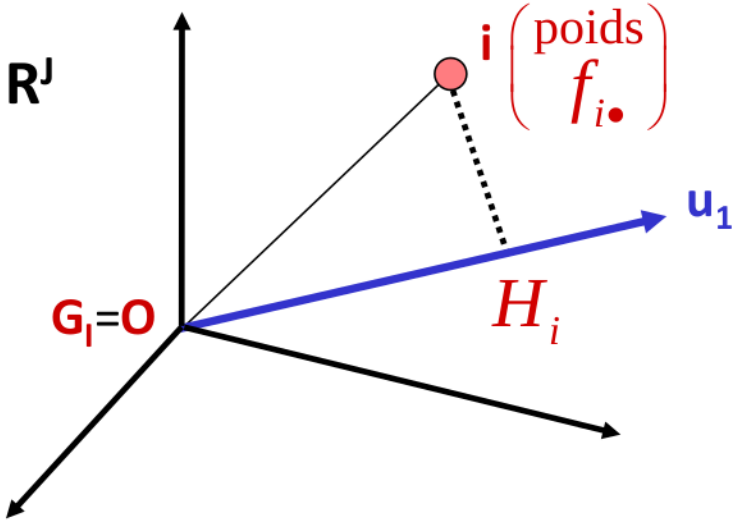
Principe

Ajustement des deux nuages

- L'AFC recherche des **axes factoriels** sur lesquels le nuage est projeté

- L'ensemble des coordonnées des projections d'un nuage sur l'un de ses axes factoriels définit un **facteur**

$\mathbf{u_1}$ rend maximum $\sum_i f_{i\bullet} OH_i^2$



- Chaque axe possède la propriété de rendre maximum l'inertie projetée du nuage avec la contrainte d'orthogonalité aux axes déjà trouvés

- Démarche analogue à l'ajustement du nuage des individus en ACP

- Implique que le nuage soit centré : G devient l'origine des axes

- Les deux nuages N_I et N_J constituent deux représentations d'un même tableau. Les analyses de ces deux nuages ne sont pas indépendantes : les relations entre elles sont regroupées sous le terme de **dualité**.

- N_I et N_J possèdent la même inertie.

- **Propriété remarquable** :

  L'analyse des deux nuages fournit deux suites de facteurs « duaux » (une pour chaque nuage) :
  - deux axes de même rang ont la **même inertie**
  - les facteurs de même rang sur les lignes et les colonnes sont liés par des **relations de transition**

  ⇨ Les facteurs sur I et sur J de même rang doivent être **interprétés conjointement**

Relations de transition

$F_s(i)$: projection de la ligne i sur l'axe de rang s de N_I

$G_s(j)$: projection de la colonne j sur l'axe de rang s de N_J

λ_s : valeur commune de l'inertie associée à chacun de ces axes

Les relations de transitions s'écrivent :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i\bullet}} G_s(j)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{\bullet j}} F_s(i)$$

Principe

Représentation simultanée ligne-colonnes

Au coefficient $\frac{1}{\sqrt{\lambda_s}}$ près, les projections des points d'un nuage sur un axe sont les barycentres des projections des points de l'autre nuage

= propriété barycentrique

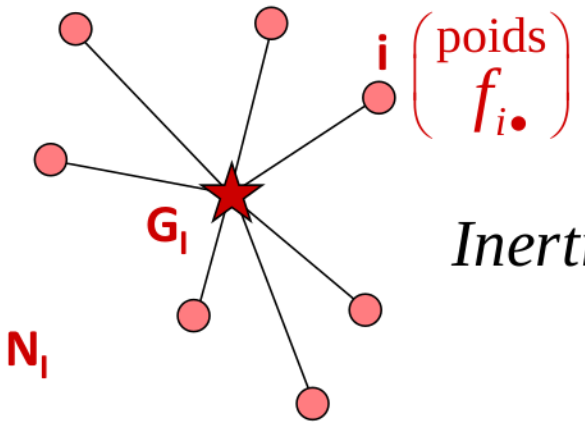
la projection de la modalité i sur un axe est le barycentre des modalités j de l'autre variable pondérées par les fréquences conditionnelles du profil de i

Les éléments « lourds » attirant le barycentre, une colonne j attire d'autant plus une ligne i que la valeur de f_{ij} est élevée

Les points éloignés de l'origine sont les profils les plus différents du profil moyen

Interprétation

Inertie des nuages et statistique du χ^2



$$Inertie(N_I) = \sum Inertie(i) = \sum_i f_{i\bullet} \sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2$$

$$\chi^2 = \sum_{ij} \frac{(effectif\ \ observé - effectif\ \ théorique)^2}{effectif\ \ théorique}$$

$$\chi^2 = \sum_{ij} \frac{(nf_{ij} - nf_{i\bullet} f_{\bullet j})^2}{nf_{i\bullet} f_{\bullet j}}$$

$$\chi^2 = n\left[Inertie(N_I)\right] = n\left[Inertie(N_J)\right]$$

La statistique du χ^2 est égale (au coefficient n près) à l'inertie totale des nuages N_I et N_J

Interprétation

Inertie des axes

Somme des valeurs propres =

$$\sum \lambda_i = \frac{1}{n}\chi^2$$ = inertie totale du nuage

Originalité en AFC : l'inertie traduit la structure du tableau $0 \leq \lambda \leq 1$

→ Chaque facteur représente une part de la liaison entre les variables

→ L'inertie d'un facteur mesure en absolu l'importance de la part de liaison représentée par l'axe

2 situations extrêmes

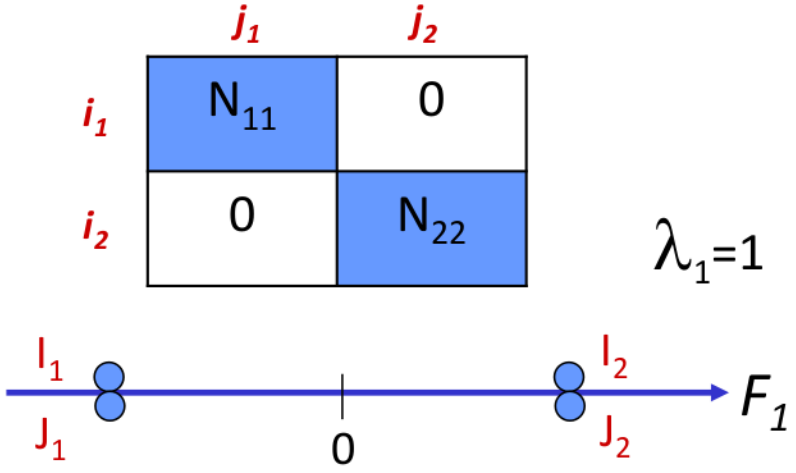
Indépendance

L'inertie de N_I et N_J est nulle

Tous les facteurs ont une inertie nulle

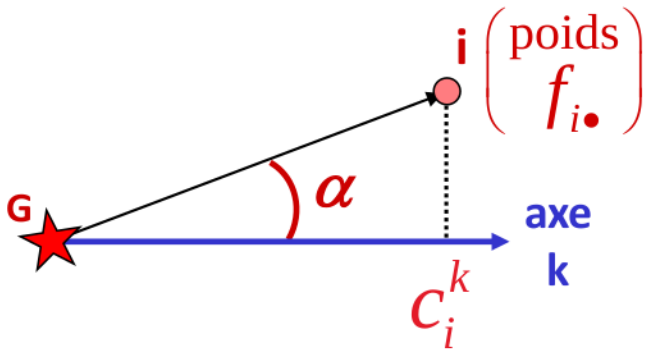
(λ=0)

Dépendance extrême



Interprétation

Cosinus et contributions



Les indices d'aides à l'interprétation définis en ACP sont valables pour un nuage quelconque et s'appliquent donc en AFC

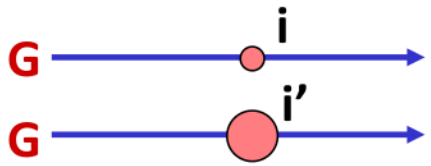
Qualité de représentation (contribution relative) $= \cos^2 \alpha$

Contribution absolue de i à l'axe k $Cr = \dfrac{f_{i\bullet}(c_i^k)^2}{\lambda_k}$

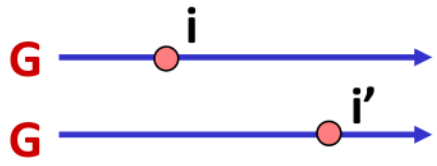
Remarque :

En ACP, les poids de tous les éléments sont égaux

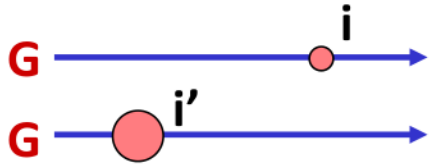
En AFC, ce n'est pas le cas et les poids interviennent dans la contribution d'un point à l'inertie d'un axe



Cr (i) < Cr(i')



Cr (i) < Cr(i')



Cr (i) = Cr(i')

Conclusion

AFC = méthode particulièrement bien adaptée à l'étude d'un tableau de contingence (historiquement imaginée pour ce type de tableau)

Propriétés remarquables de la méthode

L'AFC est une méthode couramment appliquée à d'autres tableaux
(ex : **espèces** x **échantillons** en écologie)
Condition : valeurs positives

On ne raisonne plus en terme de liaison entre deux variables qualitatives
→ **Typologie des lignes et des colonnes à travers leurs profils**

Université Paul Sabatier
TOULOUSE III

MASTER 1 ECOLOGIE
MASTER 1 MABS

BECNM – Analyses de Données Multivariées

Analyse des Correspondances Multiples (ACM)

- **1941** - Guttman
- **1950** – Burt
- **1956** – Hayashi

- *Homogeneity Analysis*
- *Dual scaling*
- *Multiple Correspondence Analysis*

Extension du domaine d'application de l'AFC
Procédures de calcul et règles d'interprétation spécifiques

Domaine d'application

Tableau
individus x variables qualitatives

Exemple : enquêtes socio-économiques

		Variable 1	…	***Variable j***	…	***Variable J***
Individus	***1***	mod_2		mod_1		mod_2
	2	mod_3		mod_3		mod_3
	3	mod_1		mod_3		mod_1
	…	mod_2	…	mod_1	…	mod_1
	…	mod_1		mod_2		mod_3
	I	mod_2		mod_1		mod_2

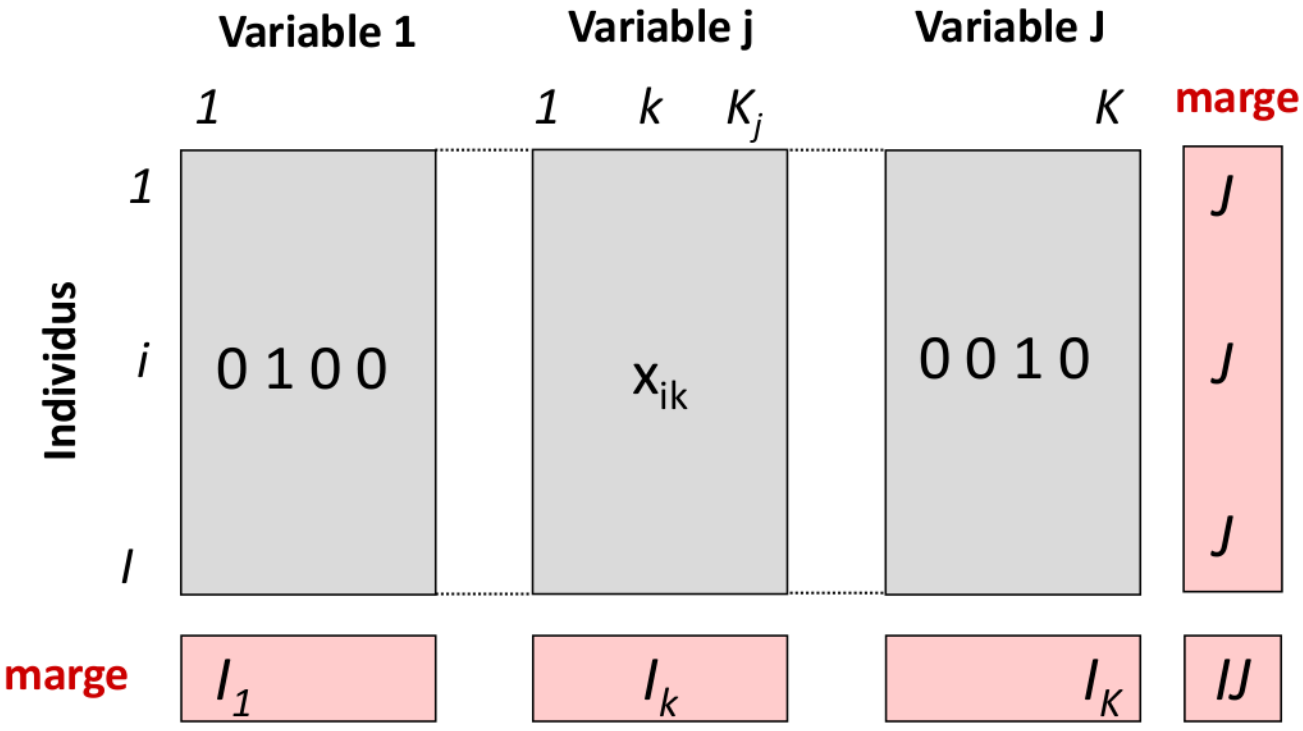
mod_k : modalité k de la variable j

Sous cette forme, le tableau n'est pas exploitable

→ **Recodage des variables**

Notions de base

Tableau disjonctif complet



$$\sum_{k=1}^{k=Kj} x_{ik} = 1$$

$$\sum_{i=1}^{i=I} x_{ik} = I_k$$

$$\sum_{k=1}^{k=K} x_{ik} = J$$

$$\sum_{k=1}^{k=Kj} I_k = I$$

K_j : nombre de modalités de la variable j
K : nombre de modalités toutes variables confondues

Les colonnes de ce tableau sont appelées « **indicatrices des modalités** »

Objectifs

Les objectifs de l'**ACM** font intervenir **trois familles d'objets** :

- **Typologie des individus**
  Basée sur une notion de ressemblance : 2 individus sont proches s'ils possèdent un grand nombre de modalités en commun

- **Liaisons entre variables**
  Cherche à résumer l'ensemble des variables par un petit nombre de variables synthétiques

- **Typologie des modalités**
  Deux modalités se ressemblent si :
  - elles sont présentes ou absentes chez un grand nombre d'individus
  - elles s'associent beaucoup ou peu aux mêmes autres modalités

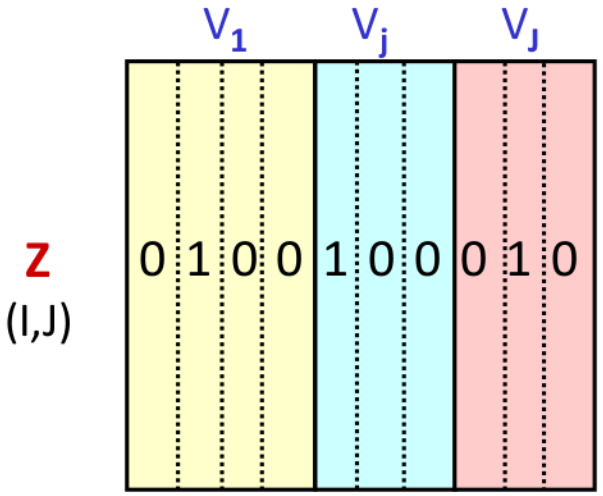
Principe

La problématique de l'ACM est apparentée à celle de l'ACP (tableau **individus** x **variables**) mais peut être considérée comme une généralisation de l'AFC
(liaisons entre plusieurs variables qualitatives)

L'**ACM** est l'**AFC** d'un tableau disjonctif complet

Ses principes sont ceux de l'AFC :

- **Transformation en profils**
- **Pondération** des points par leurs profils marginaux
- **Distance du χ^2**



Problématique riche et complexe qui s'articule autour de la typologie des modalités

Principe

Nuage des individus

- Les poids affectés aux individus sont uniformes
  (la marge sur I est constante = J)

- La transformation en profils-lignes ne modifie guère les données

  Ce profil ne prend que deux valeurs : 0 ou $\frac{1}{J}$

- **Distance entre deux individus**

$$d^2(i,i') = \sum_k \frac{IJ}{I_k}\left(\frac{x_{ik}}{J} - \frac{x_{i'k}}{J}\right)^2 = \frac{1}{J}\sum_k \frac{I}{I_k}\left(x_{ik} - x_{i'k}\right)^2$$

- **Remarque** : $\left(x_{ik} - x_{i'k}\right)$ vaut 0 ou 1

  Une modalité k intervient dans cette distance avec le poids $\frac{I}{I_k}$

  → La présence pour un individu d'une modalité rare éloigne cet individu des autres

Nuage des modalités

- La modalité k est représentée par le profil de la colonne k
- Ce profil ne prend que deux valeurs : 0 ou $\frac{1}{I_k}$
- **Distance entre deux modalités**

$$d^2(k,k') = \sum_i I \left(\frac{x_{ik}}{I_k} - \frac{x_{ik'}}{I_{k'}} \right)^2$$

- Centre de gravité du nuage des modalités
  = profil de la marge sur I (profil plat)

→ Le profil de la colonne k ressemble d'autant plus au profil moyen que l'effectif de la modalité k est grand

→ Une modalité rare sera toujours loin du centre de gravité du nuage des modalités

Relations de transition

$F_s(i)$: projection de l'individu i sur l'axe s de N_I

$G_s(k)$: projection de la modalité k sur l'axe s de N_K

λ_s : valeur commune de l'inertie associée à chacun de ces axes

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K} \frac{x_{ik}}{J} G_s(k)$$

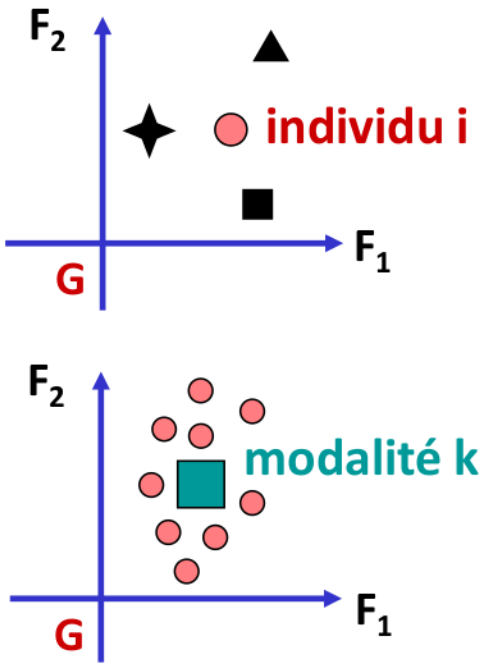
$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{x_{ik}}{I_k} F_s(i)$$

Interprétation

Représentation simultanée

Au coefficient $\frac{1}{\sqrt{\lambda_s}}$ près,

- l'individu i se trouve au point moyen (barycentre) des modalités qu'il possède
- la modalité k se trouve au point moyen des individus qui la possèdent



➡ On peut considérer une modalité comme :
- l'indicatrice d'une variable
- le barycentre d'une classe d'individus

➡ Interprétation de la **proximité entre modalités : 2 points de vue**

- modalités de variables différentes → association de modalités
- modalités d'une variable → ressemblance entre classes d'individus

Interprétation

Les variables à travers leurs modalités

Les variables ne sont pas introduites explicitement dans l'analyse
Elles n'apparaissent qu'à travers l'ensemble de leurs modalités

Barycentre des modalités d'une variable

$$\sum_{k \in K_j} \frac{I_k}{I} \frac{x_{ik}}{I_k} = \frac{1}{I}$$

Le barycentre des modalités d'une même variable se confond avec celui de l'ensemble du nuage

La projection conserve cette propriété :
L'ensemble des modalités d'une même variable est centré sur l'origine

Sous-espace des modalités d'une variable

Caractère complet du tableau disjonctif :

→ L'inertie d'une variable à r modalités est répartie dans un sous-espace à (r-1) dimensions

→ Pour représenter parfaitement les r modalités, (r-1) facteurs sont nécessaires

La quantité maximisée par les axes factoriels est l'inertie projetée du nuage de l'ensemble des modalités

L'ACM fournit des facteurs = **variables numériques synthétiques** liées le plus possible aux variables qualitatives initiales

Influence relative d'une variable en ACM :

Pour un axe donné, l'importance a priori de chaque variable est la même mais le nombre d'axes sur lesquels une variable peut influer est directement lié au nombre de ses modalités

La contribution d'une variable à l'inertie d'un facteur est la somme des contributions de toutes ses modalités

Conclusions

Codage des variables qualitatives

- Les variables qualitatives étudiées en ACM résultent souvent d'une transformation de variables numériques
- Les résultats de l'ACM dépendent du choix du codage

Intérêt du codage en variables qualitatives :

- Rendre homogènes des données numériques et qualitatives
- Pour variables numériques → autre approche des données / ACP
- Etudier des variables dont les distributions sont très irrégulières (un codage en classe neutralise l'influence d'éléments extrêmes)

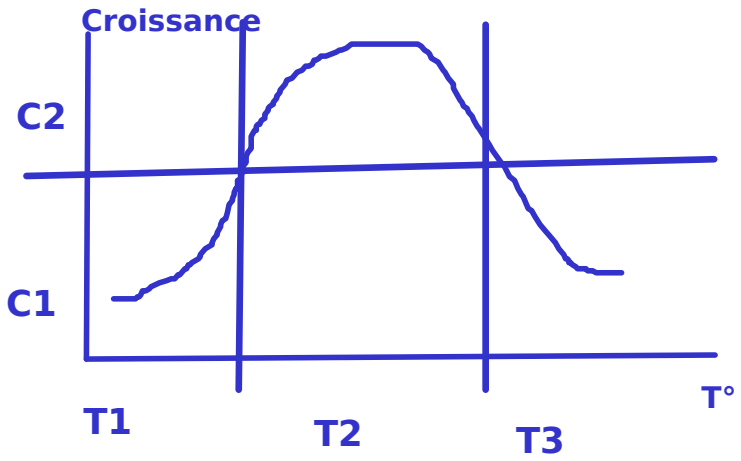
L'ACM peut mettre en évidence des **liaisons non linéaires** entre les variables (graphiquement : proximité de modalités extrêmes), phénomènes invisibles en ACP qui ne tient compte que des liaisons linéaires

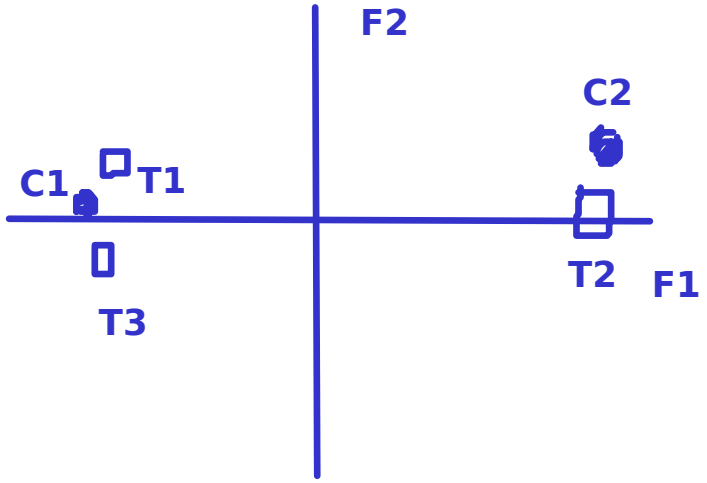
→ En réduisant l'information, on augmente la richesse du résultat !

Conclusions

Codage des variables qualitatives

Ex :





L'ACM peut mettre en évidence des **liaisons non linéaires** entre les variables (graphiquement : proximité de modalités extrêmes), phénomènes invisibles en ACP qui ne tient compte que des liaisons linéaires

→ En réduisant l'information, on augmente la richesse du résultat !

Codage des variables qualitatives

Choix du nombre de classes

- Nb **trop petit** : regroupe des individus différents (perte d'information)
- Nb **trop grand** : risque d'obtenir des classes d'effectif faible
  risque de mettre en évidence des liaisons ponctuelles entre quelques modalités
  → perte de l'aspect synthétique de l'analyse

  En pratique : pas utile de dépasser 8 modalités
  4-5 souvent bien suffisantes

Principe à respecter :

Obtenir des classes de même effectif plutôt que des intervalles de même amplitude (évite d'avoir de faibles effectifs)