ADD

Analyse de données

Chap I Objectif de l'ADD.

Domaine qui regroupe diverses méthodes mathématiques. On souhaite obtenir une organisation de ces données

- Analyse en composantes principales (ACP)
- Analyse factorielle des correspondances (AFC)
- Analyse des composantes multiples (ACM)

} Analyse factorielle

Problématique:

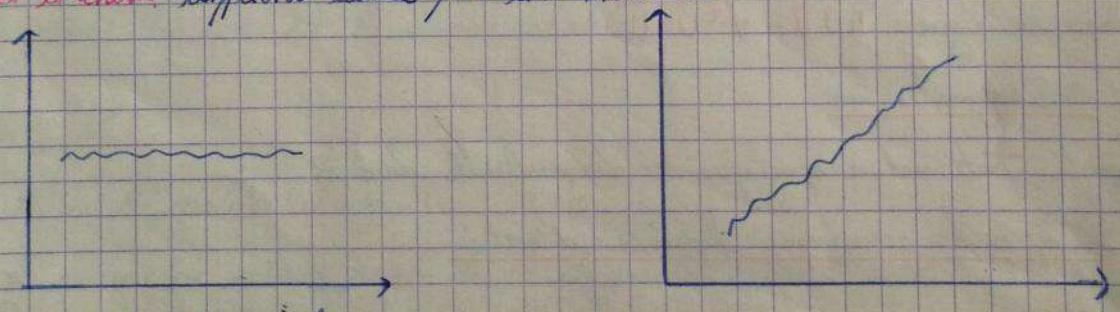
Comment visualiser et interpréter un nuage de points dans un espace de dimension très grande ?

Idée: Définir la "meilleure" projection possible de nuage de points.

Démarche:

Traitement de données initiales afin d'éliminer les biais statiques et les effets d'échelle.

Effet d'échelle: différentes allures pour une même courbe



- Quantifier la notion d'information
  - notion de variance, covariance, corrélation.
  - introduire 1 notion de distance.
- Réaliser des changements d'axes: axes hiérarchisés selon la part d'information contenue
- Choisir les axes de projection en contrôlant la perte d'information
- Réaliser la projection et effectuer l'analyse en la validant par les 7 paramètres que l'on aura obtenus lors de la méthode.

II Type de données

1) Tableau de données

- Les lignes correspondent aux individus statistiques (sauf AFC)
- Les colonnes : * variables statistiques
  * modalités des variables statistiques.

2) Variables.

* En ACP : variables quantitatives de même importance (≈) → discrète
* En AFC : 2 variables sur une population (qualitative ou quantitative)
* En ACM : des variables sur une population (qualitative ou ")

III Calcul matriciel et ADD

1. Matrices symétriques

Propriété : Soit M une matrice quelconque alors les matrices ${}^tM \times M$ et $M \times {}^tM$ sont symétriques

2. Distance

Si u et v, deux vecteurs colonnes dans une base orthonormée

Produit scalaire de U et V : ${}^tU \times V$

Distance euclidienne : $\|U\| = \sqrt{{}^tU \times U}$

3. Diagonalisation

Propriété : Toute matrice symétrique est diagonalisable dans $\mathbb{R}$

4. Approche géométrique ellipsoïde d'inertie

Q° annexe.

Chapitre II — **Paramètres d'un nuage de points. Approche statistique.**

I Paramètres statistiques

Soit n l'effectif total, X variable, x modalité de X.

moyenne : $\overline{X} = E(X) = \frac{1}{n} \sum x$

Variance : $V(x) = E((X - E(x))^2) = \frac{1}{n} \sum (x - \overline{x})^2$

Ecart-type : $\sigma_x = \sqrt{V(x)}$

Covariance : $cov(X, Y) = E(X - E(X))(Y - E(Y)) = \frac{1}{n} \sum (x - \overline{x})(y - \overline{y})$

Corrélation $\rho(X, Y) = \frac{cov(X, Y)}{\sigma_x \cdot \sigma_y}$ avec $-1 \leq \rho(X, Y) \leq 1$

Variable centrée : $X^c = X - \overline{X}$ $\quad \overline{X}^c = 0$

variable centrée et réduite $X^* = \frac{X - \overline{X}}{\sigma_x}$ $\quad \overline{X}^* = 0$, $\sigma_{X^*} = 1$

II Matrices fondamentales

1. Matrice de variance - covariance.

Définition : X^c désignant la matrice des données centrées. On appelle matrice de variance-covariance la matrice $\boxed{\Sigma = \frac{1}{n} {}^t X^c . X^c}$

Remarque : Σ matrice symétrique.

Matrice de variance - covariance : $\Sigma = \begin{pmatrix} var(X_1^c) & cov(X_1^c, X_2^c) & cov(X_1^c, X_3^c) \\ cov(X_2^c, X_1^c) & var(X_2^c) & cov(X_2^c, X_3^c) \\ cov(X_3^c, X_1^c) & cov(X_3^c, X_2^c) & var(X_3^c) \end{pmatrix}$

$M = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{bmatrix}$

$M^c = \begin{bmatrix} a_1 - \overline{A} & b_1 - \overline{B} & c_1 - \overline{C} \\ a_2 - \overline{A} & b_2 - \overline{B} & c_2 - \overline{C} \end{bmatrix}$

${}^t M^c = \begin{bmatrix} a_1 - \overline{A} & a_2 - \overline{A} \\ b_1 - \overline{B} & b_2 - \overline{B} \\ c_1 - \overline{C} & c_2 - \overline{C} \end{bmatrix}$

$\Sigma = \begin{bmatrix} var(A) & cov(A, B) & cov(A, C) \\ cov(B, A) & var(B) & cov(B, C) \\ cov(C, A) & cov(C, B) & var(C) \end{bmatrix}$

trace $\Sigma = var(A) + var(B) + var(C)$

Matrice symétrique, carré, d'ordre égal au nombre de variable

Une covariance peut être négative

Une variance est toujours positive.

Chapitre II

Paramètres d'un nuage de points. Approche statistique.

I Paramètres statistiques

Soit n l'effectif total, X variable, x modalité de X.

moyenne : $\overline{X} = E(X) = \frac{1}{n} \sum x$

Variance : $V(x) = E((x - E(x))^2) = \frac{1}{n} \sum (x - \overline{x})^2$

Ecart-type : $\sigma_x = \sqrt{V(x)}$

Covariance : $cov(x,y) = E(X - E(x))(Y - E(Y)) = \frac{1}{n} \sum (x - \overline{x})(y - \overline{y})$

Corrélation $\rho(x,y) = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$ avec $-1 \leq \rho(x,y) \leq 1$

Variable centrée : $X^c = X - \overline{X}$ $\quad \overline{X}^c = 0$

Variable centrée et réduite $X^\Delta = \frac{X - \overline{X}}{\sigma_x}$ $\quad \overline{X}^\Delta = 0$, $\sigma_{x^\Delta} = 1$

II Matrices fondamentales

1. Matrice de variance - covariance.

Définition : X^c désignant la matrice des données centrées. On appelle matrice de variance-covariance la matrice $\Sigma = \frac{1}{n} {}^t X^c . X^c$

Remarque : Σ matrice symétrique.

Matrice de variance-covariance : $\Sigma = \begin{pmatrix} var(x_1^c) & cov(x_1^c, x_2^c) & cov(x_1^c, x_3^c) \\ cov(x_2^c, x_1^c) & var(x_2^c) & cov(x_2^c, x_3^c) \\ cov(x_3^c, x_1^c) & cov(x_3^c, x_2^c) & var(x_3^c) \end{pmatrix}$

$M = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{bmatrix}$

$M^c = \begin{bmatrix} a_1 - \overline{A} & b_1 - \overline{B} & c_1 - \overline{C} \\ a_2 - \overline{A} & b_2 - \overline{B} & c_2 - \overline{C} \end{bmatrix}$

${}^t M^c = \begin{bmatrix} a_1 - \overline{A} & a_2 - \overline{A} \\ b_1 - \overline{B} & b_2 - \overline{B} \\ c_1 - \overline{C} & c_2 - \overline{C} \end{bmatrix}$

$\Sigma = \begin{bmatrix} var(A) & cov(A,B) & cov(A,C) \\ cov(B,A) & var(B) & cov(B,C) \\ cov(C,A) & cov(C,B) & var(C) \end{bmatrix}$

trace $\Sigma = var(A) + var(B) + var(C)$

Matrice symétrique, carré, d'ordre égal au nombre de variable

Une covariance peut être négative

Une variance est toujours positive.

2. Matrice des corrélations R.

Déf. X^s (matrice centrée réduite). On appelle la matrice de corrélation R :

$$R = \frac{1}{M} {}^t X^s \cdot X^s$$

avec $$R = \begin{bmatrix} 1 & f(x_1^s, x_2^s) & f(x_1^s, x_3^s) \\ f(x_2^s, x_1^s) & 1 & f(x_2^s, x_3^s) \\ f(x_3^s, x_1^s) & f(x_3^s, x_2^s) & 1 \end{bmatrix}$$

AFC à récupérer.

ADD

Chapitre III

Analyse en composantes principales (ACP)

I Aperçu général

p variables quantitatives $X_1, X_2 \dots X_p$. population de n individus

Si $X_1, X_2, \dots X_p$ sont corrélés, il existe des redondances.

Idée de l'ACP

éliminer les redondances : Transformer les p variables (initiales) $X_1 \dots X_p$ en p nouvelles variables → facteurs : $f_1 \dots f_p$.

- conserver k facteurs ($k \leq p$) en conservant un maximum d'information.

redondance = corrélation.

→ Il faut que les facteurs ne soient pas corrélés.

1) La transformation

transformation : $(X_1 \dots X_p) \rightarrow (f_1 \dots f_p)$

compression : extraire des p facteurs, k facteurs → paramètre : qualité globale d'explication = qqà

transformation : Σ ou R → matrice diagonale → DIAGONALISATION de la matrice Σ ou R.

Soit U la base de vecteurs propres → facteurs

Changement de base sur les données initiales : $\begin{matrix} A^c \\ A^s \end{matrix} \longrightarrow F = \begin{matrix} A^c U \\ A^s U \end{matrix}$ Matrice des composantes principales.

↳ = coordonnées des individus dans la nouvelle base.

On a donc :

* A^c ou A^s matrice d'origine.
* U matrice de vecteurs propres de la diagonalisation de Σ ou de R.
* Λ matrice diagonale des valeurs propres.
* F matrice des composantes principales.

2) L'interprétation

* "identifier" les facteurs = donner une signification concrète aux facteurs

idée : calculer les corrélations entre les variables et les facteurs.

→ matrice de saturation $S = \frac{1}{n} {}^t X^s F \Lambda^{-1/2}$

II ACP d'un exemple.

Cf Annexe avec X : prix d'un litre de lait
Y : " " d'eau
Z : " " d'huile.

A° → Matrice Centrée réduite → moyennes = 0.
Ecart-type = 0

matrice corrélation → termes de la diagonale = 1.

trace = 1+1+1 = 3 = p. = variance totale (permet de mesurer la qté totale d'info)

3ème étape : diagonalisation.

déterminer ses valeurs propres puis vecteurs propres.

$\det(R-\lambda I) = 0$ (det valeurs propres). On obtient p valeurs propres.

On classe de la + grande à la + petite : [? ? ?]
grande moyen petit

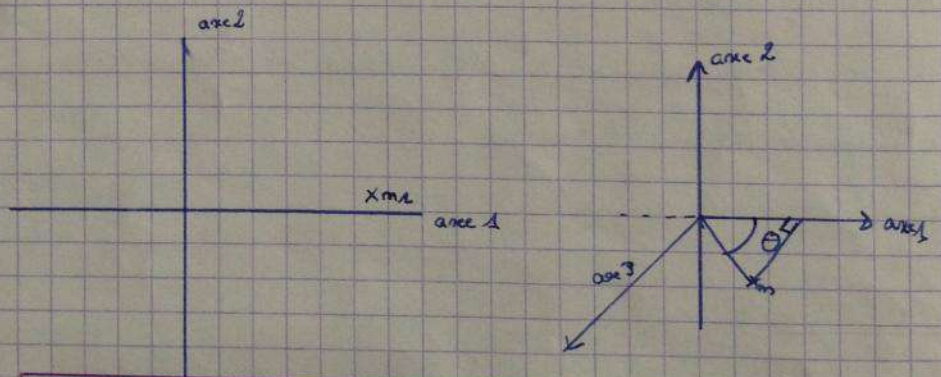
* Vecteurs propres : Si λ est une valeur propre d'une matrice M, un vecteur propre associé à λ est un vecteur u tel que $MU = \lambda U$

Règle : On choisi des vecteurs propres unitaires ou normés, càd de norme = 1.

Propriété : Les vecteurs propres sont orthogonaux 2 à 2.

4ème étape : calcul des composantes principales.

Qualité de représentation d'un individu sur un axe. (qlt ou $\cos^2$)



$$\cos^2(mi ; \text{axe } k) = \frac{\text{coord}(mi, \text{axe } k)^2}{\sum \text{coord}(mi, \text{axe } j)^2}$$

Si qlt ≈ 1 alors mi proche de l'axe k.
Si qlt ≈ 0 " " éloigné " k.

Matrice de saturation: - toujours carré

- indique les corrélations entre les variables et les facteurs
- IDC (indice de Corrélation) toujours compris entre -1 et 1.

Corrélations: facteur 1 facteur 2 facteur 3

X^s

Y^s

Z^s

Carte de corrélation.

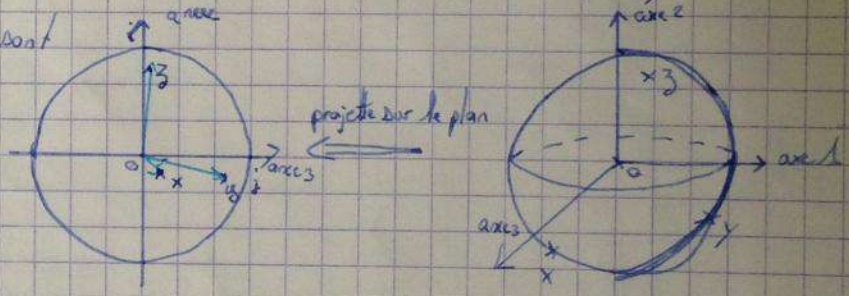
$S = [a_{ij}]_{1 \le i \le p \atop 1 \le j \le p}$ $\quad \sum_{j=1}^{p} a_{ij}^2 = 1$ (ligne)

$\sum_{i=1}^{p} a_{ij}^2 = \lambda_j$ (colonne)

$p=3$ $\boxed{a_{i1}^2 + a_{i2}^2 + a_{i3}^2 = 1}$.

$\Rightarrow x^2 + y^2 + z^2 = 1$

- Les points variables sont situés sur une hypersphère de centre O et de rayon 1.
- Les projetés des points variables sont situés à l'intérieur du cercle de centre O de rayon 1.



Interprétation:

En ACP essaye d'identifier les axes. Essaye de donner un nom aux différents axes

Avec quelles variables les axes sont corrélés.

A proximité des axes et des points

2 variables corrélés: si l'une augmente, l'autre augmente

Matrice de corrélations - élément de contrôle

ACP: - variables quantitatives

- variables à peu près la même importance.
- nature des données à observer

ACP centrée X^c; Σ

ACP centrée réduite X^Δ, R

3 cas de figure pour une ACP : - variable longueur (m: entre 3 et 5, cm: 300 et 500) unités différentes
centrée réduite ⇒ variance + grande dans le 2e cas que dans le 1er)

Matrice centrée réduite: Σ colonne = 0

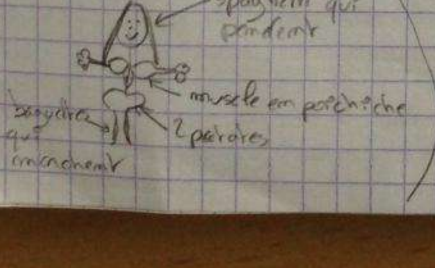
Σ valeur propre = nb de variables = trace de R.

qge : qualité globale d'explication. (mesure la qualité de représentation d'un individu par rapport à l'axe)

Matrice de corrélat : corrélation entre les variables -1 < 1 (forte corrélation)

Σ carré (des coefs)^2 = 1 (des lignes)
colonne = (valeur propre)^2

Matrice des saturations:



→ Sextime alias Sex'

Analyse factorielle des correspondances (AFC)

Objectif : Etudier les liens eventuels entre 2 variables sur une population.

On est face à 2 variables generalement qualitatives

I Tableau de données

1) Le tableau des effectifs (k)

2 variables X et Y, possedant un certain nombre de modalités

n = nombre de modalités de X → nombre de lignes

p = " " de Y → colonnes.

N = effectif total de la population

k_{ij} = effectif de la modalité $X = x_i$ et $Y = y_j$.

$k_{i\cdot}$ =

$k_{\cdot j}$

X \ Y	y_1	y_2	...	y_p	
x_1	k_{11}	k_{12}		k_{1p}	$k_{1\cdot}$
x_2	k_{21}	k_{22}			$k_{2\cdot}$
⋮					
x_n	k_{n1}	k_{n2}		k_{np}	$k_{n\cdot}$
	$k_{\cdot 1}$	$k_{\cdot 2}$		$k_{\cdot p}$	

$X = x_i$

$Y = y_j$.

$$\sum_{i=1}^{n}\sum_{j=1}^{p} k_{ij} = N$$

$$\sum_{i=1}^{n} k_{i\cdot} = N \qquad \sum_{j=1}^{p} k_{\cdot j} = N$$

Données initiales

		C (chanson)	J (jazz)	$k_{i\cdot}$
Jeune →	J	40	20	60
	AF	40	10	50
	AM	30	20	50
Vieux →	V	10	30	40
	$k_{\cdot j}$	120	80	200

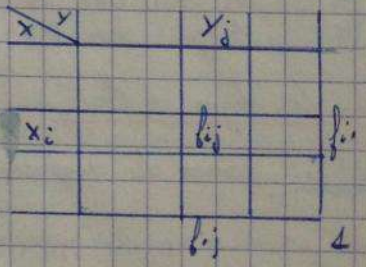
2) Le tableau des fréquences

tableau des fréquences par rapport à l'effectif total

f_{ij} = fréquence de $X = x_i$ et $Y = y_j$

$f_{ij} = p(X = x_i \cap Y = y_j)$

$$f_{ij} = \frac{k_{ij}}{N}$$



$f_{i\cdot}$ = fréquence de $X = x_i$

$$f_{i\cdot} = \frac{k_{i\cdot}}{N} = \sum_{j=1}^{p} f_{ij}$$

$f_{\cdot j}$ = fréquence de $Y = y_j$

$$f_{\cdot j} = \frac{k_{\cdot j}}{N} = \sum_{i=1}^{n} f_{ij}$$

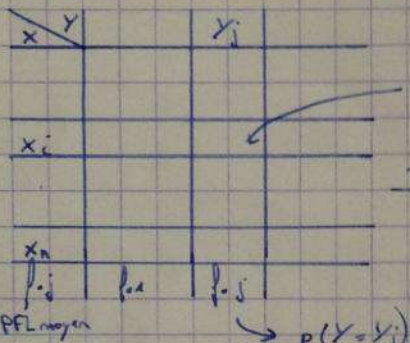
$f_{ij} = p(Y = y_j)$

(Cf tableau des fréquences)

	C	J	
J	0,18	0,12	$f_{1\cdot}=0,3$
AF	0,15	0,1	$f_{2\cdot}=0,25$
AM	0,15	0,1	$f_{3\cdot}=0,25$
V	0,12	0,08	0,2
	$f_{\cdot 1}=0,6$	$f_{\cdot 2}=0,4$	1

3. Les tableaux des profils-lignes et des profils colonnes.

a) Le tableau des PFL



$$\frac{k_{ij}}{k_{i\cdot}} = \frac{\frac{k_{ij}}{N}}{\frac{k_{i\cdot}}{N}} = \frac{f_{ij}}{f_{i\cdot}} = \frac{p(X=x_i \cap Y=y_j)}{p(X=x_i)} = p(Y=y_j / X=x_i)$$

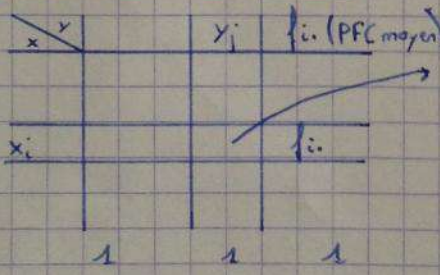
Indépendance $\iff p(X=x_i \cap Y=y_j) = p(X=x_i) \times p(Y=y_j)$

$\iff p(Y=y_j / X=x_i) = p(Y=y_j)$

Y \ X	C	J	
J	0,6/0,6	0,4/0,4	1
AF	0,6	0,4	1
AM	0,6	0,4	1
V	0,25	0,75	1
PFM	0,6	0,4	1

poids des colonnes : $f_{\cdot j}$

b) Le tableau des PFC



$$\frac{k_{ij}}{k_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}} = \frac{p(X=x_i \cap Y=y_j)}{p(Y=y_j)} = p(X=x_i / Y=y_j)$$

X \ Y	C	J	
J			0,3
AF			0,25
AM			0,25
V			0,2
	1	1	1

PFL : 1 point dans $\mathbb{R}^p$

PFC : 1 point dans $\mathbb{R}^n$

4. Les degrés de liberté

PFL : Somme des composantes d'un PFL = 1 $\Rightarrow$ chaque PFL vérifie

$x_1 + x_2 + \dots + x_p = 1$ hyperplan de dim = p-1.

$\Rightarrow$ p-1 degrés de liberté pour les PFL.

II AFC

1. La distance du χ^2 (KHI-Deux)

cf Annexe

2. L'ajustement du nuage des n PFL dans l'espace des p colonnes : matrice à diagonaliser

$\|u\|^2 = {}^t u \times u$.

$$d^2_{\chi^2}(i,e) = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{ej}}{f_{e\cdot}} \right)^2 = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \frac{f_{ej}}{f_{e\cdot}\sqrt{f_{\cdot j}}} \right)^2$$

$$A = \left[\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \right]_{\substack{1 \le i \le n \\ 1 \le j \le p}} \quad \text{sur PFC} \quad B = \frac{f_{ij}}{f_{\cdot i}\sqrt{f_{j\cdot}}}$$

En pratique $X^* = \left[\frac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \right]_{\substack{1 \le i \le n \\ 1 \le j \le p}}$

La matrice à diagonaliser : $\boxed{T^* = {}^t X^* X^*}$

$X^* \rightarrow$ dimension "np" $\rightarrow$ n lignes, p colonnes.

${}^t X^* \rightarrow$ "pn"

T^* = matrice carrée, symétrique, de dimension "p"

3. Propriété des valeurs propres.

Prop: la matrice à diagonaliser possède toujours la valeur propre 1, appelée valeur propre triviale.

Ex: $X^* = \left[\frac{f_{ij}}{\sqrt{f_{i\cdot} f_{\cdot j}}} \right]_{\substack{1 \le i \le n \\ 1 \le j \le p}}$ ${}^t X^* = \left[\frac{f_{ji}}{\sqrt{f_{i\cdot} f_{\cdot i}}} \right]$

$[a_{ik}]_{\substack{1 \le i \le p \\ 1 \le k \le n}} \times [b_{kj}]_{\substack{1 \le k \le n \\ 1 \le j \le p}} = [\sum a_{ik} b_{kj}]_{\substack{1 \le i \le p \\ 1 \le j \le p}}$.

$$T^* = {}^t X^* X^* = \left[\frac{f_{ki}}{\sqrt{f_{k\cdot} f_{\cdot j}}} \right]_{\substack{1 \le k \le n \\ 1 \le j \le p}} \left[\frac{f_{ki}}{\sqrt{f_{k\cdot} f_{\cdot i}}} \right]_{\substack{1 \le i \le p \\ 1 \le k \le n}} =$$

$$= \left[\sum_{k=1}^{n} \frac{f_{ki} \times f_{kj}}{f_{k\cdot}\sqrt{f_{\cdot i} f_{\cdot j}}} \right]_{\substack{1 \le i \le p \\ 1 \le j \le p}}$$

$$= \left[\frac{1}{\sqrt{f_{\cdot i} f_{\cdot j}}} \sum_{k=1}^{n} \frac{f_{ki} \times f_{kj}}{f_{k\cdot}} \right]_{\substack{1 \le i \le p \\ 1 \le j \le p}}.$$

$$T^* G = [t_{ik}]_{\substack{1 \le i \le p \\ 1 \le k \le p}} \times [\sqrt{f_{\cdot k}}]_{1 \le k \le p} = \left[\sum_{k=1}^{p} t_{ik} \times \sqrt{f_{\cdot k}} \right]_{1 \le i \le p}$$

$\overset{?}{=} G = [\sqrt{f_{\cdot i}}]_{1 \le i \le p}$.

$$= \sum_{k=1}^{p} t_{ik} \sqrt{f_{\cdot k}}$$

ACP centrée réduite X^A, R

2 cas de figure pour une ACP : - variable longueur (m : entre 3 et 5, cm : 300 et 500) unités différentes
- Lié réduite ⇒ variance + grande dans le 2e cas que dans le 1er

$$\sum_{j=1}^{p} T_{ij} \times \sqrt{f_{\bullet j}} = \sum_{j=1}^{p} \frac{1}{\sqrt{f_{\bullet i} \times f_{\bullet j}}} \sum_{k=1}^{n} \frac{f_{ki} \times f_{kj}}{f_{k\bullet}} \times \sqrt{f_{\bullet j}} = \frac{1}{\sqrt{f_{\bullet i}}} \sum_{k=1}^{n} \frac{f_{kj}}{f_{k\bullet}} \sum_{j=1}^{p} f_{ki}$$

$$= \frac{1}{\sqrt{f_{\bullet i}}} \sum_{k=1}^{n} \frac{f_{ki}}{f_{k\bullet}} \times f_{k\bullet} = \frac{1}{\sqrt{f_{\bullet j}}} \sum_{k=1}^{n} f_{ki} = \frac{f_{\bullet i}}{\sqrt{f_{\bullet i}}} = \sqrt{f_{\bullet i}} \text{ donc } T^t G = G.$$

Inertie totale : L'inertie totale notée R est égale à la somme des valeurs propres non triviales $\boxed{R = \sum \lambda_i - 1}$

4. Coordonnées des points-ligne.

Coordonnée du point ligne i sur l'axe α (engendré par le vecteur propre noté u_α)

$$\underbrace{\psi_{\alpha i}}_{\text{projection du point i sur l'axe engendré par le vec. p. } \alpha = \text{coordonnée de i sur } \alpha} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\bullet}\sqrt{f_{\bullet j}}} \times u_{\alpha j} \qquad \boxed{\Psi = AU}$$

5. Etude des PFC / Ajustement du nuage des p colonnes dans l'espace des n lignes

Matrice à diagonaliser : $W^* = X^{*t} X^*$ ⇒ matrice symétrique

Matrice des données : $B = \left[\frac{f_{ji}}{f_{\bullet i}\sqrt{f_{j\bullet}}} \right]$

6. Coordonnés des points-colonnes

$$\varphi_{\alpha j} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{\bullet j}\sqrt{f_{i\bullet}}} \times v_{\alpha i} \qquad \boxed{\Phi = BV}$$

Analyse des Correspondances Multiples

Methode permettant d'étudier les liaisons entre plusieurs variables qualitatives ou quantitatives, c'est donc une generalisation de l'ACP.

Q_i = Variable i.

Q_{ij} = valeur de Q_i pour l'individu j.

n = nombre d'individus

q = nombre de variables.

m_k = nombre de modalité de Q_k.

M = nombre total de modalités.

1. Tableau disjonctif complet.

On associe à chaque modalité une variable appelée variable indicatrice, ne prenant les valeurs 0 et 1 selon que la modalité est verifié ou non.

X_j = variable indicatrice de la modalité j

$X_j(k)$: valeur (0 ou 1) de la variable indicatrice X_j pour l'individu k.

→ Σ termes de chaque ligne = q.

Σ " la colonne associée à X_j ← n_j

→ La frequence associée, appelé frequence marginale ← p_j

$$p_j = \frac{n_j}{n_q}$$

ACM = si $\frac{\text{nb d'individus très grand}}{\text{nb de modalité}}$

Les individus doivent être répartis assez regulierement.

2. Descript° de la methode.

Basé sur l'étude des distances entre PFL et PFC. et sur l'analyse des PFL et PFC.

⚠ On ne s'intéresse pas à l'indépendance.

• Notion de distance (du χ^2)

$$d^2(i;l) = \sum_{j=1}^{m} \frac{\left(\frac{x_j(i)}{q} - \frac{x_j(l)}{q}\right)^2}{p_j} = \frac{1}{q^2} \sum_{j=1}^{m} \frac{(x_j(i) - x_j(l))^2}{p_j}$$

Ordre de grandeur : Fréquence marginale colonne très faible (très peu stable) → p_j très petit, donc quand on divise donne un nombre très grande → poid très fort.

→ D'où regroupement de modalités.

Modalités

$$d^2(j;l) = \sum_{i=1}^{n} \frac{\left(\frac{x_j(i)}{n_j} - \frac{x_l(i)}{n_l}\right)^2}{\frac{1}{n}} = n \sum_{i=1}^{n} \left(\frac{x_j(i)}{n_j} - \frac{x_l(i)}{n_l}\right)^2$$

Prime à la rareté à éviter (en regroupant les effectifs)

+ modalité fq faible, + individus éloignés.

+ eff. modalité faible, + distance entre mod. est grande.

• Analyse mathematiques

Même methode qu'en AFC.

L'inertie totale est égale à $\boxed{\frac{m}{q} - 1}$