Méthodes factorielles

Mohamed Nadif

Université Paris Descartes

Outline

Méthodes factorielles

- Objectifs
- Place dans le contexte data science

Méthodes factorielles

- Analyse en composantes principales (ACP)
- Analyse des correspondances (AC ou AFCM)
- Analyse des correspondances multiples (ACM ou AFCM)
- Analyse factorielle des données mixtes
- Analyse factorielle multiple

Autres Méthodes

- MDS
- Isomap
- LLE
- ICA

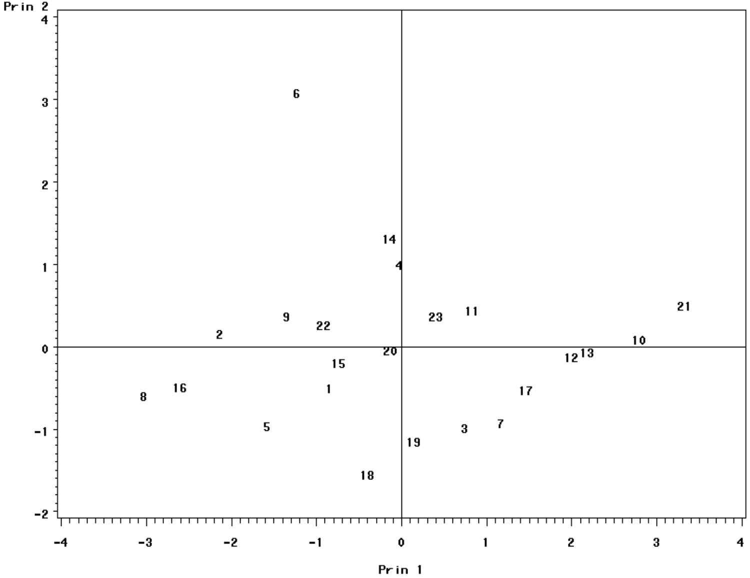
Mesures sur 23 papillons

num	Z1	Z2	Z3	Z4
1	22	35	24	19
2	24	31	21	22
3	27	36	25	15
4	27	36	24	23
5	21	33	23	18
6	26	35	23	32
7	27	37	26	15
8	22	30	19	20
9	25	33	22	22
10	30	41	28	17
11	24	39	27	21
12	29	39	27	17
13	29	40	27	17
14	28	36	23	24
15	22	36	24	20
16	23	30	20	20
17	28	38	26	16
18	25	34	23	14
19	26	35	24	15
20	23	37	25	20
21	31	42	29	18
22	26	34	22	21
23	24	38	26	21

Mesures sur 23 papillons

- Problème de classification ou de visualisation
- Dimension 4
- Analyse exploratoire pour proposer une solution

ACP sur les papillons



Distance d: $A \times A \to \mathbb{R}^+$

- $\forall \boldsymbol{x}, \boldsymbol{y} \in A,\ d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
- $\forall \boldsymbol{x}, \boldsymbol{y} \in A,\ d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$
- $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in A, d(\boldsymbol{x}, z) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$

Ultramétrique

- $\forall \boldsymbol{x}, \boldsymbol{y} \in A,\ d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
- $\forall \boldsymbol{x}, \boldsymbol{y} \in A,\ d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$
- $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in A,\ d(x, z) \leq max(d(x, y), d(y, z))$

Produit scalaire: $E \times E \to \mathbb{R}$ (E: espace vectoriel)

- $\forall \boldsymbol{x} \in E, < \boldsymbol{x}, \boldsymbol{x} >= 0 \Rightarrow x = 0$
- $\forall \boldsymbol{x}, \boldsymbol{y} \in E, < \boldsymbol{x}, \boldsymbol{y} >=< \boldsymbol{y}, \boldsymbol{x} >$
- $\forall \boldsymbol{x} \in E, < \boldsymbol{x}, \boldsymbol{x} >\geq 0$

Extension : $< x, \boldsymbol{y} >_M = x^T My$, M matrice $(p \times p)$

- symétrique $M^T = M$
- définie $\forall \boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x}^T M\boldsymbol{x} = 0 \Rightarrow x = 0$
- positive $\forall \boldsymbol{x}, \boldsymbol{y} \in E$, $\boldsymbol{x}^T M\boldsymbol{x} \geq 0$

Norme: E (espace vectoriel) $||.|| : E \to \mathbb{R}^+$

- $\forall \boldsymbol{x} \in E, \lambda \in \mathbb{R}, ||\lambda x|| = |\lambda||x||$
- $\forall \boldsymbol{x} \in E, ||\boldsymbol{x}|| = 0 \Rightarrow x = 0$
- $\forall \boldsymbol{x}, \boldsymbol{y} \in E, ||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$

Norme euclidienne et distance

- E espace euclidien, on définit la norme euclidienne $||\boldsymbol{x}||_M^2 =< x, x >_M$
- $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$ est une distance dans E
- $d_M^2(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_M^2 =< \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} >_M = (\mathbf{x} - \mathbf{y})^T M(\mathbf{x} - \mathbf{y})$
- Par exemple, $M = I$ $d_M^2(\mathbf{x}, \mathbf{y}) = \sum_j (x_j - y_j)^2$, $M = (1/s_j^2)$, $d_M^2(\mathbf{x}, \mathbf{y}) = \sum_j (\frac{x_j}{s_j} - \frac{y_j}{s_j})^2$

Autres distances

- Distance de Manahattan: $d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^p |x_j - y_j|$
- Distance de Minkowski : $d(\mathbf{x}, \mathbf{y}) = (\sum_{j=1}^p |x_j - y_j|^p)^{1/p}$
- Distance de Mahalanobis (prend en compte les corrélations entre variables) (Σ matrice de covariance)

$$d_{\Sigma^{-1}}^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})$$

- Données

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- Expression matricielle de $d^2_M(\mathbf{x}_i, \mathbf{x}_{i'}) = ||\mathbf{x}_i - \mathbf{x}_{i'}||^2_M$ avec $M = Diag(m^1, \dots, m^p)$

$$d^2_M(\mathbf{x}_i, \mathbf{x}_{i'}) = ((x_{i1} - x_{i'1}), \dots, (x_{ip} - x_{i'p})) \begin{pmatrix} m^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & m^p \end{pmatrix} \begin{pmatrix} (x_{i1} - x_{i'1}) \\ \vdots \\ \vdots \\ \vdots \\ (x_{ip} - x_{i'p}) \end{pmatrix}$$

- Expression de la distance : $d^2_M(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_j m^j (x_{ij} - x_{i'j})^2$,
- Expressions vectorielles : $\mathbf{X}^T\mathbf{X} = \sum_i^n \mathbf{x}_i \mathbf{x}_i^T$, $\mathbf{X}\mathbf{X}^T = \sum_j^p (\mathbf{x}^j)(\mathbf{x}^j)^T$

Nuages des individus

- On note $I = \{1, \dots, n\}$ et $J = \{1, \dots, p\}$
- L'ensemble des individus peut être représenté par le nuage $N(\Omega)$

$$N(\Omega) = \{(\boldsymbol{x}_i; \pi_i); i \in I\}$$

  - Ce nuage inclus dans $\mathbb{R}^p$ muni de la métrique $(D_J)_{p \times p}$
  - $(D_J)_{p \times p}$: pondérations pour toutes les variables
- Distance entre deux individus
  - $D_J = Id$ alors $d^2_{D_J}(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_j 1 \times (x_{ij} - x_{i'j})^2$,
  - $D_J = Diag(1/s_1^2, \dots, 1/s_j^2)$ alors $d^2_{D_J}(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_j \frac{1}{s_j^2}(x_{ij} - x_{i'j})^2 = \sum_j (\frac{x_{ij}}{s_j} - \frac{x_{i'j}}{s_j})^2$
- L'ensemble des variables peut être représenté par le nuage $N(V)$

$$N(V) = \{(\boldsymbol{x}^j; \rho^j); j \in J\}$$

  - Ce nuage inclus dans $\mathbb{R}^n$ muni de la métrique $(D_I)_{n \times n}$
  - $(D_I)_{n \times n}$ représente les pondérations pour tous les individus

Nuages des variables

- Tableau centré

$$Cov(\mathbf{x}^j, \mathbf{x}^{j'}) = \sum_{i=1}^{n} \pi_i x_{ij} x_{ij'} = (\mathbf{x}^j)^T D_I (\mathbf{x}^{j}{}') = < \mathbf{x}^j, \mathbf{x}^{j'} >_{D_I}$$

$$Var(x^j) = \sum_{i=1}^{n} \pi_i (x_{ij})^2 = ||\mathbf{x}^j||^2_{D_I} = d^2_{D_I}(\mathbf{O}, \mathbf{x}^j)$$

$$Cor(\mathbf{x}^j, \mathbf{x}^{j}{}') = \frac{< \mathbf{x}^j, \mathbf{x}^{j'} >_{D_I}}{||\mathbf{x}^j||_{D_I} ||\mathbf{x}^{j'}||_{D_I}}$$

$$d^2_{D_I}(\mathbf{x}^j, \mathbf{x}^{j'}) = Var(\mathbf{x}^j) + Var(\mathbf{x}^{j'}) - 2Cov(\mathbf{x}^j, \mathbf{x}^{j'})$$

- Tableau centré-réduit

$$Cor(\mathbf{x}^j, \mathbf{x}^{j'}) = Cov(\mathbf{x}^j, \mathbf{x}^{j'}) = < \mathbf{x}^j, \mathbf{x}^{j'} >_{D_I}$$

$$Var(x^j) = d^2_{D_I}(\mathbf{O}, \mathbf{x}^j) = ||\mathbf{x}^j||^2_{D_I} = 1 \quad \text{et} \quad Cor(\mathbf{x}^j, \mathbf{x}^{j}{}') = < \mathbf{x}^j, \mathbf{x}^{j'} >_{D_I}$$

$$d^2_{D_I}(\mathbf{x}^j, \mathbf{x}^{j'}) = 2(1 - Cor(\mathbf{x}^j, \mathbf{x}^{j'}))$$

Outline

Objectif : Nuage fidèle

- Recherche d'un un espace H (droite, plan le plus souvent) permettant de rendre compte de la forme du nuage en minimisant les déformations de la projection
- Visualisation + Inteprétation
- Formulations (sans pondérations et avec pondérations):

$$Max_H\{\sum_{i,i'} d^2(\mathbf{h}_i, \mathbf{h}_{i'})\}$$

  $\mathbf{h}_i$ est la projection de $\mathbf{x}_i$ sur H

$$Max_H\{\sum_{i,i'} \pi_i \pi_{i'} d^2(\mathbf{h}_i, \mathbf{h}_{i'})\}$$

- Problème équivalent à

$$Max_H\{\sum_{i} \pi_i d^2(\mathbf{h}_i, \mathbf{g})\}$$

  $\mathbf{g}$ centre de gravité de H
- Problème ancien et purement numérique, traité par (Sylvester, 1889)
- Plus tard par Eckart and Young , 1936, 1939: SVD

Inertie ou Variance

- L'inertie de $N(\Omega)$ par rapport à un point **a**

$$I_{\mathbf{a}} = \sum_{i=1}^{n} \pi_i d^2_{D_J}(\mathbf{x}_i, \mathbf{a})$$

- L'inertie par rapport au centre de gravité $\mathbf{g} = \bar{\mathbf{x}} = \sum_{i=1}^{n} \pi_i \mathbf{x}_i$

$$I_{\mathbf{g}} = \sum_{i} \pi_i d^2_{D_J}(\mathbf{x}_i, \mathbf{g})$$

- I est l'inertie totale et souvent notée I, elle est dite aussi Variance
  - lorsque $\pi_i = \frac{1}{n}$ et $D_J = Id$ alors $I = \sum_{i=1}^{n} \frac{1}{n} \sum_{j=1}^{p} (x_{ij} - \bar{x}^j)^2$
  - $I = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}^j)^2 = \sum_{j=1}^{p} Var(x^j)$
- L'inertie du nuage $N(\Omega)$ par rapport à l'espace E_k s'écrit :

$$I_{E_k} = \sum_{i=1}^{n} \pi_i d^2_{D_J}(\boldsymbol{x}_i, E_k)$$

- Objectif de l'ACP : Trouver le s.e E_k de dimension $(k < p)$ tel que I_{E_k} soit minimum

Recherche de E_k

- On peut montrer que E_k contient nécessairement le centre de gravité **g** (Théorème Huygens)
- **X** centrée implique E_k est s.e.v contenant l'origine O
- A partir de la relation $I = I_{E_k} + I_{E_k^\perp}$, le problème consiste à chercher E_k maximisant $I_{E_k^\perp}$
- Solution de problème (2 théorèmes en Algèbre): E_k est formé de $\Delta u_1 \oplus \Delta u_2 \ldots \oplus \Delta u_k$ ($\Delta_{u_\alpha} \perp \Delta u_\beta$ for $\alpha \neq \beta$)
- projection

Expressions matricielles

- Inertie totale

$$
\begin{aligned}
I &= \sum_i \pi_i d^2_{D_J}(\mathbf{x}_i, \mathbf{O}) \\
&= \sum_i \pi_i ||\mathbf{x}_i||^2_{D_J} \\
&= \sum_i \pi_i \mathbf{x}_i^T D_J \mathbf{x}_i \\
&= trace \sum_i \pi_i \mathbf{x}_i^T D_J \mathbf{x}_i \\
&= trace \sum_i \pi_i D_J \mathbf{x}_i \mathbf{x}_i^T \\
&= trace(\mathbf{X}^T D_I \mathbf{X} D_J)
\end{aligned}
$$

- $S = \mathbf{X}^T D_I \mathbf{X} D_J$
- Si $\mathbf{D}_I = \frac{1}{n}\mathbf{I}$ et $D_J = Id$ alors $\mathbf{S} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ est la matrice de covariance

- Montrer que $I_{\Delta u^{\perp}} = \mathbf{u}^T D_J \mathbf{X}^T D_I \mathbf{X} D_J \mathbf{u} =< \mathbf{X} D_J \mathbf{u}, \mathbf{X} D_J \mathbf{u} >_{D_I}$

u maximisant $I_{\Delta_{u^\perp}}$

- On sait que $I_{\Delta_{u^\perp}} = \mathbf{u}^T D_J \mathbf{X}^T D_I \mathbf{X} D_J \mathbf{u}$
- Trouver **u** maximisant $I_{\Delta_{u^\perp}}$ avec $||\mathbf{u}||^2_{D_J} = 1$
- Formulation du problème

$$\begin{cases} Max_{\Delta_{u^\perp}} \mathbf{u}^T (D_J \mathbf{X}^T D_I \mathbf{X} D_J) \mathbf{u} \\ ||\mathbf{u}||^2_{D_J} = 1 \end{cases} \quad \text{ou} \quad \begin{cases} Max_{\Delta_{u^\perp}} \mathbf{u}^T A \mathbf{u}, \text{ avec } A = D_J \mathbf{X}^T D_I \mathbf{X} D_J \\ ||\mathbf{u}||^2_{D_J} = 1 \end{cases}$$

- Le langrangien : $Lag = \mathbf{u}^T A \mathbf{u} - \lambda(\mathbf{u}^T D_J \mathbf{u} - 1)$
- $\frac{\partial Lag}{\partial \mathbf{u}} = 2A\mathbf{u} - 2\lambda D_J \mathbf{u} = 0$ d'où $A\mathbf{u} = \lambda D_J \mathbf{u}$,
- sachant $\mathbf{u}^T \mathbf{D}_J \mathbf{u} = 1$, on déduit $\lambda = \mathbf{u}^T A \mathbf{u}$, λ est donc le maximum recherché
- D_J étant inversible car définie positive $D_J^{-1} A \mathbf{u} = \lambda \mathbf{u}$
- **u** est vecteur propre de $\mathbf{D}_J^{-1} A = \mathbf{X}^T \mathbf{D}_I \mathbf{X} \mathbf{D}_J = \mathbf{S}$ correspondant à la plus grande valeur propre

Inertie expliquée par $I_{\Delta u^{\perp}}$

- Appelons $\mathbf{u_1}$ le vecteur $\mathbf{u}$ correspondant à la plus grande valeur propre appelée λ_1,
- Quel est le deuxième vecteur $\mathbf{u_2}$ qui

$$\begin{cases} Max_{\Delta_{u_2^{\perp}}} \mathbf{u_2}^T A \mathbf{u_2} \\ ||\mathbf{u_2}||^2_{D_J} = 1 \text{ et } \mathbf{u_2}^T D_J \mathbf{u_1} = 0 \end{cases}$$

- $Lag = \mathbf{u_2}^T A\mathbf{u_2} - \lambda_2(\mathbf{u}^T D_J \mathbf{u_2} - 1) - \mu \mathbf{u_2}^T D_J \mathbf{u_1}$

- $\frac{\partial Lag}{\partial \mathbf{u_2}} = 2A\mathbf{u_2} - 2\lambda_2 D_J \mathbf{u_2} - \mu D_J \mathbf{u_1} = 0$

- En multipliant par $\mathbf{u_1}^T$, on en déduit que $\mu = 0$

- $\mathbf{u_2}$ est le second vecteur propre de $\mathbf{D}_J^{-1} A = \mathbf{S}$ relatif à la seconde plus grande valeur λ_2

- Pour tout $\alpha \leq p$, $\mathbf{u}_\alpha$ est le vecteur propre de $\mathbf{D}_J^{-1} A = \mathbf{S}$ relatif à λ_α

Conséquence

1. On sait maintenant comment définir $E_k = \Delta u_1 \oplus \Delta u_2 \ldots \oplus \Delta u_k$
2. Les axes Δ_{u_α} sont appelés axes factoriels ou axes principaux
3. Cette étape passe par la diagonalisation de la matrice **S**
4. Il suffit de déterminer les coordonnées de la projection de tous les points du nuage sur chaque axe factoriel

Les composantes principales

- On notera dans la suite $\mathbf{c}_\alpha$ la αème composante principale
- $\mathbf{c}^\alpha = (c_1^\alpha, \ldots, c_n^\alpha)^T$, $\alpha = 1..p$ $\quad c_i^\alpha = < \boldsymbol{x}_i, u_\alpha >_{D_J}$
- $c^\alpha = \mathbf{X} D_J u_\alpha$ et la matrice des composantes principales $C = \mathbf{X} D_J U$
- Les composantes principales sont de nouvelles variables
- Formule de reconstitution : $X = CU^T D_J^{-1}$

$$X \approx \tilde{\mathbf{X}} = \sum_{\alpha=1}^{k} c^\alpha \mathbf{u}_\alpha^T \text{ avec } D_J = Id$$

- Conséquence : il s'agit d'un problème d'approximation Minimiser $||\mathbf{X} - \tilde{\mathbf{X}}||^2$

Application

Exercice

- Soit I un ensemble de 4 individus décrits par 3 variables quantitatives. On suppose que $D_I = \frac{1}{4} Id$. Soit A la matrice des données initiales. Le tableau obtenu après centrage en colonne est noté X. En effectuant l'ACP des individus (avec $D_J = Id$), on constate que le nuage des 4 individus est exactement dans le premier plan factoriel. Sachant que
  - les deux premiers vecteurs propres normés sont $u_1 = \frac{1}{\sqrt{2}}(1,1,0)^T$ et $u_2 = \frac{1}{\sqrt{3}}(1,-1,1)^T$
  - les deux coordonnées de ces individus sont obtenus à l'aide des deux composantes $c_1 = 2\sqrt{2}(1,1,-1,-1)^T$ et $c_2 = 2\sqrt{3}(-1,1,-1,1)^T$
- Déterminer le tableau X. En déduire la matrice intiale A sachant que les moyennes des 3 variables sont respectivement 4, 6 et 2.

Propriétés des composantes principales

1. Les composantes principales sont des combinaisons linéaires des variables initiales

$$c^{\alpha} = \mathbf{X}(D_J u_{\alpha}) = (\mathbf{x}^1, \ldots, \mathbf{x}^j)(D_J u_{\alpha}) = \sum_{j=1}^{p} a^j x^j \text{ avec } \mathbf{a} = (a^1, \ldots, a^p)^T = D_J \mathbf{u}_{\alpha}$$

2. Les composantes principales c^{α} sont centrées $\bar{c}^{\alpha} = 0$
3. Les composantes principales ne sont pas corrélées

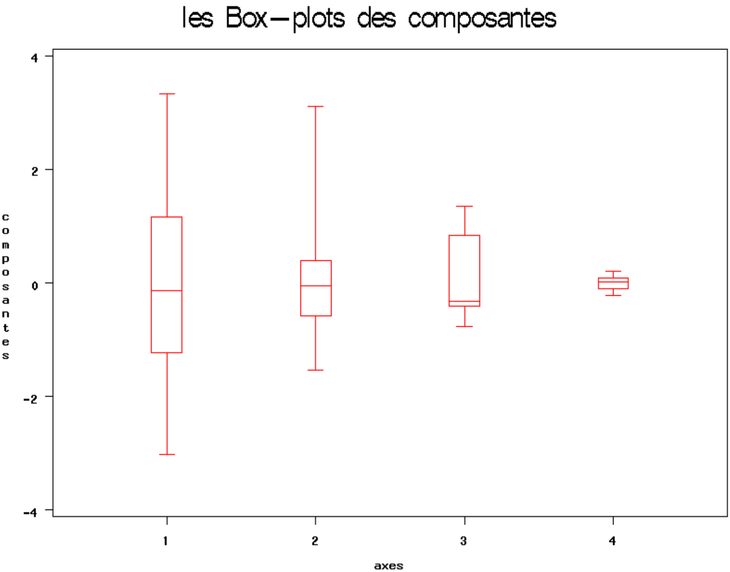
$$\begin{aligned}
Cov(c^{\alpha}, c^{\beta}) &= < \mathbf{X} D_J \mathbf{u}_{\alpha}, \mathbf{X} D_J \mathbf{u}_{\beta} >_{D_I} \\
&= (\mathbf{X} D_J \mathbf{u}_{\alpha})^T D_I \mathbf{X} D_J \mathbf{u}_{\beta} \\
&= \mathbf{u}_{\alpha}^T D_J (\mathbf{X}^T D_I \mathbf{X} D_J) \mathbf{u}_{\beta} \\
&= \mathbf{u}_{\alpha}^T D_J (S D_J \mathbf{u}_{\beta}) \\
&= \mathbf{u}_{\alpha}^T D_J (\lambda_{\beta} \mathbf{u}_{\beta}) \\
&= \lambda_{\beta} < \mathbf{u}_{\alpha}, \mathbf{u}_{\beta} >_{D_I} \\
&= \lambda_{\beta} \times 0 \text{ si } (\alpha \neq \beta)
\end{aligned}$$

4. La variance de $\mathbf{c}^{\alpha}$ est égale à λ_{α}

$$Var(c^{\alpha}) = \lambda_{\alpha} = \sum_{i=1}^{n} \pi_i (c_i^{\alpha})^2$$

Variance des composantes

- Exemple des données papillons



Indices d'aide à l'interprétation

- Qualité de la projection sur E_k (pourcentage d'inertie pris en compte par E_k) est évaluée par

$$\frac{\sum_{\alpha=1}^{k} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha}$$

- Part d'inertie expliquée par une axe Δu_α est exprimée donc par

$$\frac{\lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha}$$

- Qualité de représentation de $\boldsymbol{x}_i$ sur Δu_α est évaluée par

$$\frac{(c_i^\alpha)^2}{||\boldsymbol{x}_i||_{D_J}^2} = cos^2(\mathbf{x}_i, \mathbf{u}_\alpha)$$

- Contribution relative de $\boldsymbol{x}_i$ à un axe Δu_α (part d'inertie de Δu_α prise en compte ou expliquée par l'individu $\boldsymbol{x}_i$) est définie par

$$\frac{\pi_i(c_i^\alpha)^2}{\sum_i \pi_i(c_i^\alpha)^2} = \frac{\pi_i(c_i^\alpha)^2}{\lambda_\alpha}$$

ACP des individus

1. Calul de la matrice S
2. Diagonalisation de S
3. Normalisation des vecteurs propres
4. Calcul des composantes principales
5. Représentation graphique des individus
6. Construction des plans factoriels $(\Delta\mathbf{u}_\alpha, \Delta\mathbf{u}_\beta)$, $(\Delta\mathbf{u}_1, \Delta\mathbf{u}_2)$ est appelé premier plan factoriel.
7. exemple

ident	x^1	x^2	x^3	x^4
$i1$	19	49	33	39
$i2$	22	48	32	38
$i3$	23	49	29	39
$i4$	19	52	27	42
$i5$	21	49	31	39
$i6$	17	52	29	42
$i7$	17	51	31	41
$i8$	21	49	31	39
$i9$	21	51	27	41

Etapes de calcul, $D_I = \frac{1}{n} Id$ et $D_J = Id$

1. Centrage de la matrice des données
2. Calcul de la matrice **S**, dans ce cas $S = \frac{1}{9}\mathbf{X}^T\mathbf{X}$

$$\begin{pmatrix} 4 & -2 & 0 & -2 \\ -2 & 2 & -2 & 2 \\ 0 & -2 & 4 & -2 \\ -2 & 2 & -2 & 2 \end{pmatrix}$$

3. Diagonalisation de **S**
   - 8 et 4 sont deux valeurs propres simples et 0 est une valeur propre double
   - les vecteurs $(-1, 1, -1, 1)^T$ et $(-1, 0, 1, 0)^T$ sont vecteurs propres associés respectivement à 8 et 4
4. Normalisation de ces vecteurs $\mathbf{u_1} = \frac{1}{2}(-1, 1, -1, 1)^T$ et $\mathbf{u_2} = \frac{1}{\sqrt{2}}(-1, 0, 1, 0)^T$, Part d'inertie expliquée par $\Delta\mathbf{u_1} = 8/12$ et $\Delta\mathbf{u_2} = 4/12$
5. Calcul des deux composantes principales

$$c^1 = -\frac{1}{2}x^1 + \frac{1}{2}x^2 - \frac{1}{2}x^3 + \frac{1}{2}x^4$$

et

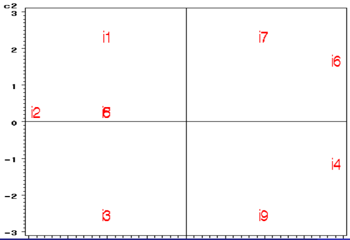
$$c^2 = -\frac{1}{\sqrt{2}}x^1 + \frac{1}{\sqrt{2}}x^3$$

Suite

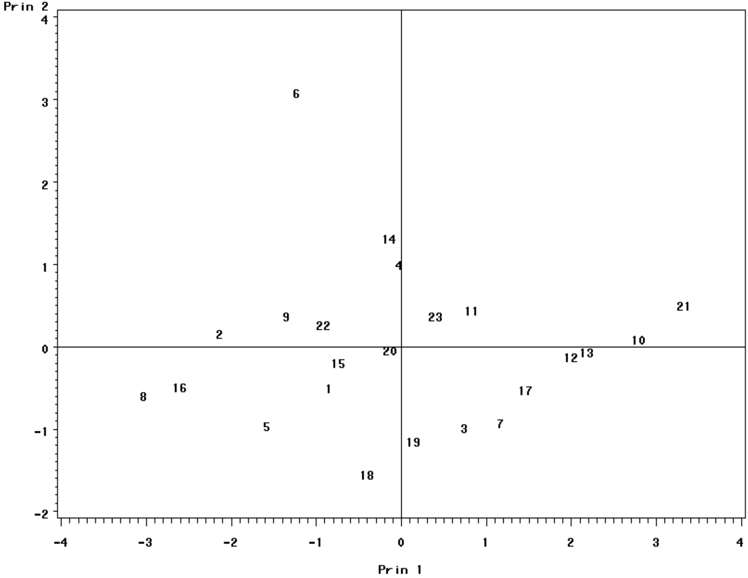
- Calcul de $C = (\mathbf{c}^1, \mathbf{c}^2)$

$$\begin{pmatrix} -2 & 2\sqrt{2} \\ -4 & 0 \\ -2 & -2\sqrt{2} \\ 4 & -\sqrt{2} \\ -2 & 0 \\ 4 & \sqrt{2} \\ 2 & 2\sqrt{2} \\ -2 & 0 \\ 2 & -2\sqrt{2} \end{pmatrix}$$

- Représentation graphique



ACP des papillons



ACP des variables

- Dans l'ACP des individus on diagonalisait $\mathbf{S} = X^T D_I X D_J$. Si $D_J = Id$ et $D_I = \frac{1}{n} Id$ alors la matrice à diagonaliser est :

$$\frac{1}{n} X^T X$$

- Dans l'ACP des variables on cherchera à diagonaliser $\mathbf{W} = X D_J X^T D_I$ (On notera $\mathbf{v}_\alpha$ vecteur propre associé λ_α (voir plus loin, SVD). Si $\mathbf{D}_J = Id$ et $\mathbf{D}_I = \frac{1}{n} Id$ alors la matrice à diagonalisée est :

$$\frac{1}{n} X X^T$$

Formules de transition $\mathbb{R}^p$ et $\mathbb{R}^n$

$$\begin{cases} \mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X} \mathbf{u}_\alpha \\ \mathbf{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}^T \mathbf{v}_\alpha \end{cases}$$

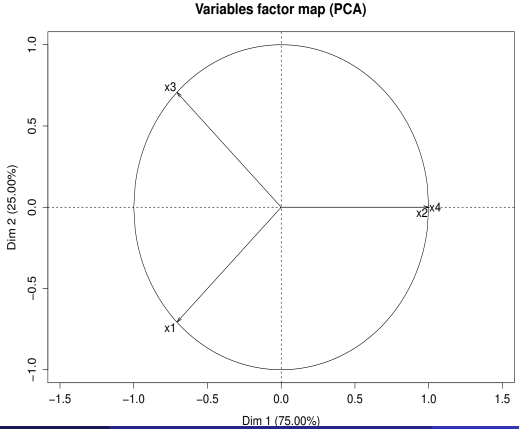
Les coordonnées des variables

$$\mathbf{d}_\alpha = \sqrt{\lambda_\alpha} \mathbf{u}_\alpha$$

$$\mathbf{d}^j_\alpha = < \mathbf{x}^j, c^\alpha >_{D_I}$$

Coordonnées des variables

$Variable$	$\|\|x^j\|\|_{D_I}$	d^1	d^2	$\rho(x^j, c^1)$	$\rho(x^j, c^2)$
x^1	2	$-\sqrt{2}$	$-\sqrt{2}$	$-\sqrt{2}/2$	$-\sqrt{2}/2$
x^2	$\sqrt{2}$	$\sqrt{2}$	0	1	0
x^3	2	$-\sqrt{2}$	$\sqrt{2}$	$-\sqrt{2}/2$	$\sqrt{2}/2$
x^4	$\sqrt{2}$	$\sqrt{2}$	0	1	0



Décomposition spectrale d'une matrice carrée

- Décomposition de Σ de taille $n \times n$: $\Sigma = U \Delta U^T$
  - les colonnes de U (matrice orthogonale) sont les vecteurs propres de Σ
  - les valeurs de Γ sont les valeurs propres

Décomposition en valeurs singulières d'une matrice, SVD

- Décomposition de $\mathbf{X}$ de taille $n \times p$: $\mathbf{X} = U \Gamma V^T$
  - U est une matrice orthogonale d'ordre $n \times p$ ($U^T U = Id$)
  - V est une matrice orthogonale d'ordre $p \times p$ ($V^T V = Id$)
  - Γ est une matrice diagonale d'ordre $p \times p$ avec des valeurs positives ou nulles appelées valeurs singulières
- Conséquences ($n \geq p$)
  1. les matrices carrées $\mathbf{X}^T\mathbf{X}$ d'ordre $p \times p$ et $\mathbf{X}\mathbf{X}^T$ d'ordre $n \times n$ s'écrivent :
  $$\mathbf{X} = U \Gamma V^T \Rightarrow \mathbf{X}\mathbf{X}^T = U \Gamma^2 U^T \text{ et } \mathbf{X}^T\mathbf{X} = V \Gamma^2 V^T$$
  2. Les vect. propres de $\mathbf{X}^T\mathbf{X}$ sont les colonnes de V et les val. propres sont les carrées des valeurs singulières
  3. $\mathbf{X}^T\mathbf{X}$ et $\mathbf{X}\mathbf{X}^T$ partagent p mêmes valeurs propres, les $(n-p)$ sont égales à 0
  $$(\mathbf{X}^T\mathbf{X})\mathbf{v} = \gamma^2 \mathbf{v} \text{ implique } (\mathbf{X}\mathbf{X}^T)\mathbf{X}\mathbf{v} = \gamma^2 \mathbf{X}\mathbf{v}$$
  4. ACP par SVD ($n << p$)

Vecteurs propres et valeurs singulières versus valeurs propres

- $\mathbf{X}$ centré et $k \leq min(n, p)$
- $\mathbf{X}_{n\times p} = \mathbf{U}_{n\times k}\Gamma_{k\times k}\mathbf{V}_{p\times k}^T$
- $\mathbf{X}^T\mathbf{X} = \mathbf{V}\Gamma^2\mathbf{V}^T = n\Sigma \Rightarrow \mathbf{V}\Gamma^2\mathbf{V}^T = n\mathbf{U}\Gamma\mathbf{U}^T = \mathbf{U}n\Delta\mathbf{U}^T$
- Conséquence : même vecteurs propres et valeurs singulières σ_i sont égales à $\sqrt{n\lambda_i}$

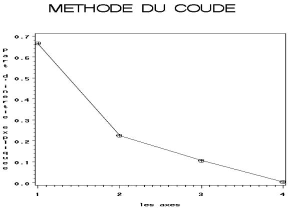
```
#ACP non normée de X_{9×4}
res.pca=PCA(data, scale=FALSE)
#centrer la matrice
centre=scale(data, center = TRUE, scale = FALSE)
#calculer la matrice de covariance
cov(centre)*8/9
#Comparer valeurs propres et vecteurs propres : Même vecteurs propres et relations entre valeurs propres

res.svd=svd(centre)
```

Latent Semantic Analysis (LSA/LSI)

- Soit $\mathbf{X}$ une matrice document-terme
- $\mathbf{X}_{n\times p} = U_{n\times k}\Gamma_{k\times k}V_{p\times k}^T$
- LSA est une SVD
- Les n lignes $U_{n\times k}\Gamma_{k\times k}$ sont les nouvelles coordonnées de chaque document après réduction de la dimension

Choix du nombre d'axes

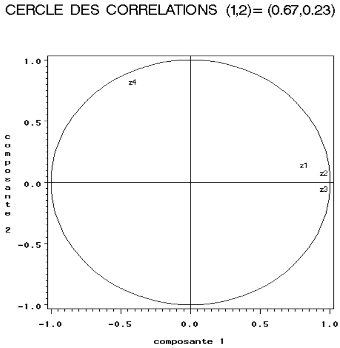
- Choisir un % est dénué de tout fondement: 10% sur 20 variables n'est pas le même intérêt sur un tableau à 500 variables
- Visualisation de la décroissance des valeurs propres : méthode du coude



- En cas de difficultés essayer $-log\frac{\lambda_k}{\lambda_1}$
- Dans le cas de données centrées-réduites
  - Critère de Kaiser : retenir uniquement les composantes correspondant aux valeurs propres > 1 (car Inertie moyenne$=I/p = p/p = 1$)
  - Karlis, Saporta and Spinakis (2003) proposent de garder les composantes correspondant aux valeurs vérifiant

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$$

Cercle des corrélations



$z1, z2, z3$ fortement corrélées avec la première composante et $z4$ avec la deuxième composante.

Choix du nombre d'axes

- Transformation est souvent nécessaire, passer nécessairement par la statistique exploratoire uni et bi-dimenssionnelle
- Sélection des variables, voire sélection des individus
- Introduction d'éléments supplémentaires (un élément supplémentaire ne participe pas à la création d'un axe)
  1. individus supplémentaires : Soit y_s appliquer d'abord la transformation, par exemple le centrage $x_s = (y_s^1 - \bar{x}^1, \ldots, y_s - \bar{x}^p)^T$ puis calculer les coordonnées
  $$< x_s, \mathbf{u}_\alpha >_{D_J}$$
  2. variables quantitative : Soit y^s, si on note $\bar{y}^s$ sa moyenne, par exemple le centrage implique $x^s = (y_s^1 - \bar{y}^1, \ldots, y_s - \bar{y}^p)^T$, les coordonées sont définis par
  $$< x^s, \mathbf{v}_\alpha >_{D_I}$$
  3. variable qualitative : centres de gravité pour chaque classe
  4. individus et variables supplémentaires
- Interpréter les axes factoriels à l'aide du cercle des corrélations
- Sélection les individus ayant le plus contribuer à la création des axes
- Utiliser les qualités de représentation pour enrichir votre interprétation et aussi pour une éventuelle appréciation des proximités entre les individus

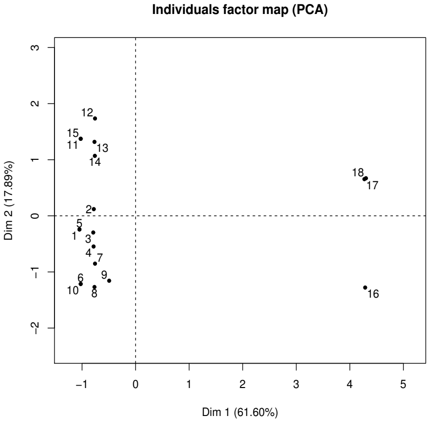
Effet de taille

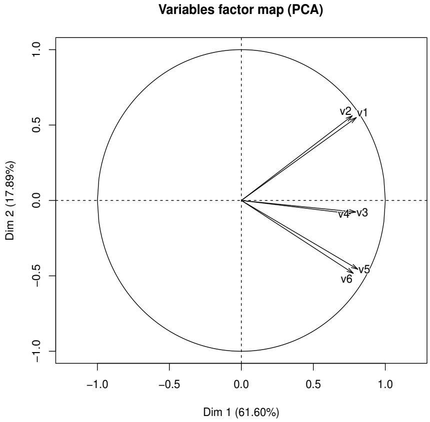
- Les variables peuvent être toutes situées du même côté de l'un axe factoriel. Une telle disposition apparaît lorsque toutes les variables sont corrélées positivement entre elles. Si pour un individu, une variable prend une valeur forte, toutes les autres variables prennent également une valeur forte
- Cette caractéristique est présente le plus souvent sur le premier axe factoriel. On parle d'effet de "taille" ou facteur de taille. L'orthogonalité des axes fait qu'il ne peut exister qu'un seul facteur taille. La deuxième composante est un caractère "forme"
- Recheche de groupes naturels de variables (problème de classification de variables) = pivoter les axes de l'ACP, ce qui implique un changement de répartition de la variance (illustration)

Classification à partir des composantes principales ?

- Souvent on a tendance à faire une ACP et à partir de des premiers axes on applique une classification. Attention une structure en classes peut être évidente en le plan factoriel (1,5) et complétement inexistante en utilisant (1,2)
- Souvent on cherche à visualiser les classes sur les plans factoriels alors qu'il est plus logique de le faire à l'aide d'une méthode d'analyse factorielle discriminante (apprentissage supervisé).

```
v1 <- c(1,1,1,1,1,1,1,1,1,1,3,3,3,3,3,4,5,6)
v2 <- c(1,2,1,1,1,1,2,1,2,1,3,4,3,3,3,4,6,5)
v3 <- c(3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,5,4,6)
v4 <- c(3,3,4,3,3,1,1,2,1,1,1,1,2,1,1,5,6,4)
v5 <- c(1,1,1,1,1,3,3,3,3,3,1,1,1,1,1,6,4,5)
v6 <- c(1,1,1,2,1,3,3,3,4,3,1,1,1,2,1,6,5,4)
m1 <- cbind(v1,v2,v3,v4,v5,v6)
```





Analyse factorielle

Rotation des axes

library(FactoMineR)
res.pca=PCA(m1)

	$Dim.1$	$Dim.2$	$Dim.3$	$Dim.4$	$Dim.5$
$v1$	0.7948541	0.53569995	−0.2338656	−0.1070259	−0.10333277
$v2$	0.7650331	0.53822219	−0.3062290	0.1258816	0.10470588
$v3$	0.7976654	−0.01559018	0.5518224	−0.2265563	0.08632718
$v4$	0.7594380	−0.02259678	0.6004697	0.2382473	−0.07199173
$v5$	0.8112499	−0.48166033	−0.2903922	−0.1111402	−0.06132028
$v6$	0.7794997	−0.51290340	−0.3219653	0.1153344	0.05427318

#Analyse factorielle avec varimax
factanal(m1, factors = 3)

	$Factor1$	$Factor2$	$Factor3$
$v1$	0.944	0.182	0.267
$v2$	0.905	0.235	0.159
$v3$	0.236	0.210	0.946
$v4$	0.180	0.242	0.828
$v5$	0.242	0.881	0.286
$v6$	0.193	0.959	0.196

Outline

Table de contingence

- Ensemble d'individus décrits par deux variables qualitatives
- Tableau croisé
- Tableau à valeurs positives assimilé à un tableau de contingence si la somme des lignes ou des clonnes a un sens
  - Tableau : individus décrits par une variable prenant des valeurs à differentes étapes
  - Tableau de pourcentages

exemple

	prof	tran	home	child	shop	wash	meal	sleep	tv	leis
maus	610	140	60	10	120	95	115	760	175	315
waus	475	90	250	30	140	120	100	775	115	305
wnaus	10	0	495	110	170	110	130	785	160	430
mnsus	615	141	65	10	115	90	115	765	180	305
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
mnsea	652	133	134	22	68	94	102	762	122	310
msea	627	148	68	0	88	92	86	770	58	463
wsea	433	86	296	21	128	102	94	758	58	379

I : types of population and J : variety of activities, x_{ij}: amount of time spent on a variety of activities j by i during a given time period j

Données normalisées

- $\mathbf{X}$ d'ordre $n \times p$ est définie par $\mathbf{X} = \{(x_{ij}); i \in I, j \in J\}$ I est une variable qualitative à n modalités (catégories) et J une variable qualitative à p modalités
- Notons la somme des lignes et des colonnes de $\mathbf{X}$ par $x_{i.} = \sum_{j=1}^{p} x_{ij}$ and $x_{.j} = \sum_{i=1}^{n} x_{ij}$ et la somme totale des valeurs toutes les données $N = \sum_{ij} x_{ij}$.
- La matrice des pourcentages $\mathbf{F} = \{(f_{ij} = x_{ij}/N); i \in I, j \in J\}$
- Les fréquences marginales $f_{i.} = \sum_j f_{ij}$ et $f_{.j} = \sum_i f_{ij}$
- On note $\mathbf{D}_I = Diag(f_{1.}, \dots, f_{n.})$ la matrice associée aux poids des lignes et par $\mathbf{D}_J = Diag(f_{.1}, \dots, f_{.p})$ la matrice diagonale asscociée aux poids des colonnes.

Nuages des profils lignes

- Nuage des lignes $N(I) = \{(f_i^J; f_{i.}); i = 1, \dots, n\}$, les profils lignes sont définis par $f_{iJ} = (f_{i1}/f_{i.}, \dots, f_{ip}/f_{i.})^T$
- Matrice des profils lignes : $\mathbf{D}_I^{-1}\mathbf{F}$
- Centre de gravité $f_J = (f_{.1}, \dots, f_{.p})^T$
- Sur $N(I)$ nous utilisons la métrique $Diag(1/f_{.1}, \dots, 1/f_{.p}) = \mathbf{D}_J^{-1}$
- Distance entre deux modalités $d_{\chi^2}(i, i') = d_{\chi^2}(f_i^J, f_{i'}^J) = \sum_{j=1}^{p} \frac{1}{f_{.j}} (\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}})^2$

Etude de la dispersion de $N(I)$ et $N(J)$

- Critère de $\chi^2(I,J) = \sum_{i,j} \frac{\text{(effectifs observés}-\text{effectifs théoriques)}^2}{\text{effectifs théoriques}}$
- Inertie du nuage $N(I)$

$$
\begin{aligned}
Inertie(N(I)) &= \sum_i f_{i.} d_{\chi^2}(f_{iJ}, fJ) \\
&= \sum_i f_{i.} \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2 \\
&= \sum_{i,j} \frac{f_{i.}}{f_{.j}} \left(\frac{f_{ij} - f_{i.} f_{.j}}{f_{i.}}\right)^2 \\
&= \sum_{i,j} \frac{1}{f_{i.} f_{.j}} (f_{ij} - f_{i.} f_{.j})^2 \\
&= \frac{1}{N} \sum_{i,j} \frac{1}{N f_{i.} f_{.j}} (N f_{ij} - N f_{i.} f_{.j})^2 \\
&= \frac{1}{N} \chi^2(I,J)
\end{aligned}
$$

- Inertie du nuage $N(J)$: $Inertie(N(J)) = \sum_j f_{.j} d_{\chi^2}(f_{Ij}, f_I) = Inertie(N(I))$

Nuages des profils colonnes

- Nuage des colonnes $N(J) = \{(f_I^j; f_{.j}); j = 1, \ldots, p\}$ Les profils colonnes sont définis par $f_{Ij} = (f_{1j}/f_{.j}, \ldots, f_{nj}/f_{.j})^T$
- Matrice des profils colonnes $\mathbf{D}_J^{-1}\mathbf{F}^T$
- Sur $N(J)$ nous utilisons la métrique $Diag(1/f_{1.}, \ldots, 1/f_{n.})^T = \mathbf{D}_I^{-1}$
- Distance entre deux modalités

$$d_{\chi^2}(j,j') = d_{\chi^2}(f_I^j, f_I^{j'}) = \sum_{i=1}^{n} \frac{1}{f_{i.}}(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j}})^2$$

AFC est une double ACP

- Les résultats sont obtenus à partir de la diagonalisation de $\mathbf{S} = \mathbf{F}^T\mathbf{D}_I^{-1}\mathbf{F}\mathbf{D}_J^{-1}$, les résultats à partir de $\mathbf{W} = \mathbf{F}\mathbf{D}_J^{-1}\mathbf{F}^T\mathbf{D}_I^{-1}$ peuvent être déduits.
- Tous les valeurs propres sont inférieures ou égales à 1
- Nous avons ces relations $\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{F}\mathbf{D}_J^{-1}\mathbf{u}_\alpha$ et $\mathbf{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{F}^T\mathbf{D}_I^{-1}\mathbf{v}_\alpha$
- centrage ou pas ?

Formules de transition $\mathbb{R}^p$ et $\mathbb{R}^n$

- Composantes principales $\mathbf{c}^\alpha = \mathbf{D}_I^{-1}\mathbf{F}\mathbf{D}_J^{-1}\mathbf{u}_\alpha$ et $\mathbf{d}^\alpha = \mathbf{D}_J^{-1}\mathbf{F}^T\mathbf{D}_I^{-1}\mathbf{v}_\alpha$

$$\begin{cases} c_i^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \frac{f_{ij}}{f_{i.}f_{.j}} \mathbf{u}_{\alpha j} \\ d_j^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{f_{ij}}{f_{i.}f_{.j}} \mathbf{v}_{\alpha i} \end{cases}$$

- On en déduit les relations quasi-barycentriques suivantes

$$\begin{cases} c_i^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \frac{f_{ij}}{f_{i.}} d_j^\alpha \\ d_j^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{f_{ij}}{f_{.j}} c_i^\alpha \end{cases}$$

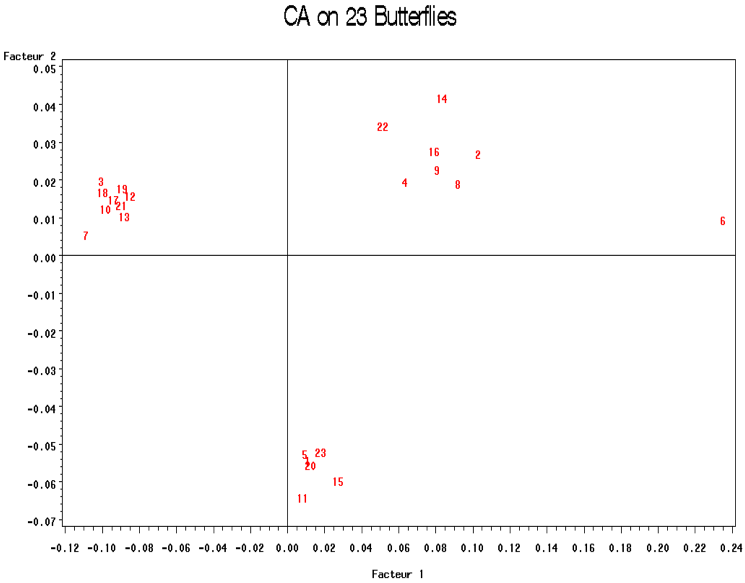
Formule de reconstitution

$$f_{ij} = f_{i.}f_{.j} \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha} c_i^\alpha d_j^\alpha$$

R : FactoMineR

res.ca=CA(papillons)

AFC des papillons



Outline

Individus décrits par des variables qualitatives

- L'analyse factorielle des correspondances multiples permet d'étudier le lien entre plusieurs variables de type qualitatif. Cette méthode est en fait une extension de l'AFC simple qui étudie le lien entre deux variables qualitatives.
- AFCM=AFC en cas de deux variables qualitatives
- Soit un ensemble d'individus décrits par plus de deux variables:
  - $\mathbf{X}$ de taille $n \times p$: n individus décrites par p variables comme c'est le cas des questionnaires. On notera m_q le nombre de modalités de la variable q et m étant le nombre total de modalités.
  - $\mathbf{Z}$ tableau disjonctif complet de taille $n \times m$
  - $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$ tableau de Burt de taille $m \times m$. Une juxtaposition de tables de contingence. Soit $\mathbf{D} = Diag(B)$; $\mathbf{D}$ exprime les effectifs de modalités
- $\mathbf{Z}$ et $\mathbf{B}$ peuvent être considérés comme des des tables de contingence

A partir de Z

- On a $\sum_{i,j} z_{ij} = np$
- $z_{i.} = \sum_j z_{ij} = p$
- $p_j = \frac{z_{.j}}{n}$ est la proportion des individus ayant choisi la modalité j

Inertie totale

- T : Inertie de $N(J)$) est égale à $\sum_{j=1}^{m} f_{i.} d^2_{\chi^2}(f_{Ij}, f_I) = \frac{m}{p} - 1$, que peut-on dire ?
- $T = \sum_{j=1}^{m} \frac{n - z_{.j}}{np}$ ou encore $T = \sum_{q=1}^{p} \frac{m_q - 1}{p}$ que peut-on dire ?

Matrice à diagonaliser

- Soit $\mathbf{F} = \frac{1}{np}\mathbf{Z}$ de terme général $f_{ij} = \frac{z_{ij}}{np}$
- Dans AFC on avait à diagonaliser $\mathbf{S} = \mathbf{F}^T \mathbf{D}_I^{-1} \mathbf{F} \mathbf{D}_J^{-1}$
- ici on $D_I = \frac{1}{n}\mathbf{I}$ car $\frac{1}{np} z_{i.} = \frac{p}{np} = \frac{1}{n}$ et $D_J = \frac{1}{np}\mathbf{D}$ car $f_{.j} = \frac{z_{.j}}{np}$
- On en déduit $\mathbf{D}_I^{-1} = n\mathbf{I}$ et $\mathbf{D}_J^{-1} = np\mathbf{D}^{-1}$
- Par conséquent et comme $\mathbf{Z} = np\mathbf{F}$ on a $\mathbf{S} = \mathbf{F}^T \mathbf{D}_I^{-1} \mathbf{F} \mathbf{D}_J^{-1} = \frac{1}{p} \mathbf{Z}^T \mathbf{Z} \mathbf{D}^{-1}$

Formules de transition

$$\begin{cases} c_i^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \frac{z_{ij}}{z_{i.}} d_j^\alpha \\ d_j^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{z_{ij}}{z_{.j}} c_i^\alpha \end{cases}$$

$$\begin{cases} \mathbf{c}^\alpha = \frac{1}{p\sqrt{\lambda_\alpha}} \mathbf{Z} \mathbf{d}^\alpha \\ \mathbf{d}^\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{D}^{-1} \mathbf{Z}^T \mathbf{c}^\alpha \end{cases}$$

Outline

AFDM

- Soit $\mathbf{X}$ matrice de données de n individus décrits par p variables quantitatives et d variables qualitatives (DM: Données mixtes)
- Les variables qualitatives sont transformées en tableaux de modalités : tableau disjonctif complet.
- Objectif de l'AFDM : Trouver les composantes liées aux variables originales au sens de la fonction objective à maximiser :

$$\sum_k \rho^2(\mathbf{c}, k) + \sum_q \eta^2(\mathbf{c}, q) \quad \rho\text{: coefficient de corrélation}, \eta\text{: rapport de corrélation}$$

- Chaque modalité d'une variable q est codée par 0 ou 1. Soit m le nombre total de modalités
- Chaque variable $k = 1, p$ quantitative est centrée réduite: $\frac{x_{ij} - \hat{x}}{s_j}$
- Chaque modalité $j = 1, m$ est normalisée $\frac{z_{ij}}{\sqrt{p_j}}$ où $p_j = \frac{z_{.k}}{n}$
- AFDM est une extension de ACP normée et AFCM
- Dans FactoMineR, utiliser FADM

Outline

AFM

Description

- Cette fois-ci **X** est décrite par des variables quantitatives/qualitatives structurées en groupes
- AFM est une extension de l'ACP si les variables sont quantitatives et l'AFCM si les variables sont qualitatives
- Cette méthode permet de confronter et comparer l'information apportée par plusieurs sources d'information
- Elle s'appuie d'abord sur des ACP/AFCM séparées
- Principe : Chaque groupe de variables est normalisée en divisant par $\sqrt{\lambda_1}$ la première valeur propre à partir d'une ACP réalisée en retenant uniquement le groupe de variables en question

Package et exemples d'application

- Dans FactoMineR, utiliser MFA
- Exemple : Des évaluations par plusieurs jury (Données vin)
- Exemple : Words décrits par word2vec, Glove etc. (A faire)

Outline

Méthodes factorielles

Objectifs : Réduction de la dimension + Interprétation

Description

- ACP appliquée à des données quantitatives
- AFC appliquée à tout tableau de contingence ou assimilé à un tableau de contingence
- AFCM est une extension de l'AFC avec plus de deux variables qualitatives
- AFDM sur des données mixtes (variables quantitatives et qualitatives)
- AFM sur des des variables quantitatives/qualitatives structurées en groupes
- AFD : Analyse factorielle discriminante (cours apprentissage supervisé)

Introduction des éléments supplémentaires dit actifs

- Importance des variables quantitatives ou qualitative supplémentaires dans les méthodes factorielles
- Importance des individus (lignes) supplémentaires
- Intérêt des rotations des axes dans l'ACP