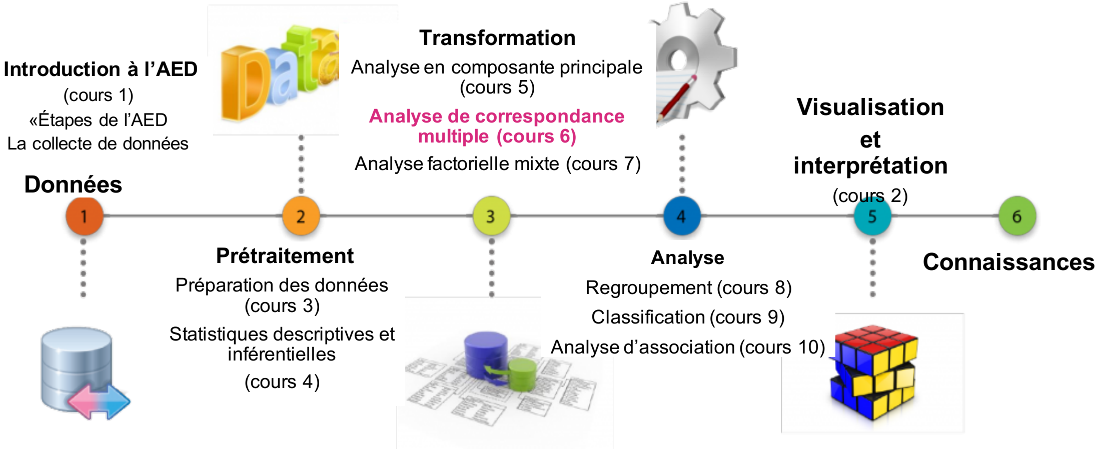
Cours 6 - Analyse des correspondances multiples (ACM)

Neila Mezghani

23 mars 2021

COLLÈGE DE
BOIS-DE-BOULOGNE


Introduction à l'AED
(cours 1)
«Étapes de l'AED
La collecte de données
Données
Transformation
Analyse en composante principale (cours 5)
Analyse de correspondance multiple (cours 6)
Analyse factorielle mixte (cours 7)
Visualisation
et
interprétation
(cours 2)
1
2
3
4
5
6
Prétraitement
Préparation des données (cours 3)
Statistiques descriptives et inférentielles
(cours 4)
Analyse
Regroupement (cours 8)
Classification (cours 9)
Analyse d'association (cours 10)
Connaissances

Plan du cours

Plan du cours

Définition

- L'analyse des correspondances multiples (ACM) est une méthode d'analyse factorielle adaptée aux données qualitatives (aussi appelées catégorielles).
- Une méthode de description statistique multidimensionelle d'un ensemble de données qualitatives. (L'équivalent de l'ACP pour les données quantitatives)

ACP — Analyse en composante principale : permet de traiter simultanément un nombre quelconque de variables, toutes quantitatives.

AFC — Analyse factorielle des correspondance : permet d'analyser la liaison existant entre deux variables qualitatives

ACM — Analyse de correspondance multiple : permet de traiter simultanément un nombre quelconque de variables, toutes qualitatives.

AFDM — Analyse factorielle de données mixtes (AFDM) permet de traiter un ensemble d'individus décrit par un ensemble de variables quantitatives et qualitatives simultanément.

Objectifs de l'ACM

L'ACM permet de faire des représentations graphiques du contenu d'un ensemble de données pour extraire des connaissance :

- Sur des **similitudes** ou **ressemblance** entre les individus
- Sur des **relations** ou **liaisons** entre les modalités des différentes variables qualitatives.

Questionnements

Les questions auxquelles permet de répondre une ACM sont :

- Quels sont les individus qui se ressemblent ? (proximité entre les individus)
- Sur quelles variables sont fondées ces ressemblances ?
- Quelles sont les associations entre les modalités ?

Plan du cours

Plan du cours

Les données : Matrice de données qualitatives $\mathbf{H}$

$$\mathbf{H} = \begin{array}{c|ccccc|}
 & 1 & \cdots & j & \cdots & p \\
\hline
1 & & & & & \\
\vdots & & & \vdots & & \\
i & \cdots & & h_{ij} & \cdots & \\
\vdots & & & \vdots & & \\
n & & & & & \\
\hline
\end{array}$$

- n est le nombre d'individus, p le nombre de variables qualitatives
- $h_{ij} \in M_j$ avec M_j qui l'ensemble des modalités de la $j$ème variable
- $m_j = card(M_j)$ le nombre de modalités de la variable j et $m = m_1 + \cdots + m_p$ le nombre total de modalités.

Tableau disjonctif complet $\mathbf{K}$ (TDC)

Sachant que nous disposons de $m = m_1 + \cdots + m_p$ modalités. On peut représenter les données sous la forme d'un tableau disjonctif complet $\mathbf{K}$:

$$\mathbf{K}= \begin{array}{c|ccccc|}
 & 1 & \cdots & s & \cdots & m \\
\hline
1 & & & & & \\
\vdots & & & \vdots & & \\
i & \cdots & & k_{is} & \cdots & \\
\vdots & & & \vdots & & \\
n & & & & & \\
\hline
\end{array}$$

Chaque colonne s est l'indicatrice de la modalité s (`pandas.get_dummies`)

- $k_{is} = 1$ si l'individu i possède la modalité s
- $k_{is} = 0$ sinon

Tableau de contingence

Le tableau disjonctif complet est traité comme un tableau de contigence :

$$\mathbf{K}=\begin{array}{c|ccccc|c} & 1 & \cdots & s & \cdots & m & \\ \hline 1 & & & & & & \\ \vdots & & & \vdots & & & \\ i & \cdots & & k_{is} & \cdots & & k_{i.}=p \\ \vdots & & & \vdots & & & \\ n & & & & & & \\ \hline & & & k_{.s}=n_s & & & k_{..}=np \end{array}$$

Dans ce cas on a :

- Les totaux en ligne sont constants $=$ au nombre p de variable
- Les totaux en colonne $=$ au nombre d'individus possédant la modalité s
- Le total général vaut $k_{..}=np$

Exemple : Matrice de données qualitatives

Chien	Taille	Velocite	Affection	Cote	Fonction
Beauceron	Taille++	Veloc++	Affec+	2	utilite
Basset	Taill-	Velo-	Affe-	4,5	chasse
Berger All	Taille++	Veloc++	Affec+	2,5	utilite
Boxer	Taille+	Veloc+	Affec+	3	compagnie
Bull-Dog	Taill-	Velo-	Affec+	1,5	compagnie
Bull-Mastif	Taille++	Velo-	Affe-	1	utilite
Caniche	Taill-	Veloc+	Affec+	4	compagnie
Labrador	Taille+	Veloc+	Affec+	3,5	chasse

← Données actives →

Exemple : Tableau disjonctif complet (TDC)

	m_1			m_2			m_3	
Chien	Taill-	Taille+	Taille++	Velo-	Veloc+	Veloc++	Affe-	Affec+
Beauceron	0	0	1	0	0	1	0	1
Basset	1	0	0	1	0	0	1	0
Berger All	0	0	1	0	0	1	0	1
Boxer	0	1	0	0	1	0	0	1
Bull-Dog	1	0	0	1	0	0	0	1
Bull-Mastif	0	0	1	1	0	0	1	0
Caniche	1	0	0	0	1	0	0	1
Labrador	0	1	0	0	1	0	0	1

Exemple : Tableau de contingence

$$k_{i.} = p$$

	m_1			m_2			m_3			
Chien	Taill-	Taille+	Taille++	Velo-	Veloc+	Veloc++	Affe-	Affec+		Somme
Beauceron	0	0	1	0	0	1	0	1		3
Basset	1	0	0	1	0	0	1	0		3
Berger All	0	0	1	0	0	1	0	1		3
Boxer	0	1	0	0	1	0	0	1		3
Bull-Dog	1	0	0	1	0	0	0	1		3
Bull-Mastif	0	0	1	1	0	0	1	0		3
Caniche	1	0	0	0	1	0	0	1		3
Labrador	0	1	0	0	1	0	0	1		3
Somme	3	2	3	3	3	2	2	6		24

$$k_{.s} = n_s$$

La matrice des profils-lignes $\mathbf{L}$

La matrice des profils-lignes est obtenue en suivant la relation :

$$\frac{k_{is}}{p} \tag{1}$$

$c = (\cdots, \frac{n_s}{np}, \cdots)^t \in \mathbb{R}^m$ est le vecteur des poids des modalités (aussi appelé profil-moyen).

$$\mathbf{L} = \begin{array}{c|ccccc|}
 & 1 & \cdots & s & \cdots & m \\ \hline
1 & & & & & \\
\vdots & & & \vdots & & \\
i & \cdots & & \frac{k_{is}}{p} & \cdots & \\
\vdots & & & \vdots & & \\
n & & & & & \\ \hline\hline
c^t & \cdots & & \frac{n_s}{np} & & \cdots
\end{array}$$

Exemple : La matrice des profils-lignes $\mathbb{L}$

Chien	Taill-	Taille+	Taille++	Velo-	Veloc+	Veloc++	Affe-	Affec+
Beauceron	0	0	0,333333	0	0	0,333333	0	0,333333
Basset	0,333333	0	0	0,333333	0	0	0,333333	0
Berger All	0	0	0,333333	0	0	0,333333	0	0,333333
Boxer	0	0,333333	0	0	0,333333	0	0	0,333333
Bull-Dog	0,333333	0	0	0,333333	0	0	0	0,333333
Bull-Mastif	0	0	0,333333	0,333333	0	0	0,333333	0
Caniche	0,333333	0	0	0	0,333333	0	0	0,333333
Labrador	0	0,333333	0	0	0,333333	0	0	0,333333
n_s	3	2	3	3	3	2	2	6
c	0,125	0,083333	0,125	0,125	0,125	0,083333	0,083333	0,25

La matrice des profils-colonnes $\mathbf{C}$

La matrice des profils-colonnes est obtenue en suivant la transformation suivante :

$$\frac{k_{is}}{n_s} \tag{2}$$

$r = (\cdots, \frac{1}{n}, \cdots)^t \in \mathbb{R}^n$ est le vecteur des poids des individus.

$$\mathbf{C}= \begin{array}{c|ccccc||c}
 & 1 & \cdots & \mathsf{s} & \cdots & \mathsf{m} & \mathbf{r} \\
\hline
1 & & & & & & \\
\vdots & & & \vdots & & & \\
\mathsf{i} & \cdots & & \frac{k_{is}}{n_s} & \cdots & & \frac{1}{n} \\
\vdots & & & \vdots & & & \\
\mathsf{n} & & & & & & \\
\hline
 & & & & & &
\end{array}$$

Exemple : La matrice des profils-colonnes C

Chien	Taill-	Taille+	Taille++	Velo-	Veloc+	Veloc++	Affe-	Affec+		r
Beauceron	0	0	0,333333	0	0	0,5	0	0,166667		0,125
Basset	0,333333	0	0	0,333333	0	0	0,5	0		0,125
Berger All	0	0	0,333333	0	0	0,5	0	0,166667		0,125
Boxer	0	0,5	0	0	0,333333	0	0	0,166667		0,125
Bull-Dog	0,333333	0	0	0,333333	0	0	0	0,166667		0,125
Bull-Mastif	0	0	0,333333	0,333333	0	0	0,5	0		0,125
Caniche	0,333333	0	0	0	0,333333	0	0	0,166667		0,125
Labrador	0	0,5	0	0	0,333333	0	0	0,166667		0,125

Distance du χ^2 (1)

- En ACP, les individus et les variables étaient les lignes et les colonnes d'une même matrice (la matrice des données centrées réduites)
- En ACM, les individus et les modalités sont les lignes et les colonnes de deux matrices différentes
- On construit ainsi la matrice des profils-lignes $\mathbf{L}$ et la matrice des profils-colonnes $\mathbf{C}$ pour faire les analyses

Distance du χ^2 (2)

- On utilise la distance χ^2 pour comparer deux individus décrits par deux points de $\mathbb{R}^m$ (deux profils lignes) ou deux modalités décrites par deux points de $\mathbb{R}^n$ (deux profils colonnes).
  - Distance du χ^2 entre deux individus : métrique $\mathbf{D}_c^{-1}$
  - Distance du χ^2 entre deux modalités : : métrique $\mathbf{D}_r^{-1}$

Distance du χ^2 entre deux individus : métrique $\mathbf{D}_c^{-1}$

La distance entre deux individus :

$$d^2(i,i') = \sum_{s=1}^{m} \frac{1}{f_{.s}} \left(\frac{k_{is} - k_{i's}}{p}\right)^2 = \frac{n}{p} \sum_{s=1}^{m} \frac{1}{n_s} (k_{is} - k_{i's})^2 \tag{3}$$

- Deux individus sont proches s'ils possèdent les même modalités

Exemple : Distance entre deux individus :

$$d^2(i,i') = \sum_{s=1}^{m} \frac{1}{f_{.s}} \left(\frac{k_{is} - k_{i's}}{p}\right)^2 = \frac{n}{p} \sum_{s=1}^{m} \frac{1}{n_s} (k_{is} - k_{i's})^2$$

$$d^2(\text{Beauceron, Basset}) = \frac{8}{3} \sum_{s=1}^{8} \frac{1}{n_s} (k_{is} - k_{i's})^2 = 5.77$$

$$d^2(\text{Caniche, Basset}) = \frac{8}{3} \sum_{s=1}^{8} \frac{1}{n_s} (k_{is} - k_{i's})^2 = 3.55$$

V

Distance du χ^2 entre deux modalités : : métrique $\mathbf{D}_r^{-1}$

La distance entre deux modalités :

$$d^2(s,s') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{k_{is}}{n_s} - \frac{k_{is'}}{n_{s'}}\right)^2 = n \sum_{i=1}^{n} \left(\frac{k_{is}}{n_s} - \frac{k_{is'}}{n_{s'}}\right)^2 \qquad (4)$$

- Deux modalités sont proches si elle sont possédées par les mêmes individus.

Exemple : Distance entre deux modalités

La distance entre deux modalités :

$$d^2(s,s') = \sum_{s=1}^{m} \frac{1}{f_{i.}} (\frac{k_{is}}{n_s} - \frac{k_{is'}}{n_{s'}})^2 = n \sum_{s=1}^{m} (\frac{k_{is}}{n_s} - \frac{k_{is'}}{n_{s'}})^2$$

$$d^2(\text{Taille-, Veloc-}) = 8 \sum_{i=1}^{8} (\frac{k_{i1}}{n_1} - \frac{k_{i4}}{n_4})^2 = 1,77$$

$$d^2(\text{Taille-, Veloc+}) = 8 \sum_{i=1}^{8} (\frac{k_{i1}}{n_1} - \frac{k_{i5}}{n_5})^2 = 3,55$$

Les individus qui partagent les caractéristiques (taille-, velocité-) sont plus nombreux que (taille-, vélocité+)

Plan du cours

Principe général de l'ACM

- Le principe général de la réalisation d'une ACM consiste à :
  - considérer les profils-lignes si on s'intéresse à la variabilité des individus et de leurs profils.
  - considérer les profils-colonnes si on s'intéresse aux liens entre les modalités.
- En pratique, on réalise une ACP du nuage des profils-lignes, en utilisant la métrique du chi-deux.
- On obtient alors la projection de la matrice X dans l'espace factoriel

$$X_r = \mathbf{D}_r^{-1} X \tag{5}$$

- (`fanalysis.mca`)

Plan du cours

Inertie totale

- Nous disposons d'une matrice des profils-lignes et d'une matrice des profils-colonnes.
- L'inertie de ces deux matrices sont égales.
- Elle peut être calculée selon la relation :

$$I = I(\mathbf{L}) = I(\mathbf{C}) = \frac{m}{p} - 1 \tag{6}$$

  m étant le nombre de modalités.
  p le nombre de variables.
- L'inertie dépend donc du nombre moyen de modalités par variable.

Nombre d'axes à retenir (1)

- De par la nature des données du TDC (variables binaires, colonnes démultipliées par le nombre de modalités, redondance de l'information), il est difficile de concentrer l'inertie sur les premiers axes factoriels de l'ACM.
- Cependant, on peut proposer un premier seuil pour éliminer certaines valeurs propres.
- En effet, en supposant que la matrice $\mathbf{K}$ est de plein rang, on a :

$$r = m - p$$

valeurs propres non nulles $\Longrightarrow$ On garde au plus r axes factoriels.

Nombre d'axes à retenir (2)

- Le pourcentage d'inertie expliquée par un axe α est :

$$I = \frac{\lambda_\alpha}{\lambda_1 + \lambda_2 + \cdots + \lambda_r} \times 100 \tag{7}$$

- Dans ce cas :

$$I = \sum_{k=1}^{r} \lambda_k = \frac{m}{p} - 1 \Leftrightarrow \frac{1}{r} \sum_{k=1}^{r} \lambda_k = \frac{1}{p} \tag{8}$$

- Autrement dit, une valeur propre vaut en moyenne $\frac{1}{p} \implies$ On peut donc sélectionner dans un premier temps les valeurs propres supérieures à ce seuil, c'est-à-dire celles qui contribuent plus que la moyenne.

Nombre d'axes à retenir : règle de Kaiser (3)

- En ACP, la règle de Kaiser consiste à retenir les vecteurs propres (axes) dont les valeurs propres sont $\lambda > 1$
- En ACM, on retient les valeurs propres λ_q telle que :

$$\lambda_q > \frac{1}{p} \tag{9}$$

Nombre d'axes à retenir : règle de coude (4)

- Il est difficile de concentrer l'inertie sur les premiers axes factoriels de l'ACM.
- Comme les pourcentages d'inertie expliquées par chaque axes factoriel sont petits par construction de l'ACM $\Longrightarrow$ on n'utilise pas en pratique le critère du pourcentage d'inertie pour choisir le nombre d'axes à retenir.

Exemple : Axes à retenir (1)

Nombre max de facteurs à retenir = 5 :

$$r = m - p = 5 \tag{10}$$

Les valeurs propres non nulles :

Les valeurs propres en pourcentage cumulé de la variance totale

Les valeurs propres en valeurs absolue

Id	Val.P	%	Cumul %
1	0,71	42,48	42,48
2	0,59	35,49	77,97
3	0,26	15,72	93,69
4	0,07	4,18	97,88
5	0,04	2,12	100,00

Les valeurs propres en pourcentage de la variance totale

Exemple : Axes à retenir (3)

- **Selon la règle de Kaiser** : on doit retenir les facteurs principaux dont les valeurs propres sont supérieurs à

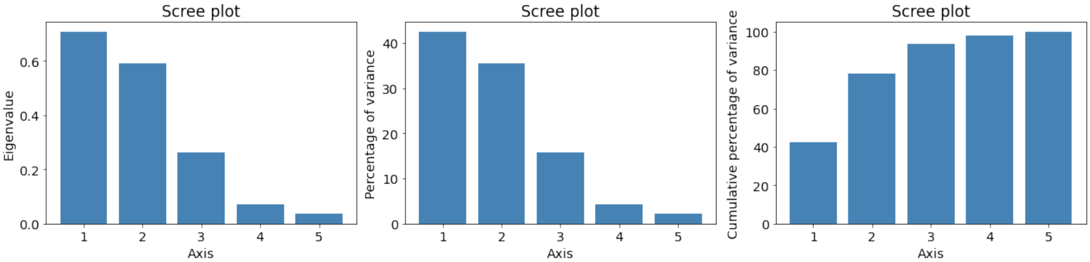
$$\lambda_i > \frac{1}{p} \implies \lambda_i > \frac{1}{3} = 0.33$$

- Les 2 premiers facteurs seraient donc les seuls pertinents avec des variances égales respectivement à 0.71 et 0.59.
- Ces 2 facteurs préservent 77.97% de l'information disponible.

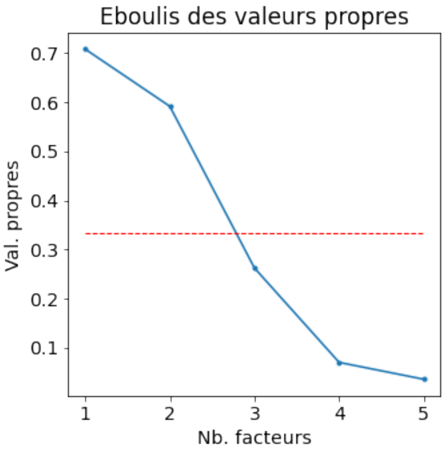
Id	Val.P	%	Cumul %
1	0,71	42,48	42,48
2	0,59	35,49	77,97
3	0,26	15,72	93,69
4	0,07	4,18	97,88
5	0,04	2,12	100,00

Exemple : Axes à retenir (2)

- **Selon la règle de Coude** : on peut se limiter à 3 ou 4 facteurs



Exemple : Axes à retenir (3)



- On observe une dispersion apparente des proportions d'inertie restituées sur les premiers facteurs.
- Ceci est dû entre autre aux colonnes démultipliées dans le codage disjonctif $\Longrightarrow$ certaines colonnes sont redondantes

Nombre d'axes à retenir : Correction de Benzécri (5)

- Il n'est pas possible en ACM d'interpréter l'inertie totale de façon aussi intéressante qu'en ACP (due entre autre à la redondance des données).
- Pour obtenir une indications plus réaliste sur la qualité des facteurs, nous devons utiliser un indicateur corrigé qui élimine le biais due à la redondance artificielle introduite dans les données suite à leur codage $\Longrightarrow$ la « correction de Benzécri ».

Nombre d'axes à retenir : Correction de Benzécri (6)

- Pour ce faire, on sélectionne les l valeurs propres supérieures au seuil $\frac{1}{p}$
- et on calcule les valeurs propres corrigées :

$$\hat{\lambda}_k = \left[\left(\frac{p}{p-1}\right)\left(\lambda_k - \frac{1}{p}\right)\right]^2 \qquad (11)$$

Exemple : Correction de Benzécri (2)

- on doit retenir les facteurs principaux dont les valeurs propres sont supérieurs à $\frac{1}{3} = 0.33 \implies$ on retient les 2 premiers facteurs
- **Suite à la correction** : on peut voire que le pourcentage d'inertie cumulé est plus important pour chacun des deux premiers axes.

Avant correction

Id	Val.P	%	Cumul %
1	0,71	42,48	42,48
2	0,59	35,49	77,97
3	0,26	15,72	93,69
4	0,07	4,18	97,88
5	0,04	2,12	100,00

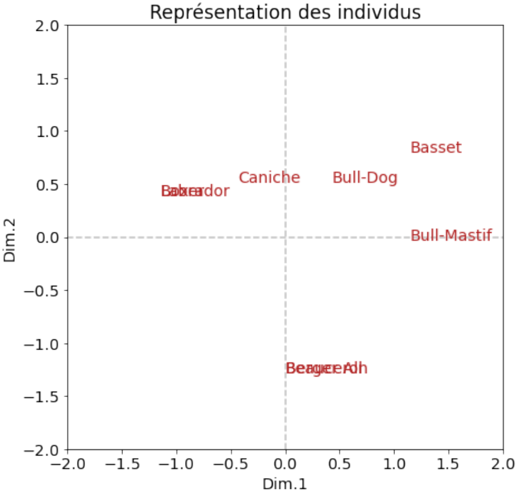
Après correction

Id	Val.P	%	Cumul %
1	0,32	67,81	67,81
2	0,15	32,19	100,00

Plan du cours

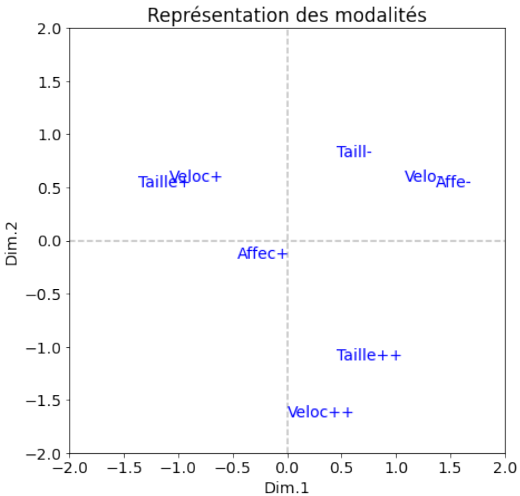
Plan du cours

Exemple : Représentation des individus



- On observe deux paires d'individus superposées : (Boxer, Labrador) et (Beauceron, Berger All).
- Ces paires partagent les mêmes descriptions dans les données initiales $\Longrightarrow$ on ne peut pas les distinguer dans la représentation factorielle.
- Diapo 24 : les caractéristiques du Caniche est plus proche de Basset que le Beauceron

Exemple : Représentation des modalités



- On observe trois groupes de variables reliés : (Velo-, Affec-), (Taille++, Veloc++) et (Taille+, Veloc+)
- Affec+ est une variable un peu isolée

Plan du cours

Contributions à l'inertie de l'axe α (1)

- Les contributions de l'individu i et de la modalité s à la l'inertie de l'axe α sont données par :

$$\begin{cases} Ctr_\alpha(i) = \frac{1}{n}\frac{x_{i\alpha}^2}{\lambda_\alpha} \\ \\ Ctr_\alpha(s) = \frac{n_s}{np}\frac{y_{s\alpha}^2}{\lambda_\alpha} \end{cases} \tag{12}$$

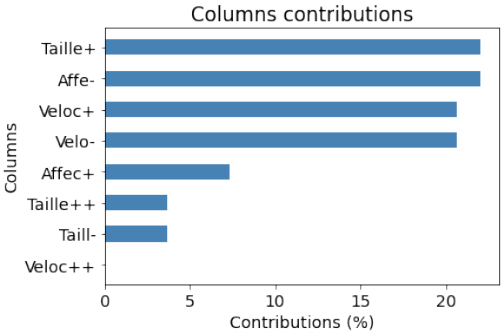
- $x_{i\alpha}$ est la coordonnée factorielle de l'individu i sur l'axe factoriel α
- $y_{s\alpha}$ est la coordonnée factorielle de la modalité s sur l'axe factoriel α

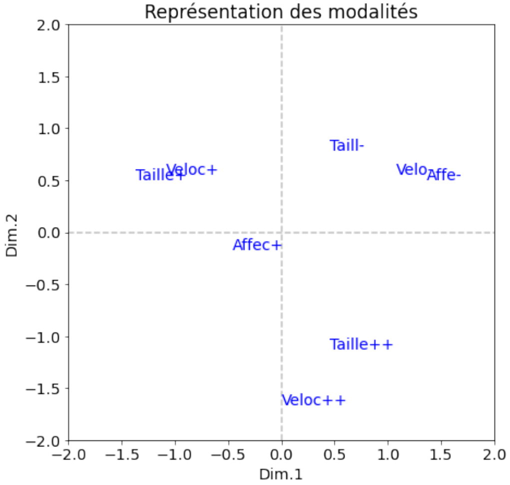
Contributions à l'inertie de l'axe α (2)

$$\begin{cases} Ctr_\alpha(i) = \frac{1}{n}\frac{x_{i\alpha}^2}{\lambda_\alpha} \\ \\ Ctr_\alpha(s) = \frac{n_s}{np}\frac{y_{s\alpha}^2}{\lambda_\alpha} \end{cases} \tag{13}$$

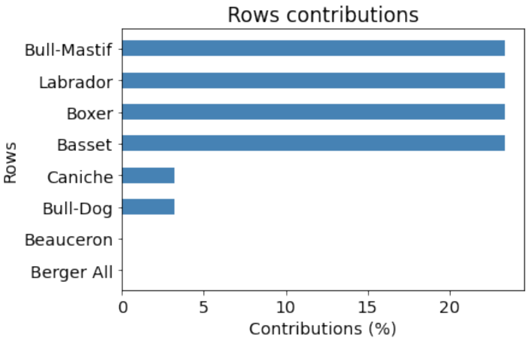
- Les individus i les plus excentrés ($x_{i\alpha} >>$) sur les plans factoriels sont ceux qui contribuent le plus.
- La contribution d'une modalité s à un axe dépend de sa fréquence n_s.
- Les modalités s les plus excentrées ($y_{s\alpha} >>$) ne sont pas nécessairement celles qui contribuent le plus : leur contribution dépend de leur fréquence n_s.

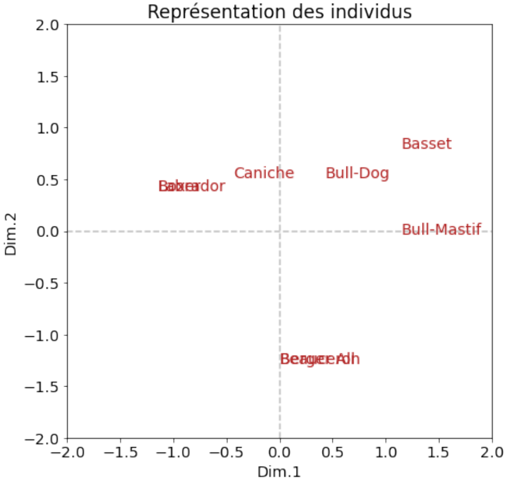
Exemple : Contributions aux axes factoriels (1)





Exemple : Contributions aux axes factoriels (2)





Plan du cours

Relations barycentriques (1)

- La réalisation d'une ACM à partir du tableau disjonctif complet (TDC) permet d'obtenir les deux relations suivantes :

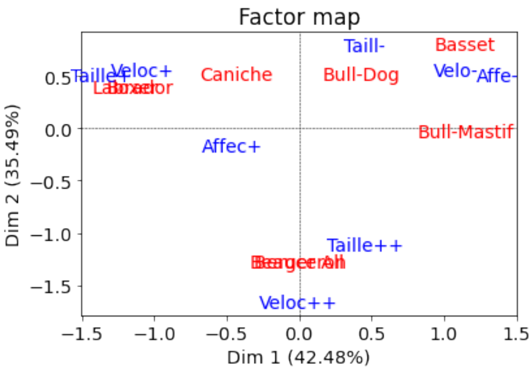
$$\begin{cases} x_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{p} \sum_{s \in S_i} y_{s\alpha} \\ \\ y_{s\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{p} \sum_{i \in I_s} x_{i\alpha} \end{cases} \tag{14}$$

avec :
- S_i est l'ensemble des modalités prises par l'individu i
- I_s est l'ensemble des individus qui possèdent la modalité s

Relations barycentriques (2)

- Ces relations quasi-barycentriques donnent la possibilés de représenter les individus et les modalités sur le même graphe $\Longrightarrow$ Sur un même graphique, on peut représenter les moyennes arthmétiques dilatées par $\frac{1}{\sqrt{\lambda_\alpha}}$
- Sur ces graphiques, à $\frac{1}{\sqrt{\lambda_\alpha}}$ près :
  - Les individus sont au centre des modalités qu'ils possèdent.
  - Les modalités sont au centre des individus qui les possèdent.

Exemple : Relations barycentriques



- Les individus sont en rouge et les modalités sont en bleu
- On peut voir les individus qui se ressemblent. Ex : BergerAll et Beaceron
- On peut voir les modalités sur lesquelles sont fondées ces ressemblances. Ex : BergerAll et Beaceron sont de Taille++ et Veloc++.

On peut aussi voir les associations entre les modalités (Taille+ et Veloc+) et (Taille++ et Veloc++)

En conclusion

Les connaissances que nous avons pu extraire de l'ACM :

- Les individus qui se ressemblent (proximité entre les individus)
- Les variables sur lesquelles sont fondées ces ressemblances
- Les associations entre les modalités