

Analyse des Correspondances Multiples – ACM

(ou Analyse Factorielle des Correspondances Multiples – AFCM)

Principes et pratique de l'ACM

Ricco RAKOTOMALALA

Université Lumière Lyon 2



PLAN

1. Position du problème
2. ACM : calculs via la matrice des indicatrices
3. ACM : calculs via le tableau de Burt
4. ACM : calculs via la matrice des profils lignes
5. **Pratique de l'ACM**
6. Les logiciels (SPAD, SAS, R et Tanagra)
7. Plus loin avec l'ACM (1) : analyse parallèle pour la détection des facteurs pertinents
8. Plus loin avec l'ACM (2) : analyse des relations non linéaires entre variables quantitatives
9. Bibliographie



Position du problème

Construire un nouveau système de représentation

(facteurs, axes factoriels : combinaisons linéaires des indicatrices des variables originelles)
qui permet synthétiser l'information



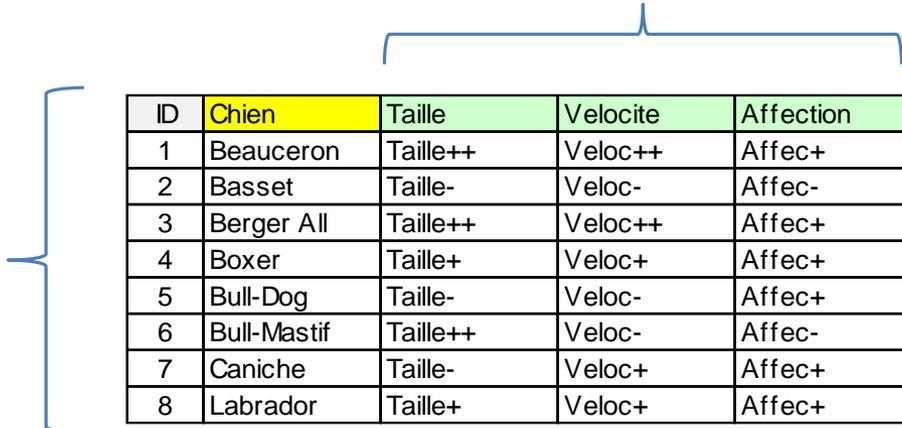
Variables « actives » qualitatives
c.-à-d. sont utilisées pour la
construction des facteurs

Extrait des données « races
canines »

(Tenenhaus, 2006 ; page 254)

$i : 1, \dots, n$
Individus actifs

$j : 1, \dots, p$



ID	Chien	Taille	Velocite	Affection
1	Beauceron	Taille++	Veloc++	Affec+
2	Basset	Taille-	Veloc-	Affec-
3	Berger All	Taille++	Veloc++	Affec+
4	Boxer	Taille+	Veloc+	Affec+
5	Bull-Dog	Taille-	Veloc-	Affec+
6	Bull-Mastif	Taille++	Veloc-	Affec-
7	Caniche	Taille-	Veloc+	Affec+
8	Labrador	Taille+	Veloc+	Affec+

Questions :

- (1) Quelles sont les chiens qui se ressemblent ? (proximité entre les individus)
- (2) Sur quelles caractéristiques sont fondées les ressemblances / dissemblances
- (3) Quelles sont les relations entre les modalités (distance)
- (4) Quelles sont les relations entre les variables



Tableau de données – Codage disjonctif complet

Le caractère ordinal de certaines variables (si elles le sont) est ignoré.

$$M = \sum_{j=1}^p m_j = 8$$

$m_1 = 3$ $m_2 = 3$ $m_3 = 2$

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	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	x_{ik}	1	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

$n = 8$

$p = 3$

$n_1 = 3$

$$\sum_{k=1}^M n_k = n \times p = 8 * 3 = 24$$



Position du problème (1)

Analyse des proximités entre les individus



Travailler sur les profils lignes

Distance du χ^2 entre les individus – Distance à l'origine

Exacerber les écarts
entre modalités rares

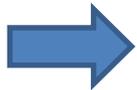
Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Beauceron	0.000	0.000	0.333	0.000	0.000	0.333	0.000	0.333
Basset	0.333	0.000	0.000	0.333	0.000	0.000	0.333	0.000
Berger All	0.000	0.000	0.333	0.000	0.000	0.333	0.000	0.333
Boxer	0.000	0.333	0.000	0.000	0.333	0.000	0.000	0.333
Bull-Dog	0.333	0.000	0.000	0.333	0.000	0.000	0.000	0.333
Bull-Mastif	0.000	0.000	0.333	0.333	0.000	0.000	0.333	0.000
Caniche	0.333	0.000	0.000	0.000	0.333	0.000	0.000	0.333
Labrador	0.000	0.333	0.000	0.000	0.333	0.000	0.000	0.333
Profil moyen	0.125	0.083	0.125	0.125	0.125	0.083	0.083	0.250

Barycentre (O)

$$\frac{n_k}{n \times p}$$

$$d^2(\text{beauceron}, \text{basset}) = \sum_{k=1}^M \frac{1}{\frac{n_k}{n \times p}} \left(\frac{x_{1k}}{p} - \frac{x_{2k}}{p} \right)^2 = \frac{1}{0.125} (0.000 - 0.333)^2 + \dots + \frac{1}{0.250} (0.333 - 0.000)^2 = 5.778$$

$$d^2(\text{basset}, \text{caniche}) = \frac{1}{0.125} (0.333 - 0.333)^2 + \frac{1}{0.083} (0.000 - 0.000)^2 + \dots + \frac{1}{0.250} (0.000 - 0.333)^2 = 3.556$$



Le basset a plus de caractères communs avec le caniche qu'avec le beauceron

$$d^2(\text{basset}, O) = \frac{1}{0.125} (0.333 - 0.125)^2 + \frac{1}{0.083} (0.333 - 0.083)^2 + \dots + \frac{1}{0.250} (0.000 - 0.250)^2 = 2.111$$

$$d^2(\text{caniche}, O) = \frac{1}{0.125} (0.333 - 0.125)^2 + \dots + \frac{1}{0.250} (0.333 - 0.250)^2 = 1.222$$



Le caniche est plus proche du « chien moyen » que le basset



Travailler sur les profils lignes

Inertie d'un individu – Inertie totale – Objectif de l'ACM

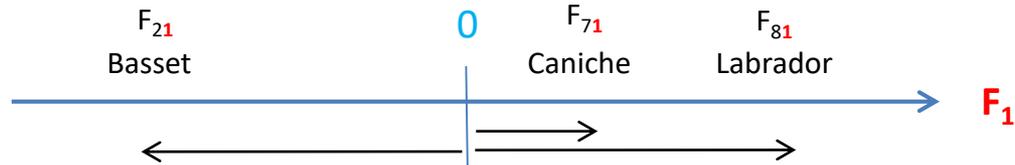
Inertie totale = distance 2 à 2 des individus = dispersion totale des individus

$$\begin{cases} I = \frac{1}{2n^2} \sum_{i=1}^n \sum_{i' \neq i}^n d^2(i, i') \\ I = \sum_{i=1}^n \frac{1}{n} d^2(i, 0) \end{cases}$$

Chien	Poids	$d^2(O)$	Inertie
Beauceron	0.125	1.667	0.208
Basset	0.125	2.111	0.264
Berger All	0.125	1.667	0.208
Boxer	0.125	1.667	0.208
Bull-Dog	0.125	1.222	0.153
Bull-Mastif	0.125	2.111	0.264
Caniche	0.125	1.222	0.153
Labrador	0.125	1.667	0.208

Inertie totale 1.667

Objectif de l'ACM : trouver un système de représentation (repère factoriel) qui préserve au mieux les distances entre les individus \Leftrightarrow qui permet de discerner le mieux possible les individus entre eux \Leftrightarrow qui maximise les (le carré des) écarts à l'origine.



1^{er} facteur : $\lambda_1 = \sum_{i=1}^n \frac{1}{n} \times F_{i1}^2$

F_{ih} : Coordonnée de l'individu i sur le facteur h
 λ_h : dispersion (inertie) associée au facteur h

$\frac{\lambda_1}{I}$ Part d'inertie restituée par le 1^{er} facteur

Nb. max de facteurs : $H_{max} = M - p$

Et : $\sum_{h=1}^{H_{max}} \lambda_h = I$ Décomposition orthogonale



Position du problème (2)

Analyse des associations entre les modalités



Travailler sur les profils colonnes

Distance du χ^2 entre les modalités – Distance à l'origine

$$\frac{x_{ik}}{n_k}$$

Barycentre : $\frac{p}{n \times p} = \frac{1}{n}$

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+	Profil moyen
Beauceron	0.000	0.000	0.333	0.000	0.000	0.500	0.000	0.167	0.125
Basset	0.333	0.000	0.000	0.333	0.000	0.000	0.500	0.000	0.125
Berger All	0.000	0.000	0.333	0.000	0.000	0.500	0.000	0.167	0.125
Boxer	0.000	0.500	0.000	0.000	0.333	0.000	0.000	0.167	0.125
Bull-Dog	0.333	0.000	0.000	0.333	0.000	0.000	0.000	0.167	0.125
Bull-Mastif	0.000	0.000	0.333	0.333	0.000	0.000	0.500	0.000	0.125
Caniche	0.333	0.000	0.000	0.000	0.333	0.000	0.000	0.167	0.125
Labrador	0.000	0.500	0.000	0.000	0.333	0.000	0.000	0.167	0.125

$$d^2(\text{taille-}, \text{velocite-}) = \sum_{i=1}^n \frac{1}{n} \left(\frac{x_{i1}}{n_1} - \frac{x_{i4}}{n_4} \right)^2 = \frac{1}{0.125} (0.000 - 0.000)^2 + \frac{1}{0.125} (0.333 - 0.333)^2 + \dots + \frac{1}{0.125} (0.000 - 0.000)^2 = 1.778$$

$$d^2(\text{taille-}, \text{velocite+}) = \frac{1}{0.125} (0.000 - 0.000)^2 + \frac{1}{0.125} (0.333 - 0.000)^2 + \dots + \frac{1}{0.125} (0.000 - 0.333)^2 = 3.556$$



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

$$d^2(\text{taille-}, O) = \frac{1}{0.125} (0.000 - 0.125)^2 + \frac{1}{0.125} (0.333 - 0.125)^2 + \dots + \frac{1}{0.125} (0.000 - 0.125)^2 = 1.6667$$

$$d^2(\text{taille+}, O) = \frac{1}{0.125} (0.000 - 0.125)^2 + \frac{1}{0.125} (0.000 - 0.125)^2 + \dots + \frac{1}{0.125} (0.500 - 0.125)^2 = 3.000$$



« Taille+ » est une caractéristique que l'on retrouve plus rarement que « Taille- » chez les chiens



Travailler sur les profils colonnes

Inertie d'une modalité – Inertie totale – Objectif de l'ACM

Inertie totale = dispersion totale des modalités

$$\omega_k = \frac{n_k}{n \times p}$$

	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Poids	0.125	0.083	0.125	0.125	0.125	0.083	0.083	0.250
d ² (Moda)	1.667	3.000	1.667	1.667	1.667	3.000	3.000	0.333

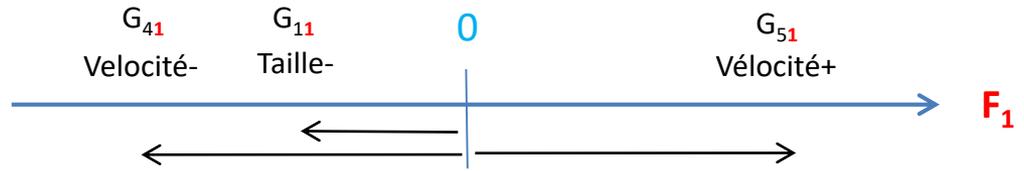
Inertie d'une modalité

Inertie	0.208	0.250	0.208	0.208	0.208	0.250	0.250	0.083
---------	-------	-------	-------	-------	-------	-------	-------	-------

Inertie totale	1.667
----------------	-------



Objectif de l'ACM : trouver un système de représentation (repère factoriel) qui préserve au mieux les distances entre les modalités ⇔ qui permet de discerner le mieux possible les modalités entre elles ⇔ qui maximise les (le carré des) écarts à l'origine.



1^{er} facteur : $\lambda_1 = \sum_{k=1}^M \omega_k \times G_{k1}^2$

G_{kh} : Coordonnée de la modalité k sur le facteur h
 λ_h : dispersion (inertie) associée au facteur h

Le 2nd facteur cherche à modéliser l'information non captée par le 1^{er} facteur c.-à-d. $(I - \lambda_1)$ etc.

Et : $\sum_{h=1}^{H_{max}} \lambda_h = I$ Décomposition orthogonale



$$\omega_k = \frac{n_k}{n \times p}$$

Le poids d'une modalité dépend de sa fréquence. Logique.

$$d^2(k) = \frac{n}{n_k} - 1$$

Une modalité est d'autant plus distante de l'origine qu'elle est rare.

$$I(k) = \omega_k \times d^2(k) = \frac{1}{p} \left(1 - \frac{n_k}{n} \right)$$

Une modalité contribue plus à l'inertie quand elle est rare.

$$I(j) = \sum_{k=1}^{m_j} d^2(k) = \frac{1}{p} (m_j - 1)$$

La contribution d'une variable à l'inertie est fonction du nombre de ses modalités.

$$I = \sum_{j=1}^p I(j) = \frac{M}{p} - 1$$

L'inertie totale ne dépend que des caractéristiques des variables : 'p' nombre de variables, 'M' nombre total de modalités, et le nombre moyen de modalités par variables (M/p).



Position du problème (3)

Analyse des variables



On peut réécrire l'analyse des modalités :

$$\begin{aligned}\lambda_1 &= \sum_{k=1}^M \omega_k \times G_{k1}^2 \\ &= \sum_{j=1}^p \sum_{\substack{k=1+ \\ l=1}}^{\sum_{l=1}^j m_l} \omega_k \times G_{k1}\end{aligned}$$



$$\lambda_1 = \frac{1}{p} \sum_{j=1}^p \eta^2(F_1, X_j)$$

Or...

- (1) G_{k1} est égale (à un facteur près) à la moyenne conditionnelle de la modalité « k » (des coordonnées des individus portant la modalité « k ») sur le facteur F1.
- (2) La moyenne des moyennes conditionnelles est nulle (les facteurs sont centrés)
- (3) La variance totale est la même quelle que soit la variable considérée (c'est celle du facteur)

L'ACM vise à maximiser la moyenne des rapports de corrélations (Tenenhaus, page 260)

Remarque : On note que, forcément, $\lambda_1 \leq 1$ puisque $\eta^2(F_1, X_j) \leq 1, \forall j$



Concrètement, le facteur est construit de manière à ce que, *globalement*, on distingue au mieux entre elles les modalités de chaque variable c.-à-d. pour chaque variable, ses modalités soient le plus étalées possible sur l'axe factoriel.



Calculs

Concrètement, comment obtenir les résultats de l'ACM à partir d'un tableau de données



Stratégie 1

Analyse factorielle des correspondances sur la matrice des indicatrices



Analyse des correspondances sur le tableau des indicatrices

Tableau des indicatrices = tableau de comptage un peu particulier

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	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

Tableau de valeurs positives :

- Les marges ont un sens
 - Les profils ont un sens
- On peut appliquer l'analyse factorielle des correspondances (AFC)

Avec l'AFC, on va analyser :

1. Les relations entre les modalités (colonnes)
2. Les proximités entre les individus (lignes)
3. Les association « individus x modalités »



C'est exactement le propos de l'ACM !



#chargement des données

```
canines <- read.table(file="canines-subset.txt",header=T,sep="\t",row.names=1,dec=".")
```

```
> summary(canines[1:3])
      Taille      Velocite      Affection
Taille- :3      Veloc- :3      Affec-:2
Taille+ :2      Veloc+ :3      Affec+:6
Taille++:3      Veloc++:2
```

#variables actives

```
summary(canines[1:3])
```

```
> print(canines01)
      TailleTaille- TailleTaille+ TailleTaille++ VelociteVeloc- VelociteVeloc+ VelociteVeloc++ AffectionAffec- AffectionAffec+
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           0
```

#codage 0/1 des variables

```
library(dummies)
```

```
canines01 <- dummy.data.frame(canines[1:3])
```

```
print(canines01)
```

AFC ne sait pas que c'est un tableau spécifique d'indicateurs, il produit 2 facteurs en trop, de v.p. nulles.

#charger le package pour l'AFC

```
library(FactoMineR)
```

#ACM par une AFC sur les indicatrices

```
canines.01.afc <- CA(canines01)
```

```
print(canines.01.afc$eig)
```

```
print(round(canines.01.afc$col$coord,4))
```

```
> print(canines.01.afc$eig)
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 7.080313e-01      4.248188e+01      42.48188
dim 2 5.914894e-01      3.548936e+01      77.97124
dim 3 2.619918e-01      1.571951e+01      93.69075
dim 4 6.974652e-02      4.184791e+00      97.87554
dim 5 3.540771e-02      2.124462e+00      100.00000
dim 6 7.785294e-33      4.671176e-31      100.00000
dim 7 3.382653e-33      2.029592e-31      100.00000

> print(round(canines.01.afc$col$coord,4))
      Dim 1   Dim 2   Dim 3   Dim 4   Dim 5
TailleTaille-  0.4559 -0.7881 -0.9004  0.1200  0.1122
TailleTaille+ -1.3676 -0.5008  0.8363 -0.3600  0.2231
TailleTaille++ 0.4559  1.1220  0.3428  0.1200 -0.2609
VelociteVeloc-  1.0815 -0.5547  0.0998 -0.4047 -0.1251
VelociteVeloc+ -1.0815 -0.5547  0.0998  0.4047 -0.1251
VelociteVeloc++ 0.0000  1.6642 -0.2993  0.0000  0.3752
AffectionAffec- 1.3676 -0.5008  0.8363  0.3600  0.2231
AffectionAffec+ -0.4559  0.1669 -0.2788 -0.1200 -0.0744
```

Coordonnées des colonnes (modalités). On peut obtenir tout aussi facilement les coordonnées des lignes (individus).

Stratégie 2

Analyse factorielle des correspondances sur le tableau de Burt



Analyse des correspondances sur le tableau de Burt

Tableau de Burt = un autre tableau de comptage un peu particulier

Tableau de Burt : tableau de croisement des variables, prises deux à deux

	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Taille-	3			2	1		1	2
Taille+		2			2			2
Taille++			3	1		2	1	2
Veloc-	2		1	3			2	1
Veloc+	1	2			3			3
Veloc++			2			2		2
Affec-	1		1	2			2	
Affec+	2	2	2	1	3	2		6

Tableau de valeurs positives :

- Les marges ont un sens
 - Les profils ont un sens
- On peut appliquer l'AFC

Avec l'AFC, on va analyser principalement les relations entre les modalités (en colonnes ET en lignes du tableau de Burt).

- (1) Mais... à l'instar de l'ACP où on diagonalise la matrice des corrélations, on pourra revenir sur les individus
- (2) Attention, les informations sont dupliquées (des individus sont comptés plusieurs fois), il faudra corriger les résultats de l'AFC.



ACM via le tableau de Burt

Calculs sous R

```
> print(canines.burt)
      Taille.Taille- Taille.Taille+ Taille.Taille++ Velocite.Veloc- Velocite.Veloc+ Velocite.Veloc++ Affection.Affec- Affection.Affec+
Taille.Taille-      3           0           0           2           1           0           1           2
Taille.Taille+      0           2           0           0           2           0           0           2
Taille.Taille++      0           0           3           1           0           2           1           2
Velocite.Veloc-      2           0           1           3           0           0           2           1
Velocite.Veloc+      1           2           0           0           3           0           0           3
Velocite.Veloc++      0           0           2           0           0           2           0           2
Affection.Affec-      1           0           1           2           0           0           2           0
Affection.Affec+      2           2           2           1           3           2           0           6
```

#construction du tableau de Burt

```
library(ade4)
```

```
canines.burt <- acm.burt(canines[1:3],canines[1:3])
```

```
print(canines.burt)
```

#AFC sur le tableau de Burt

```
library(FactoMineR)
```

```
canines.burt.afc <- CA(canines.burt)
```

#attention, des corrections sont nécessaires

μ_h sont les v.p. fournies par cette AFC

B_h sont les coordonnées factorielles des modalités

```
print(round(sqrt(canines.burt.afc$eig),5))
```

```
print(round(cbind(canines.burt.afc$col$coord[,1]/sqrt(sqrt(canines.burt.afc$eig[1,1])),ca
```

```
nines.burt.afc$col$coord[,2]/sqrt(sqrt(canines.burt.afc$eig[2,1]))),5))
```

```
eigenvalue
dim 1  0.70803
dim 2  0.59149
dim 3  0.26199
dim 4  0.06975
dim 5  0.03541
dim 6  0.00000
dim 7  0.00000
```

$$\lambda_h = \sqrt{\mu_h}$$

```
      [,1] [,2]
Taille.Taille-  0.45587 -0.78813
Taille.Taille+ -1.36762 -0.50076
Taille.Taille++ 0.45587  1.12197
Velocite.Veloc- 1.08146 -0.55474
Velocite.Veloc+ -1.08146 -0.55474
Velocite.Veloc++ 0.00000  1.66422
Affection.Affec- 1.36762 -0.50076
Affection.Affec+ -0.45587  0.16692
```

$$G_{kh} = \frac{B_{kh}}{\sqrt{\lambda_h}}$$



Stratégie 3 : passer par l'ACP

A l'instar de l'AFC, on peut obtenir les résultats de l'ACM via une ACP sur le tableau des profils lignes (ou sur le tableau des profils colonnes aussi d'ailleurs)



Distance du KHI-2 entre 2 individus, utilisée en ACM

$$d_{ACM}^2(i, i') = \sum_{k=1}^M \frac{1}{n_k} \left(\frac{x_{ik}}{p} - \frac{x_{i'k}}{p} \right)^2$$

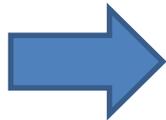
$n \times p$

Distance euclidienne en ACP normée

$$d_{ACP}^2(i, i') = \sum_{k=1}^M \frac{1}{\sigma_k^2} (x_{ik} - x_{i'k})^2$$

Où σ_k^2 est la variance de la $k^{\text{ème}}$ colonne des indicatrices

$$\sigma_k^2 = \frac{n_k(n - n_k)}{n^2}$$



Pour qu'il y ait équivalence, il faut pondérer la $k^{\text{ème}}$ indicatrice par u_k

$$u_k = \frac{n - n_k}{n \times p}$$

Ainsi, on peut obtenir les résultats de l'ACM via un programme d' ACP en appliquant cette pondération (sur les variables indicatrices)

$$d_{ACP \rightarrow ACM}^2(i, i') = \sum_{k=1}^M u_k \times \left[\frac{1}{\sigma_k^2} (x_{ik} - x_{i'k})^2 \right]$$



ACM via une ACP

ACP sur les profils lignes – Programme R

	TailleTaille-	TailleTaille+	TailleTaille++	VelociteVeloc-	VelociteVeloc+	VelociteVeloc++	AffectionAffec-	AffectionAffec+
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

#codage 0/1 des variables

library(dummies)

canines01 <- dummy.data.frame(canines[1:3])

#construction du tableau

#des profils lignes

```
profil <- fonction(x){
  res <- x/sum(x)
  return(res)
}
```

	TailleTaille-	TailleTaille+	TailleTaille++	VelociteVeloc-	VelociteVeloc+	VelociteVeloc++	AffectionAffec-	AffectionAffec+
Beauceron	0.000000	0.000000	0.333333	0.000000	0.000000	0.333333	0.000000	0.333333
Basset	0.333333	0.000000	0.000000	0.333333	0.000000	0.000000	0.333333	0.000000
Berger All	0.000000	0.000000	0.333333	0.000000	0.000000	0.333333	0.000000	0.333333
Boxer	0.000000	0.333333	0.000000	0.000000	0.333333	0.000000	0.000000	0.333333
Bull-Dog	0.333333	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000	0.333333
Bull-Mastif	0.000000	0.000000	0.333333	0.333333	0.000000	0.000000	0.333333	0.000000
Caniche	0.333333	0.000000	0.000000	0.000000	0.333333	0.000000	0.000000	0.333333
Labrador	0.000000	0.333333	0.000000	0.000000	0.333333	0.000000	0.000000	0.333333

tab.lignes <- t(apply(as.matrix(canines01),1,profil))

print(tab.lignes)

#calculer la pondération des indicatrices

n <- nrow(canines[1:3])

p <- ncol(canines[1:3])

```
> print(ponderation)
```

TailleTaille-	TailleTaille+	TailleTaille++	VelociteVeloc-	VelociteVeloc+	VelociteVeloc++	AffectionAffec-	AffectionAffec+
0.20833333	0.25000000	0.20833333	0.20833333	0.20833333	0.25000000	0.25000000	0.08333333

marge.ligne <- sapply(canines01,sum)

ponderation <- (n-marge.ligne)/(n*p)

print(ponderation)

#ACP normée sur les profils avec introduction de la pondération

library(FactoMiner)

canines.acp.lignes <- PCA(tab.lignes,ncp=2,col.w=ponderation)

print(canines.acp.lignes\$eig[,1:2])

print(canines.acp.lignes\$ind\$coord)



```
> print(round(canines.acp.lignes$eig[,1:2],5))
```

	eigenvalue	percentage of variance
comp 1	0.70803	42.48188
comp 2	0.59149	35.48936
comp 3	0.26199	15.71951
comp 4	0.06975	4.18479
comp 5	0.03541	2.12446
comp 6	0.00000	0.00000
comp 7	0.00000	0.00000

```
> print(round(canines.acp.lignes$ind$coord,5))
```

	Dim.1	Dim.2
Beauceron	0.00000	1.27992
Basset	1.15078	-0.79906
Berger All	0.00000	1.27992
Boxer	-1.15078	-0.38512
Bull-Dog	0.42841	-0.50968
Bull-Mastif	1.15078	0.02881
Caniche	-0.42841	-0.50968
Labrador	-1.15078	-0.38512

V.p. et coordonnées factorielles correspondent exactement à celles de l'ACM

Pratique de l'ACM

Que lire et comment lire les résultats de l'ACM



Sélection des facteurs pertinents
Tableau des valeurs propres et correction de Benzécri



Tableau des valeurs propres

Détection du nombre (H) de facteurs pertinents

Le tableau indique l'inertie reproduite par les facteurs

Le nombre maximum de facteurs est $H_{\max} = M - p = 8 - 3 = 5$

Axis	Lambda	% expliqué	% cumulé
1	0.7080	42.48%	42.48%
2	0.5915	35.49%	77.97%
3	0.2620	15.72%	93.69%
4	0.0697	4.18%	97.88%
5	0.0354	2.12%	100.00%

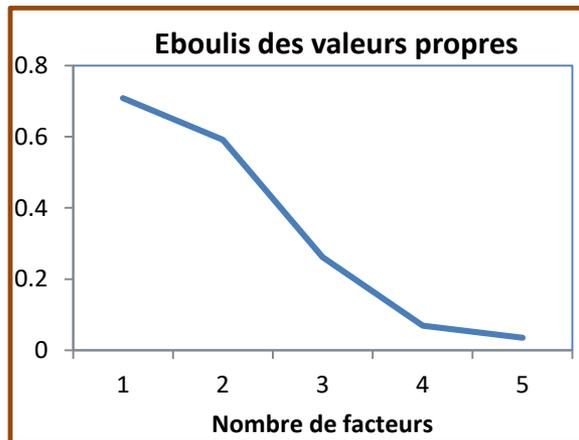
(1)

Un seuil de sélection possible : on sélectionne les facteurs dont les v.p. sont supérieures à la moyenne des v.p.

$$\frac{I}{H_{\max}} = \frac{\frac{M-1}{p}}{M-p} = \frac{M-p}{M-p} = \frac{1}{p} = \frac{1}{3} = 0.333$$



Il faudrait sélectionner H=2 facteurs apparemment



(2)

Eboulis des v.p. → utiliser la règle du coude



Sélectionner H=3 ou H=4 facteurs ?

Problème : De par la nature des données (les colonnes sont démultipliées par le codage 0/1, certaines sont redondantes), il est difficile de concentrer de l'inertie sur les premiers facteurs [comme $\lambda_1 \leq 1$, au mieux on ne disposerait que de $(1/I^*100)\%$ de l'inertie sur le 1^{er} facteur ; en pratique, l'éboulis des v.p. descend en pente douce]. Il faut utiliser un indicateur corrigé pour mieux rendre compte de l'intérêt des facteurs.



Tableau des valeurs propres

La correction de Benzécri – Point de vue du tableau de Burt

Inertie d'un tableau de contingence = ϕ^2 (cf. cours AFC)

Pourquoi une correction des v.p. ? En étudiant le tableau de Burt, on constate qu'une partie de l'information est triviale : le croisement d'une variable avec elle-même.

	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Taille-	3			2	1		1	2
Taille+		2.0	2		1.0	2		0.111
Taille++			3	1		2	1	2
Veloc-	2			3			2	1
Veloc+	1	1.0	2		2.0	3		0.556
Veloc++			2			2		2
Affec-	1	0.111		1	0.556		2	
Affec+	2		2	1		2		1.0

Inertie du tableau de Burt = Moyenne de toutes ces inerties, y compris celles sur la diagonale

$$I_{Burt} = \sum_h \mu_h = \sum_h \lambda_h^2 = \frac{2.0+1.0+0.111+1.0+\dots+1.0}{9} = 0.926$$

On souhaite s'intéresser à l'information « utile » c.-à-d. les relations entre les variables : les tableaux croisés hors diagonale

$$\frac{1.0+0.111+1.0+\dots+0.556}{6} = 0.556 = \frac{p}{p-1} \times \left(I_{Burt} - \frac{M-p}{p^2} \right)$$

En tenant compte de cette correction, on utilise une formule ajustée pour mieux rendre compte de l'inertie (l'information) restituée par les facteurs

$$\lambda'_h = \left[\left(\frac{p}{p-1} \right) \times \left(\lambda_h - \frac{1}{p} \right) \right]^2$$

uniquement pour les facteurs dont la v.p. est supérieure à la moyenne des v.p. (les autres facteurs ne sont pas intéressants)

$$\lambda_h > \frac{1}{p}$$



Démarche :

1. Calculer le seuil « 1/p »
2. Pour les facteurs dont la v.p. λ_h est supérieure à « 1/p »
 - Calculer la correction λ'_h
3. Faire la somme S' des λ'_h
4. Calculer les pourcentages d'inerties expliquées et cumulées à partir des λ'_h et de la somme S'

$$\lambda'_h = \left[\left(\frac{p}{p-1} \right) \times \left(\lambda_h - \frac{1}{p} \right) \right]^2$$

2 facteurs ont une v.p. supérieures à la moyenne

p		3			
Seuil		0.3333		Benzécri	
Axis	Lambda	Lambda'	% expliquée	%cumulée	
1	0.7080	0.3159	67.8%	67.8%	➔
2	0.5915	0.1499	32.2%	100.0%	
3	0.2620				
4	0.0697				
5	0.0354				
Somme	1.6667	0.4658461			

L'éboulis des v.p. donne des indications plus intéressantes dans ce cas, avec un « coude » plus net.

$$S' = \sum_h \lambda'_h$$



Analyse des modalités et des variables
Description et caractérisation



Analyse des modalités et des variables

Coordonnées factorielles, contributions, \cos^2

On a capté quasiment toute l'information véhiculée par les modalités (sauf pour « Taille- »)

$$\omega_k = \frac{n_k}{n \times p}$$

$$d^2(k) = \frac{n}{n_k} - 1$$

$$I(k) = \frac{1}{p_l} \left(1 - \frac{n_k}{n} \right)$$

Coordonnées de la modalité k sur le facteur h

$$G_{kh}$$

Values	Informations modalités			Coordonnées		CTR (%)		Cos ²		
Attribute = Value	Mass	Sq.Dist	Inertia	coord_1	coord_2	ctr_1	ctr_2	cos2_1	cos2_2	cumul(cos ²)
Taille = Taille++	0.125	1.667	0.208	-0.456	-1.122	3.669	26.603	0.125	0.755	0.880
Taille = Taille-	0.125	1.667	0.208	-0.456	0.788	3.669	13.127	0.125	0.373	0.497
Taille = Taille+	0.083	3.000	0.250	1.368	0.501	22.014	3.533	0.624	0.084	0.707
						Tot. ctr.	29.352	43.262		
Velocite = Veloc++	0.083	3.000	0.250	0.000	-1.664	0.000	39.021	0.000	0.923	0.923
Velocite = Veloc-	0.125	1.667	0.208	-1.081	0.555	20.648	6.503	0.702	0.185	0.886
Velocite = Veloc+	0.125	1.667	0.208	1.081	0.555	20.648	6.503	0.702	0.185	0.886
						Tot. ctr.	41.296	52.027		
Affection = Affec+	0.250	0.333	0.083	0.456	-0.167	7.338	1.178	0.624	0.084	0.707
Affection = Affec-	0.083	3.000	0.250	-1.368	0.501	22.014	3.533	0.624	0.084	0.707
						Tot. ctr.	29.352	4.710		

Informations a priori sur les modalités

Contributions : impact de la modalité sur la définition du facteur

$$CTR_{kh} = \frac{\omega_k \times G_{kh}^2}{\lambda_h}$$

Les contributions des modalités d'une variable s'additionnent → contribution d'une variable

COS² : qualité de représentation d'une modalité

$$COS_{kh}^2 = \frac{G_{kh}^2}{d^2(k)}$$

Les cos² s'additionnent d'un facteur à l'autre : qualité de la représentation sur les « h » premiers facteurs (information restituée)



Analyse des modalités et des variables

Représentation graphique - Valeurs test et rapports de corrélation

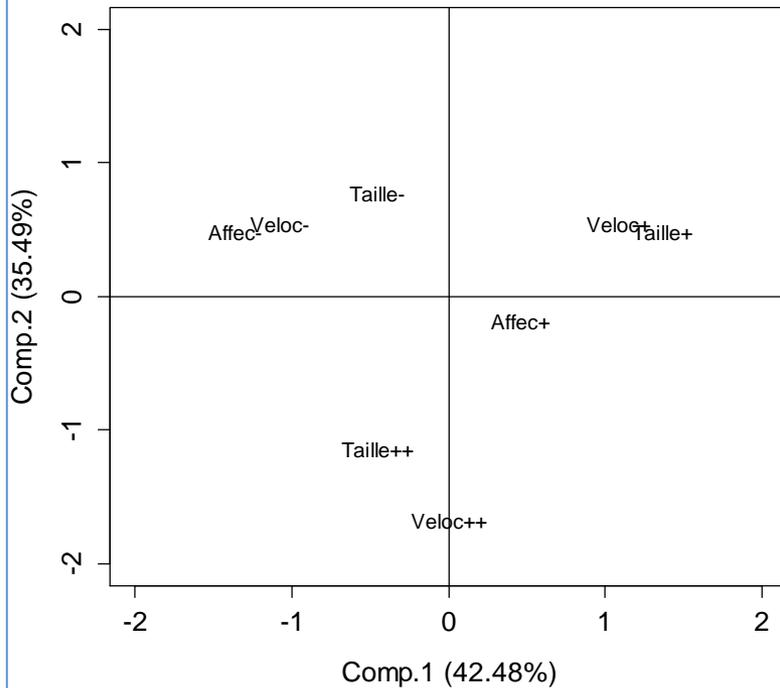
2 indicateurs supplémentaires pour chaque facteur :

- « **valeur test** » indique la significativité de l'écart de la modalité par rapport à l'origine
- « **rapport de corrélation** » indique l'étalement des modalités d'une variable (différenciation d'une ou plusieurs des modalités d'une même variable)

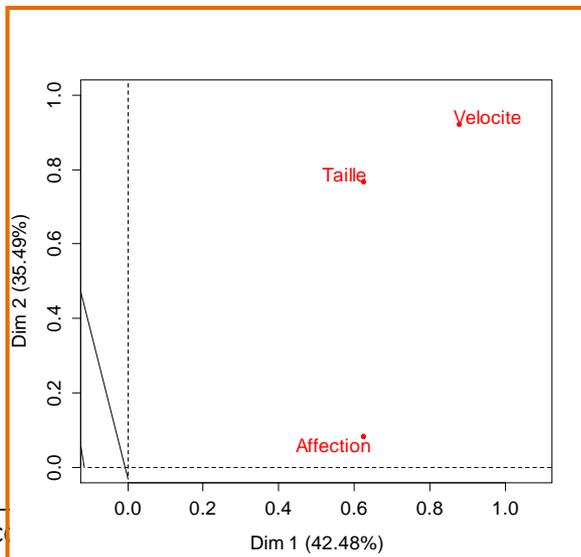
$$VT_{kh} = G_{kh} \sqrt{\frac{(n-1)n_k}{n-n_k}}$$

$$\eta_{jh}^2 = \sum_{k:X_j} \frac{n_k}{n} G_{kh}^2 = p \times \lambda_h \times CTR_{jh}$$

Projection des modalités



Values	n_k	Valeur test		Rapport de cor.	
		v-test_1	v-test_2	rap.1	rap.2
Attribute = Taille++	3	-0.934	-2.299	0.623	0.768
Attribute = Taille-	3	-0.934	1.615		
Attribute = Taille+	2	2.089	0.765		
Attribute = Veloc++	2	0	-2.542	0.877	0.923
Attribute = Veloc-	3	-2.216	1.137		
Attribute = Veloc+	3	2.216	1.137		
Attribute = Affec+	6	2.089	-0.765	0.623	0.084
Attribute = Affec-	2	-2.089	0.765		



Possibilité de représentation des variables

Analyse des modalités

Formule de reconstitution des distances

A la distance du KHI-2 utilisée dans l'espace originel, on substitue la distance euclidienne dans le repère factoriel. La précision de l'approximation dépend de la qualité de la représentation des modalités sur les facteurs considérés.

Values	Informations modalités			Coordonnées	
Attribute = Value	Mass	Sq. Dist	Inertia	coord_1	coord_2
Taille = Taille++	0.125	1.667	0.208	-0.456	-1.122
Taille = Taille-	0.125	1.667	0.208	-0.456	0.788
Taille = Taille+	0.083	3.000	0.250	1.368	0.501
Velocite = Veloc++	0.083	3.000	0.250	0.000	-1.664
Velocite = Veloc-	0.125	1.667	0.208	-1.081	0.555
Velocite = Veloc+	0.125	1.667	0.208	1.081	0.555
Affection = Affec+	0.250	0.333	0.083	0.456	-0.167
Affection = Affec-	0.083	3.000	0.250	-1.368	0.501

(À comparer avec « sq.dist. »)

Estimation de la distance à l'origine sur les 2 premiers facteurs

$$\hat{d}^2(\text{taille}++) = (-0.456)^2 + (-1.122)^2 = 1.467$$

$$\hat{d}^2(\text{taille}-) = (-0.456)^2 + (0.788)^2 = 0.829$$

$$\hat{d}^2(\text{veloc}++) = (0.000)^2 + (-1.664)^2 = 2.770$$

Faire le parallèle entre la précision de l'approximation et la colonne des \cos^2 cumulés (cf. tableau en amont)

Estimation de la distance entre modalités sur les 2 premiers facteurs

$$\hat{d}^2(\text{taille}++, \text{veloc}++) = (-0.456 - 0.000)^2 + [-1.122 - (-1.664)]^2 = 0.502$$

$$\hat{d}^2(\text{taille}++, \text{veloc}-) = [-0.456 - (-1.081)]^2 + [-1.122 - 0.555]^2 = 3.203$$

« Taille++ » est plus proche de « veloc++ » que de « veloc- ».



Analyse des individus
Description et caractérisation



Analyse des individus

Coordonnées factorielles, contributions, cos² – Représentation graphique

Coordonnées de l'individu « i » sur le facteur h

$$F_{ih}$$

↓

Chien	Coord.		CTR (%)		COS ²		Cumul
	Axe.1	Axe.2	Axe.1	Axe.2	Axe.1	Axe.2	
Beauceron	0.00	-1.28	0.00	34.62	0.00	0.98	0.98
Basset	-1.15	0.80	23.38	13.49	0.63	0.30	0.93
Berger All	0.00	-1.28	0.00	34.62	0.00	0.98	0.98
Boxer	1.15	0.39	23.38	3.13	0.79	0.09	0.88
Bull-Dog	-0.43	0.51	3.24	5.49	0.15	0.21	0.36
Bull-Mastif	-1.15	-0.03	23.38	0.02	0.63	0.00	0.63
Caniche	0.43	0.51	3.24	5.49	0.15	0.21	0.36
Labrador	1.15	0.39	23.38	3.13	0.79	0.09	0.88

Qualité de représentation des individus sur les facteurs

$$CTR_{ih} = \frac{F_{ih}^2}{d^2(i)}$$

↓

- d²(i) est la distance à l'origine de l'individu « i »
- les COS² s'additionnent d'un facteur à l'autre

Ex. Seules « Bull-Dog » et « Caniche » sont mal représentées sur les 2 premiers facteurs. Ces races présentent des spécificités non captées par les 2 premiers facteurs.

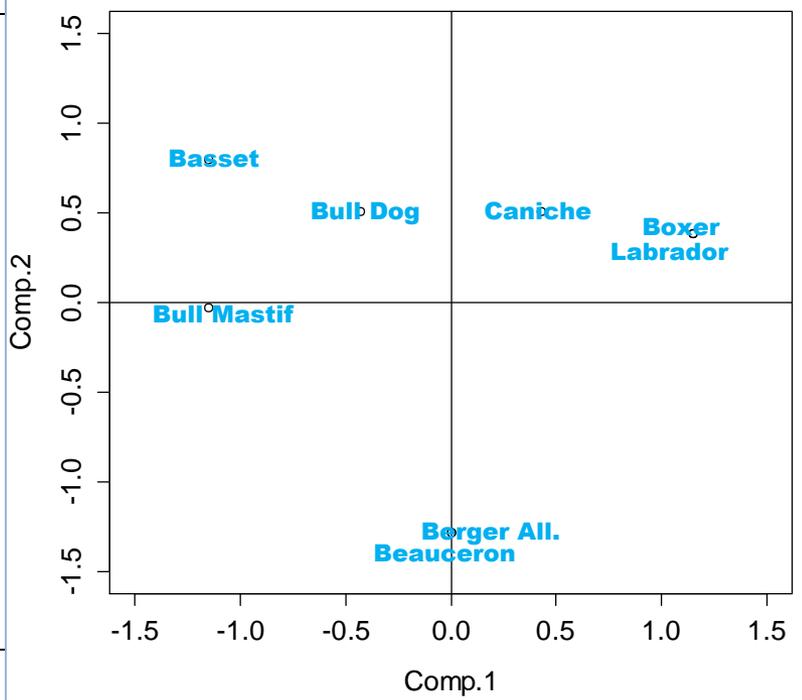
Contribution des individus dans la définition des facteurs

$$CTR_{ih} = \frac{1}{n} \times \frac{F_{ih}^2}{\lambda_h}$$

Les contribution des individus à un facteur s'additionnent.

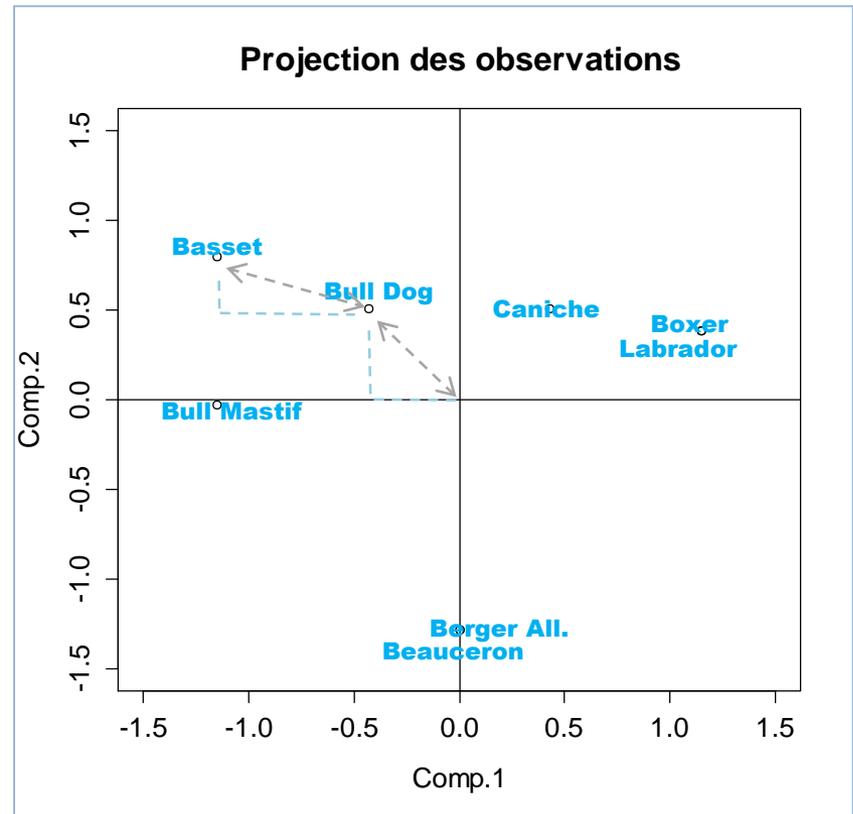
Ex. Facteur 1 est défini par l'opposition (Basset, Bull Mastif) vs. (Boxer, Labrador)

Projection des observations



Chien	Coord.	
	Axe.1	Axe.2
Beauceron	0.00	-1.28
Basset	-1.15	0.80
Berger All	0.00	-1.28
Boxer	1.15	0.39
Bull-Dog	-0.43	0.51
Bull-Mastif	-1.15	-0.03
Caniche	0.43	0.51
Labrador	1.15	0.39

Comme pour les modalités, on peut estimer les distances entre individus et les distances à l'origine. La précision de l'approximation dépend de la qualité de la représentation des modalités sur les facteurs considérés.



$$d^2(\text{Bull Dog}) = (-0.43)^2 + (0.51)^2 = 0.4433$$

$$d^2(\text{Bull Dog, Basset}) = [-0.43 - (-1.15)]^2 + (0.51 - 0.80)^2 = 0.6056$$



Représentation simultanée
Associations individus x modalités





A l'instar de l'AFC, il est possible d'obtenir les coordonnées des colonnes (modalités) à partir des lignes (individus) ; et inversement.

Coord. Individus à partir des coord. modalités

$$F_{ih} = \frac{1}{\sqrt{\lambda_h}} \sum_{k=1}^M \frac{x_{ik}}{p} G_{kh}$$

Ex. Coordonnée sur le 1^{er} facteur de « Basset » à partir de son profil ligne et des coordonnées factorielles des points modalités

Profil ligne	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Basset	0.333	0.000	0.000	0.333	0.000	0.000	0.333	0.000

Coord.Modalités	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Axe.1	-0.4559	1.3676	-0.4559	-1.0815	1.0815	0.0000	-1.3676	0.4559

$$F_{Basset,1} = \frac{1}{\sqrt{0.7080}} (0.333 \times (-0.4559) + 0.000 \times 1.3676 + \dots + 0.000 \times 0.4559) = -1.1508$$

Ex. Coordonnée sur le « Taille- » à partir de son profil colonne et des coordonnées factorielles des points individus

Coord. modalités à partir des coord. individus

$$G_{kh} = \frac{1}{\sqrt{\lambda_h}} \sum_{i=1}^n \frac{x_{ik}}{n_k} F_{ih}$$

Profil colonne	
Chien	Taille-
Beauceron	0.000
Basset	0.333
Berger All	0.000
Boxer	0.000
Bull-Dog	0.333
Bull-Mastif	0.000
Caniche	0.333
Labrador	0.000

Coord.Individus	
Chien	Axe.1
Beauceron	0.0000
Basset	-1.1508
Berger All	0.0000
Boxer	1.1508
Bull-Dog	-0.4284
Bull-Mastif	-1.1508
Caniche	0.4284
Labrador	1.1508

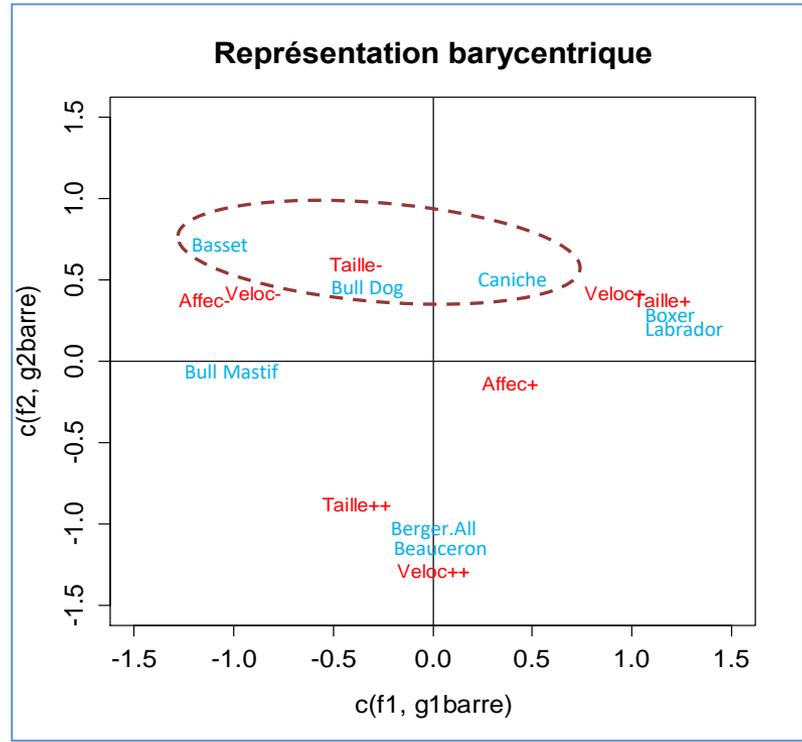
$$G_{taille-,1} = \frac{1}{\sqrt{0.7080}} \left(0.000 \times 0.000 + 0.333 \times (-1.1508) + \dots + 0.333 \times 0.4284 + 0.000 \times 1.1508 \right) = -0.4559$$



Représentation simultanée

Représentation barycentrique

Principe : On peut incorporer les modalités et les individus dans le même graphique. Le mieux dans ce cas est de placer chaque modalité au barycentre des individus qui possèdent la caractéristique.



On peut obtenir ces moyennes en les calculant à partir des coord. factorielles des individus

Ex. Moyennes pour « Taille » sur les 2 facteurs

	Données	
Taille	Moyenne de Axe.1	Moyenne de Axe.2
Taille-	-0.3836	0.6061
Taille+	1.1508	0.3851
Taille++	-0.3836	-0.8629
Moyenne générale	0.0000	0.0000

On peut les obtenir également via une transformation des coordonnées factorielles des modalités

Ex. Moyennes conditionnelles de toutes les modalités sur les 2 facteurs

Attribute = Value	Coord.Factorielles		Moyenne conditionnelles	
	coord_1	coord_2	Axe.1	Axe.2
Taille = Taille-	-0.4559	0.7881	-0.3836	0.6061
Taille = Taille+	1.3676	0.5008	1.1508	0.3851
Taille = Taille++	-0.4559	-1.1220	-0.3836	-0.8629
Velocite = Veloc-	-1.0815	0.5547	-0.9100	0.4266
Velocite = Veloc+	1.0815	0.5547	0.9100	0.4266
Velocite = Veloc++	0.0000	-1.6642	0.0000	-1.2799
Affection = Affect-	-1.3676	0.5008	-1.1508	0.3851
Affection = Affect+	0.4559	-0.1669	0.3836	-0.1284

$$\bar{G}_{kh} = G_{kh} \sqrt{\lambda_h}$$

Ex. « Basset », « Bull Dog » et « Caniche » sont les 3 chiens de « Taille- ».

Cf. relation de transition : coord. des modalités à partir des coord. des individus



Représentation simultanée

Formule de reconstitution des indicatrices associant les individus aux modalités

On peut estimer la « propension » (parce que ce n'est pas une « vraie » probabilité) de l'individu « i » à posséder le caractère « k » à partir des coordonnées factorielles (H facteurs).

Remarque : l'estimation est exacte si on prend tous les facteurs (H_{max}).

$$\hat{x}_{ik} = \frac{n_k}{n} \times \left(1 + \sum_{h=1}^H \frac{F_{ih} \times G_{kh}}{\sqrt{\lambda_h}} \right)$$

Individus

Val.Prop.	0.7080	0.5915
Chien	Axe.1	Axe.2
Beauceron	0.000	-1.280
Basset	-1.151	0.799
Berger All	0.000	-1.280
Boxer	1.151	0.385
Bull-Dog	-0.428	0.510
Bull-Mastif	-1.151	-0.029
Caniche	0.428	0.510
Labrador	1.151	0.385

Modalités

	n_k/n	axe.1	Axe.2
Taille-	0.375	-0.456	0.788
Taille+	0.250	1.368	0.501
Taille++	0.375	-0.456	-1.122
Veloc-	0.375	-1.081	0.555
Veloc+	0.375	1.081	0.555
Veloc++	0.250	0.000	-1.664
Affec-	0.250	-1.368	0.501
Affec+	0.750	0.456	-0.167

Tableau observé

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	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



Tableau reconstituée à l'aide des 2 premiers facteurs

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Beauceron	-0.12	0.04	1.08	0.03	0.03	0.94	0.04	0.96
Basset	0.92	-0.09	0.17	1.15	0.04	-0.18	0.85	0.15
Berger All	-0.12	0.04	1.08	0.03	0.03	0.94	0.04	0.96
Boxer	0.29	0.78	-0.07	-0.08	1.03	0.04	-0.15	1.15
Bull-Dog	0.66	0.16	0.18	0.72	0.31	-0.03	0.51	0.49
Bull-Mastif	0.80	-0.22	0.62	0.92	-0.19	0.27	0.71	0.29
Caniche	0.48	0.51	0.01	0.31	0.72	-0.03	0.16	0.84
Labrador	0.29	0.78	-0.07	-0.08	1.03	0.04	-0.15	1.15

$$\hat{x}_{beauceron,taille-} = 0.375 \times \left(1 + \frac{0.000 \times (-0.456)}{\sqrt{0.7080}} + \frac{(-1.280) \times 0.788}{\sqrt{0.5915}} \right) = -0.12$$

Une règle de décision possible : présence du caractère si val. estimée > (p/M = 0.375) (moyenne des val. du tableau observé)

→ 3 mauvaises associations ici (« Bull Mastif x Taille- », etc.)



Variables illustratives quantitatives et qualitatives
Renforcer l'interprétation des composantes



Ce sont des variables (qualitatives ou quantitatives) non exploitées pour la construction des composantes. Mais utilisées après coup pour mieux comprendre / commenter les résultats.

Ex. Construire l'ACM à partir des comportements d'achat des clients, et illustrer à l'aide de leurs caractéristiques signalétiques (âge, revenus, etc.).

Dans notre exemple, on cherche à illustrer les races canines avec une cote d'amour subjective attribuée par des experts (totalement inventée) et les fonctions qui leurs sont assignées.

Var. Illustrative
quantitative

Var. Illustrative
qualitative



ID	Chien	Cote	Fonction
1	Beauceron	2	utilite
2	Basset	4.5	chasse
3	Berger All	2.5	utilite
4	Boxer	3	compagnie
5	Bull-Dog	1.5	compagnie
6	Bull-Mastif	1	utilite
7	Caniche	4	compagnie
8	Labrador	3.5	chasse



Calculer les corrélations des variables supplémentaires avec les facteurs. c.-à-d. calculer le coefficient de corrélation entre les coordonnées des « n » individus sur les facteurs et les valeurs prises par la variable illustrative.

$$r_y(F_h) = \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(F_{ih} - \bar{F}_h)}{s_y \times s_{F_h}} = \frac{\frac{1}{n} \sum_{i=1}^n F_{ih} (y_i - \bar{y})}{s_y \times \sqrt{\lambda_h}}$$

Chien	Axe.1	Axe.2	Cote
Beauceron	0.000	-1.280	2
Basset	-1.151	0.799	4.5
Berger All	0.000	-1.280	2.5
Boxer	1.151	0.385	3
Bull-Dog	-0.428	0.510	1.5
Bull-Mastif	-1.151	-0.029	1
Caniche	0.428	0.510	4
Labrador	1.151	0.385	3.5

Corrélation	Axe.1	Axe.2
Cote	0.2881	0.4418

Tester la « significativité » du lien avec la statistique basée sur la transformation de Fisher

$$u_y = \sqrt{n-3} \times \left(\frac{1}{2} \ln \frac{1+r}{1-r} \right) \quad \Rightarrow \quad \text{Lien significatif à } (\sim) 5\% \text{ si } |u_y| \geq 2$$

U.Fisher	Axe.1	Axe.2
Cote	0.6630	1.0608

La « cote » ne semble corrélée avec aucun facteur, ce n'est pas étonnant : (1) elle a été créée de toute pièce ; (2) avec n = 8 observations, il est difficile de trouver quelque chose de significatif.



Variables illustratives qualitatives

Moyennes conditionnelles et coordonnées factorielles

Chien	Axe.1	Axe.2	Fonction
Beauceron	0.000	-1.280	utilite
Basset	-1.151	0.799	chasse
Berger All	0.000	-1.280	utilite
Boxer	1.151	0.385	compagnie
Bull-Dog	-0.428	0.510	compagnie
Bull-Mastif	-1.151	-0.029	utilite
Caniche	0.428	0.510	compagnie
Labrador	1.151	0.385	chasse

Solution 1 : calculer les moyennes conditionnelles et corriger avec la v.p. pour obtenir les coordonnées fact.

Fonction	Moy. conditionnelles	
	Moyenne de Axe.1	Moyenne de Axe.2
chasse	0.0000	0.5921
compagnie	0.3836	0.4682
utilite	-0.3836	-0.8629
Moy. Glob.	0.0000	0.0000

Moyennes cond. obtenues à partir du tableau des coord. fact. des individus

$$\bar{G}_{k^*h}$$

Axe	Axe.1	Axe.2
Val.Pr.	0.7080	0.5915

Coord. fact. des modalités supplémentaires après correction par les v.p.

$$G_{k^*h} = \frac{\bar{G}_{k^*h}}{\sqrt{\lambda_h}}$$

	Coordonnées factorielles	
	Axe.1	Axe.2
chasse	0.0000	0.7699
compagnie	0.4559	0.6087
utilite	-0.4559	-1.1220

Solution 2 : s'appuyer sur la relation de transition pour obtenir directement les coordonnées fact.

Codage disjonctif complet de la variable supplémentaire (on dispose des x_{ik^*})

Chien	Fonction				
	Axe.1	Axe.2	chasse	compagnie	utilite
Beauceron	0.000	-1.280	0	0	1
Basset	-1.151	0.799	1	0	0
Berger All	0.000	-1.280	0	0	1
Boxer	1.151	0.385	0	1	0
Bull-Dog	-0.428	0.510	0	1	0
Bull-Mastif	-1.151	-0.029	0	0	1
Caniche	0.428	0.510	0	1	0
Labrador	1.151	0.385	1	0	0

Application de la formule de transition :

« [somme(profil colonne x coord. individus)] / racine (valeur.propre) »

$$G_{k^*h} = \frac{1}{\sqrt{\lambda_h}} \sum_{i=1}^n \frac{x_{ik^*}}{n_{k^*}} F_{ih}$$

n_k*	2	3	3
------	---	---	---

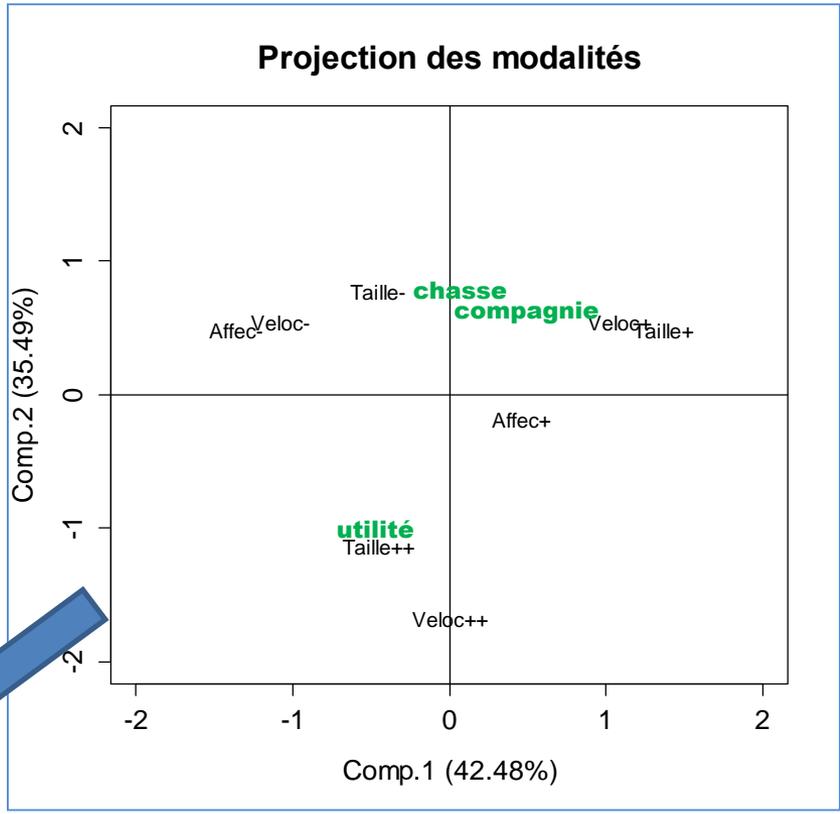
	chasse	compagnie	utilite
Axe.1	0.0000	0.4559	-0.4559
Axe.2	0.7699	0.6087	-1.1220



A partir de leurs coordonnées, on peut placer les nouvelles modalités dans le plan factoriel

Fonction	chasse	compagnie	utilite
Axe.1	0.0000	0.4559	-0.4559
Axe.2	0.7699	0.6087	-1.1220

Remarque : comme pour les modalités actives, il est possible de calculer les valeurs test et les rapports de corrélation.



On note une proximité entre certaines modalités de TAILLE et de FONCTION... confirmée par le croisement des 2 variables (« Taille++ » ⇔ « Utilité »)

Nombre de Taille	Fonction			
Taille	chasse	compagnie	utilite	Total général
Taille-	1	2		3
Taille+	1	1		2
Taille++			3	3
Total général	2	3	3	8



Individus illustratifs (supplémentaires)

Positionner de nouveaux individus



Plusieurs raisons possibles :

1. Des individus collectés après coup que l'on aimerait situer par rapport à ceux de l'échantillon d'apprentissage (les individus actifs).
2. Des individus appartenant à une population différente (ou spécifique) que l'on souhaite positionner.
3. Des observations s'avérant atypiques ou trop influentes dans l'ACM que l'on a préféré écarter. On veut maintenant pouvoir juger de leur positionnement par rapport aux individus actifs.

ID	Chien	Taille	Velocite	Affection
Supplém.	Levrier	Taille++	Veloc++	Affec-

Plutôt cas n°1 ici, on essaie de situer une race supplémentaire par rapport aux autres (les individus actifs)



Individus illustratifs (supplémentaires)

Calcul des coordonnées factorielles

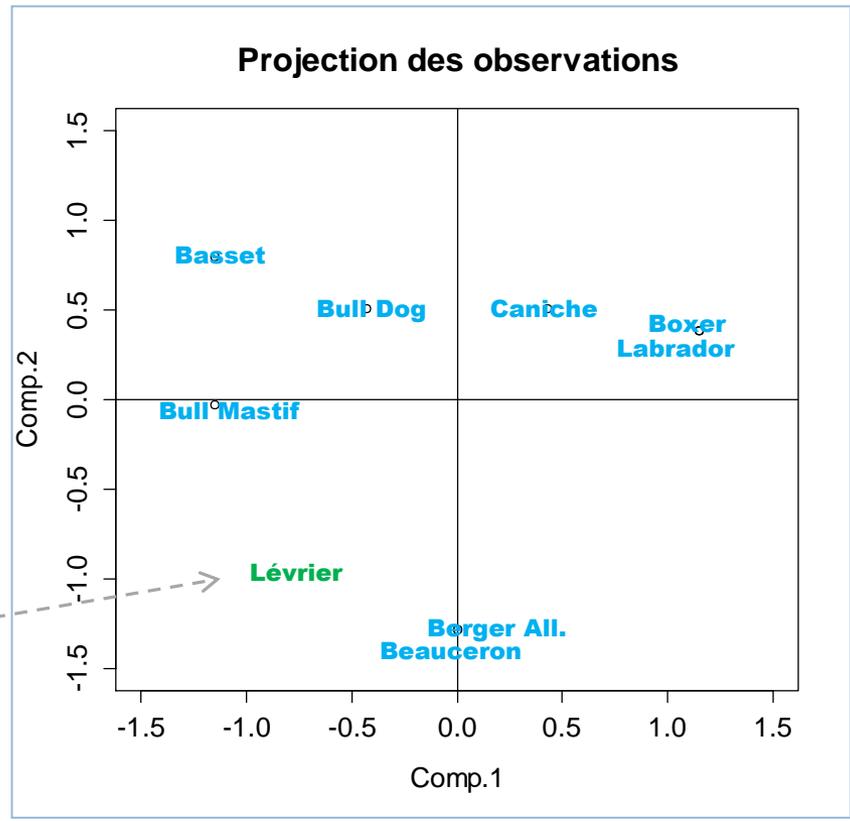
On s'appuie sur la relation de transition pour obtenir ses coordonnées factorielles à partir de son profil.

$$F_{ih} = \frac{1}{\sqrt{\lambda_h}} \sum_{k=1}^M \frac{x_{ik}}{p} G_{kh}$$

ID	Chien	Taille	Velocite	Affection
Supplém.	Levrier	Taille++	Veloc++	Affec-

Val.Propres	0.7080	0.5915		
	Axe.1	Axe.2	Desc.Levrier	Profil.Levrier
Taille-	-0.456	0.788	0	0.000
Taille+	1.368	0.501	0	0.000
Taille++	-0.456	-1.122	1	0.333
Veloc-	-1.081	0.555	0	0.000
Veloc+	1.081	0.555	0	0.000
Veloc++	0.000	-1.664	1	0.333
Affec-	-1.368	0.501	1	0.333
Affec+	0.456	-0.167	0	0.000
		Somme	3	

	Axe.1	Axe.2
Levrier	-0.7224	-0.9905



On peut observer son voisinage dans le 1^{er} plan factoriel

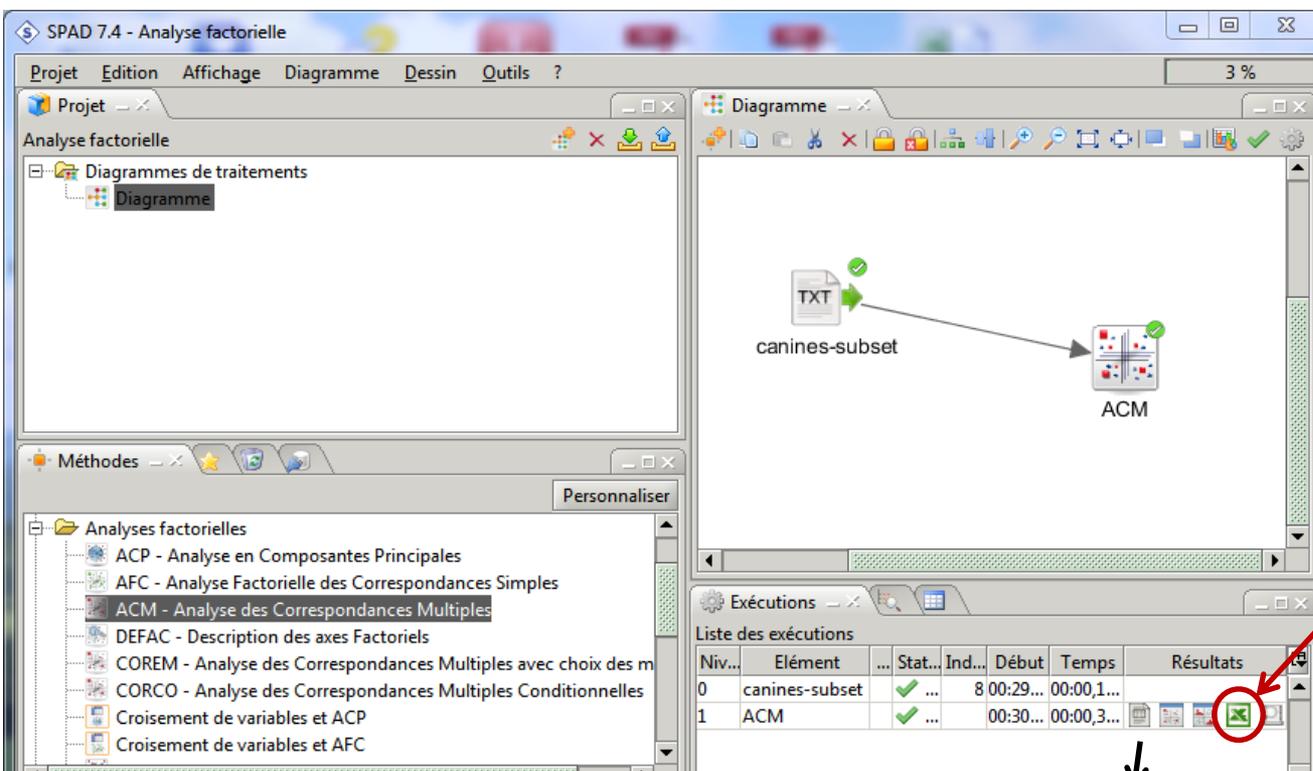
$$\frac{1}{\sqrt{0.7080}} (0.000 \times (-0.456) + 0.000 \times 1.368 + 0.333 \times (-0.456) + \dots) = -0.7224$$



Logiciels

Les signes des facteurs peuvent être différents d'un logiciel à l'autre. Ce n'est pas un problème. Ce sont les positions relatives des modalités et des individus qui importent.





SPAD

Ses sorties font référence dans les ouvrages (cf. bibliographie)

Les sorties peuvent être redirigées vers le tableur Excel. Option décisive si l'on souhaite réaliser des calculs additionnels.

Exécutions

Liste des exécutions

Niv...	Élément	Stat...	Ind...	Début	Temps	Résultats
0	canines-subset	✓	...	8 00:29...	00:00,1...	
1	ACM	✓	...	00:30...	00:00,3...	

Editeur de résultats - [unit.lst:2]

Fichier Edition Affichage Fenêtre ?

DORDONNEES, CONTRIBUTIONS ET COSINUS CARRÉS DES MODALITES ACTIVES
LIGES 1 A 5

MODALITES		COORDONNEES					CONTRIBUTIONS					COSINUS CARRÉS					
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
2 . Taille																	
m1 - Taille-	12.50	1.67	-0.46	0.79	0.90	-0.12	0.11	3.7	13.1	38.7	2.6	4.4	0.12	0.37	0.49	0.01	0.01
m2 - Taille+	8.33	3.00	1.37	0.50	-0.84	0.36	0.22	22.0	3.5	22.2	15.5	11.7	0.62	0.08	0.23	0.04	0.02
m3 - Taille++	12.50	1.67	-0.46	-1.12	-0.34	-0.12	-0.26	3.7	26.6	5.6	2.6	24.0	0.12	0.76	0.07	0.01	0.04
CONTRIBUTION CUMULEE = 29.4 43.3 66.5 20.6 40.2																	
3 . Velocite																	
m1 - Veloc-	12.50	1.67	-1.08	0.55	-0.10	0.40	-0.13	20.6	6.5	0.5	29.4	5.5	0.70	0.18	0.01	0.10	0.01
m2 - Veloc+	12.50	1.67	1.08	0.55	-0.10	-0.40	-0.13	20.6	6.5	0.5	29.4	5.5	0.70	0.18	0.01	0.10	0.01
m3 - Veloc++	8.33	3.00	0.00	-1.66	0.30	0.00	0.38	0.0	39.0	2.9	0.0	33.1	0.00	0.92	0.03	0.00	0.05
CONTRIBUTION CUMULEE = 41.3 52.0 3.8 58.7 44.2																	
4 . Affection																	
m1 - Affec-	8.33	3.00	-1.37	0.50	-0.84	-0.36	0.22	22.0	3.5	22.2	15.5	11.7	0.62	0.08	0.23	0.04	0.02
m2 - Affec+	25.00	0.33	0.46	-0.17	0.28	0.12	-0.07	7.3	1.2	7.4	5.2	3.9	0.62	0.08	0.23	0.04	0.02
CONTRIBUTION CUMULEE = 29.4 4.7 29.7 20.6 15.6																	

```

proc corresp mca data = mesdata.canines dimens = 2;
tables taille velocite affection;
run;

```

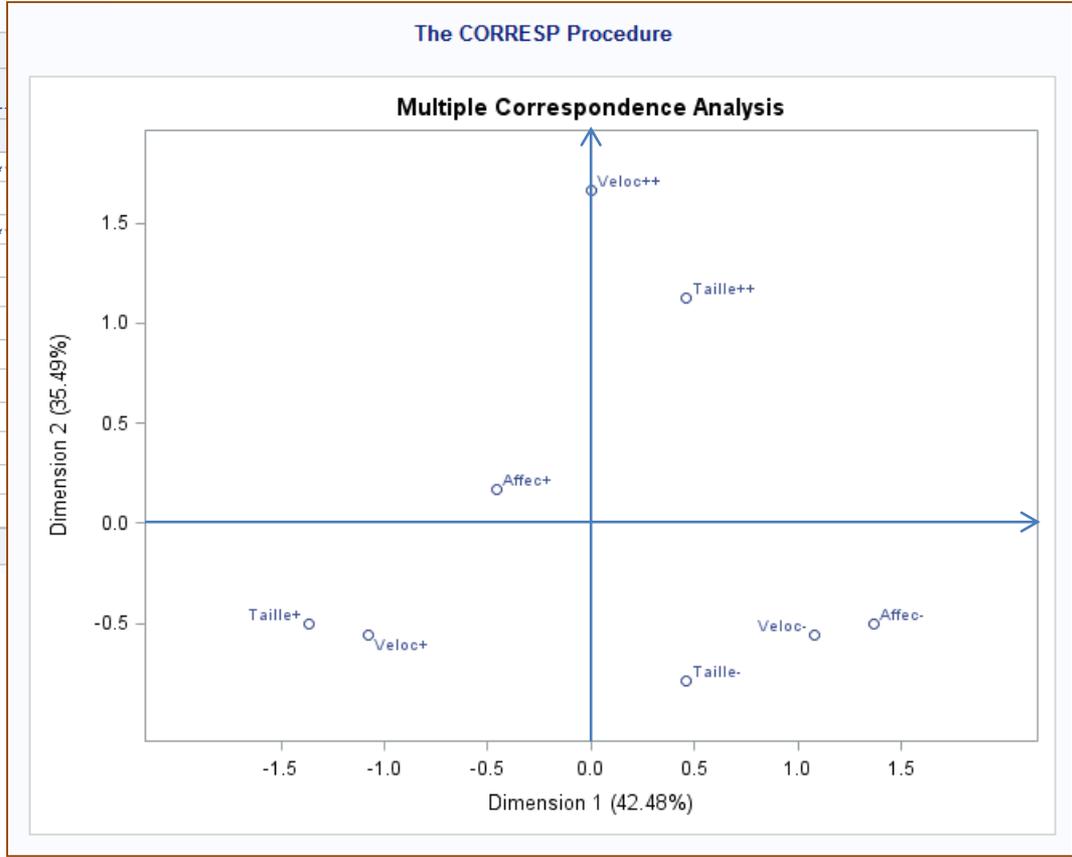
$$\sqrt{\lambda_h} \quad \lambda_h \quad n \times (M - p) \times \lambda_h$$

The CORRESP Procedure

Décomposition de l'inertie et du Khi-2

Valeur singulière	Inertie principale	Khi-2	Pourcentage	Pourcent. cumulé				
					8	16	24	32
0.84145	0.70803	28.3213	42.48	42.48	*****			
0.76908	0.59149	23.6596	35.49	77.97	*****			
0.51185	0.26199	10.4797	15.72	93.69	*****			
0.26410	0.06975	2.7899	4.18	97.88	***			
0.18817	0.03541	1.4163	2.12	100.00	*			
Total	1.66667	66.6667	100.00					

Degrés de liberté = 49



L’option **supplementary** permet de gérer les éléments supplémentaires (cf. la doc SAS). Attention cependant avec la clause **tables**, on ne peut pas tout faire.



Une multitude de packages de qualité : ade4, ca, FactoMineR, etc.

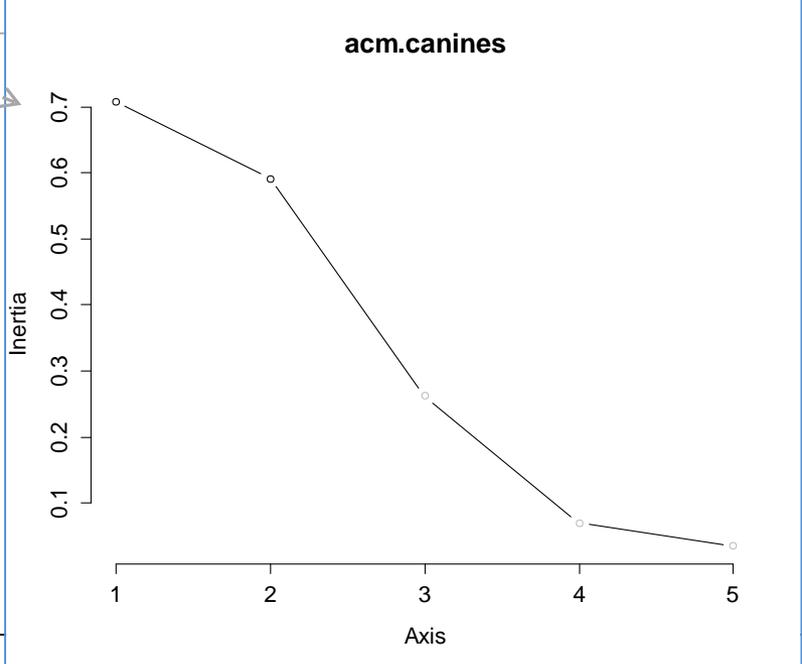
```

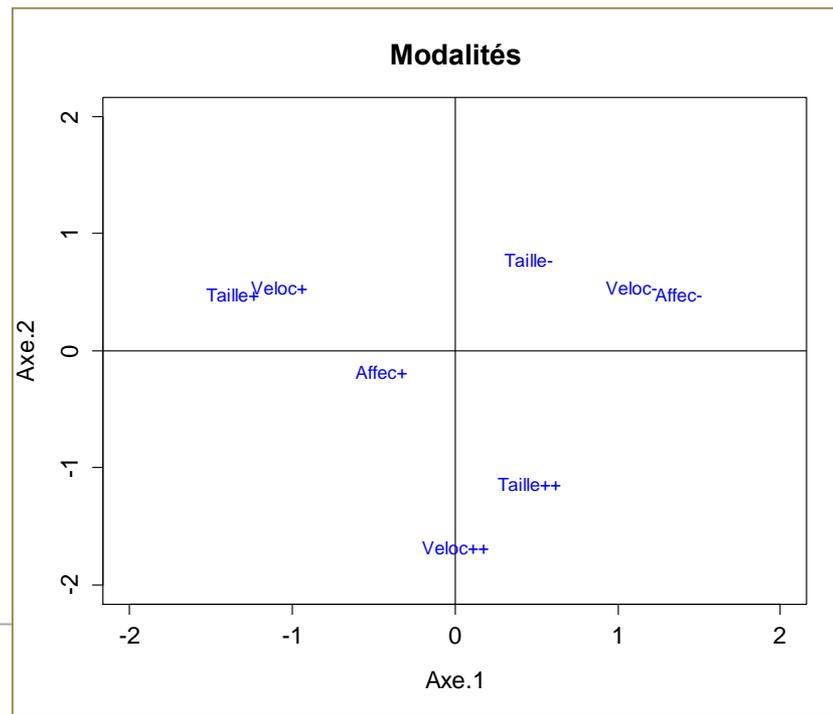
#chargement des données - y compris les 2 colonnes supplémentaires
canines <- read.table(file="canines-subset.txt",header=T,sep="\t",row.names=1,dec=".")
#chargement de la librairie
library(ade4)
#codage disjonctif des variables actives
binaires.actives <- acm.disjonctif(subset(canines,select=1:3))
colnames(binaires.actives) <- unlist(sapply(canines[1:3],function(x){levels(x)}))
print(binaires.actives)
#acm sur les indicatrices
acm.canines <- dudi.coa(binaires.actives,scannf=F,nf=2)
#éboullis des valeurs propres
screplot(acm.canines,type="lines")

```

```
> print(binaires.actives)
```

	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	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





```
#graphique des modalités
coord <- acm.canines$co
plot(coord[,1],coord[,2],xlim=c(-2,2),ylim=c(-2,2),type="n",main="Modalités",xlab="Axe.1",ylab="Axe.2")
abline(h=0,v=0)
text(coord[,1],coord[,2],labels=colnames(binaires.actives),cex=0.75,col="blue")
#evaluation des modalités => CTR et COS2
eval.acm <- inertia.dudi(acm.canines,col.inertia=T)
#contributions
contrib <- eval.acm$col.abs/100
rownames(contrib) <- colnames(binaires.actives)
print(contrib)
#cos2 (les cos2 sont signés)
cos2 <- eval.acm$col.rel/100
rownames(cos2) <- colnames(binaires.actives)
print(cos2[,1:2])
```

```
> print(contrib)
      Comp1 Comp2
Taille-  3.67 13.13
Taille+ 22.01  3.53
Taille++ 3.67 26.60
Veloc-  20.65  6.50
Veloc+  20.65  6.50
Veloc++  0.00 39.02
Affec-  22.01  3.53
Affec+   7.34  1.18
```

```
> print(cos2[,1:2])
      Comp1 Comp2
Taille- 12.47 37.27
Taille+ -62.35  8.36
Taille++ 12.47 -75.53
Veloc-  70.17 18.46
Veloc+ -70.17 18.46
Veloc++  0.00 -92.32
Affec-  62.35  8.36
Affec+ -62.35 -8.36
```



TANAGRA 1.4.48 - [Multiple Correspondence Analysis 1]

File Diagram Component Window Help

Analysis

Dataset (tanB099.txt)

- Define status 1
 - Multiple Correspondence Analysis 1

Report Chart

Axis	Eigen value	% explained	Histogram	% cumulated
1	0.708031	42.48%		42.48%
2	0.591489	35.49%		77.97%
3	0.261992	15.72%		93.69%
4	0.069747	4.18%		97.88%
5	0.035408	2.12%		100.00%

Factors characterization (active variables)

Coordinates and test-values

Values	Overall			Coordinate		Test-Value	
	Attribute = Value	Mass	Sq. Dist	Inertia	coord_1	coord_2	v-test_1
Velocite = Veloc++	0.0833	3.0000	0.2500	0.00000	-1.66422	0.000	-2.542
Velocite = Veloc-	0.1250	1.6667	0.2083	-1.08146	0.55474	-2.216	1.137
Velocite = Veloc+	0.1250	1.6667	0.2083	1.08146	0.55474	2.216	1.137
Taille = Taille++	0.1250	1.6667	0.2083	-0.451			
Taille = Taille-	0.1250	1.6667	0.2083	-0.451			
Taille = Taille+	0.0833	3.0000	0.2500	1.361			
Affection = Affec+	0.2500	0.3333	0.0833	0.451			
Affection = Affec-	0.0833	3.0000	0.2500	-1.361			

Cos² and contributions

Values	Cos ²		CTR (%)	
	Attribute = Value	cos _{2_1}	cos _{2_2}	ctr ₁
Velocite = Veloc++	0.0000	0.9232	0.000	39.021
Velocite = Veloc-	0.7017	0.1846	20.648	6.503
Velocite = Veloc+	0.7017	0.1846	20.648	6.503
-	-	Tot. ctr.	41.296	52.027
Taille = Taille++	0.1247	0.7553	3.669	26.603
Taille = Taille-	0.1247	0.3727	3.669	13.127
Taille = Taille+	0.6235	0.0836	22.014	3.533
-	-	Tot. ctr.	29.352	43.262
Affection = Affec+	0.6235	0.0836	7.338	1.178
Affection = Affec-	0.6235	0.0836	22.014	3.533
-	-	Tot. ctr.	29.352	4.710

Report Chart

Axis_2

Show supplementary variables

Contribution filtering :

Quality (cos²) filtering :

Axis_1

Coord. + Valeurs test

CTR + COS²

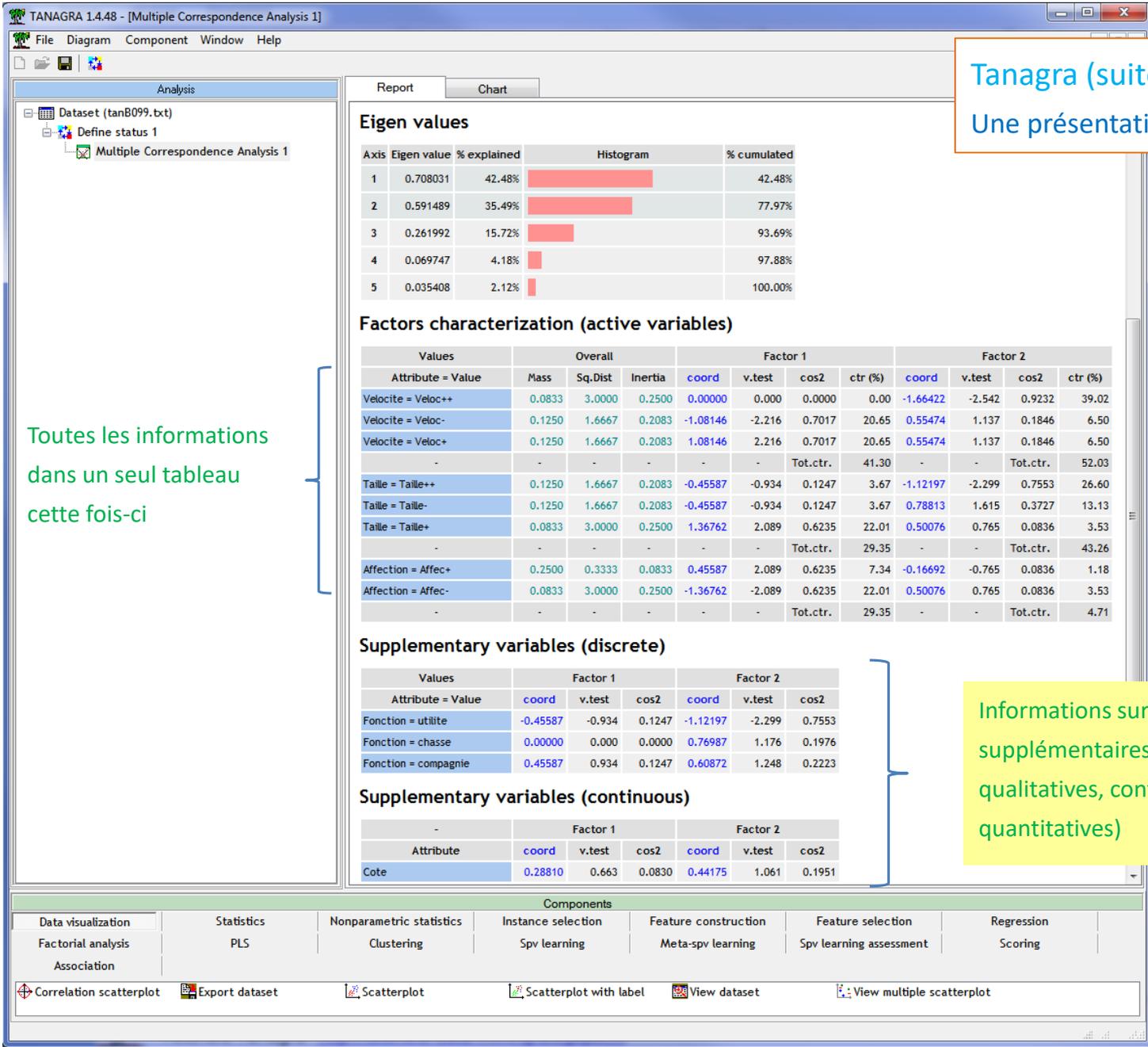
Tanagra

Axé sur la simplicité d'utilisation

Positionnement des modalités supplémentaires

Possibilité de filtrage selon la CTR et le COS²





Tanagra (suite)
Une présentation alternative

Toutes les informations dans un seul tableau cette fois-ci

Informations sur les variables supplémentaires (discrete = qualitatives, continuous = quantitatives)

Plus loin avec l'ACM (1) :

Analyse parallèle

Technique de ré-échantillonnage pour la détection des facteurs pertinents

Attention, ce type d'analyse est purement mécanique. Il faut valider les facteurs par l'interprétation.



Déterminer la distribution des λ_h sous H_0 (absence de lien entre les variables)

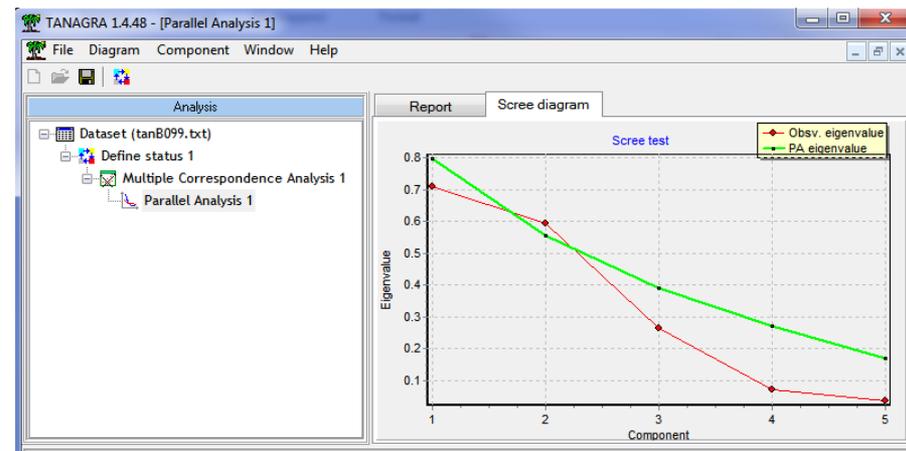
Démarche :

1. Mélanger aléatoirement les valeurs à l'intérieur des colonnes, en traitant les colonnes de manière indépendante \rightarrow le lien entre les variables est complètement cassé (on est sous H_0)
2. Réaliser l'ACM sur cette nouvelle version des données, collecter les v.p.
3. Répéter T fois les opérations (1) et (2)
4. On obtient pour chaque h une collection de v.p., on en déduit la moyenne μ_h qui sert de seuil critique
5. On décide que la composante h est pertinente si $\lambda_h > \mu_h$

Variante : Plutôt que la moyenne, on peut aussi prendre le quantile d'ordre 0.95 pour un test unilatéral à 5%

Parallel Analysis

Component	Eigenvalue	(0.95) Critical value
1	0.708031	0.796566
2	0.591489	0.555556
3	0.261992	0.388989
4	0.069747	0.270939
5	0.035408	0.168756



Données « Canines », seuil critique :
quantile d'ordre 0.95 des v.p. sous H_0

Représentation graphique des v.p. et des seuils pour chaque « h » sous Tanagra \rightarrow on sélectionnera 2 facteurs pour notre ACM



Plus loin avec l'ACM (2) :

Analyse non linéaire

Découpage en classes des variables quantitatives

Parfois, les relations entre les variables ne sont pas linéaires. Il peut être avantageux de les découper en classes (discrétisation) pour capter ce type d'information.



Description de caractéristiques de vins provenant de 3 régions différentes : 'These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines' (<http://archive.ics.uci.edu/ml/datasets/Wine>).

Variable illustrative

Variables actives (13)

Type	Alcohol	Malic_Acid	Ash	Ash_Alcalinity	Magnesium	Total_Phenols	Flavanoids	flavanoid_Phe	roanthocyanin	Color_Intensity	Hue	OD280/OD315	Proline
A	14.23	1.71	2.43	15.60	127.00	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065.00
A	13.20	1.78	2.14	11.20	100.00	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050.00
A	13.16	2.36	2.67	18.60	101.00	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185.00
A	14.37	1.95	2.50	16.80	113.00	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480.00
A	13.24	2.59	2.87	21.00	118.00	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735.00

n = 178 individus



Toutes les variables actives sont quantitatives, l'analyse en composantes principales (ACP) semble s'imposer.



Significance of Principal Components

Global critical values	
Kaiser-Guttman	1
Karlis-Saporta-Spinaki	1.52076

Eigenvalue table - Test for significance

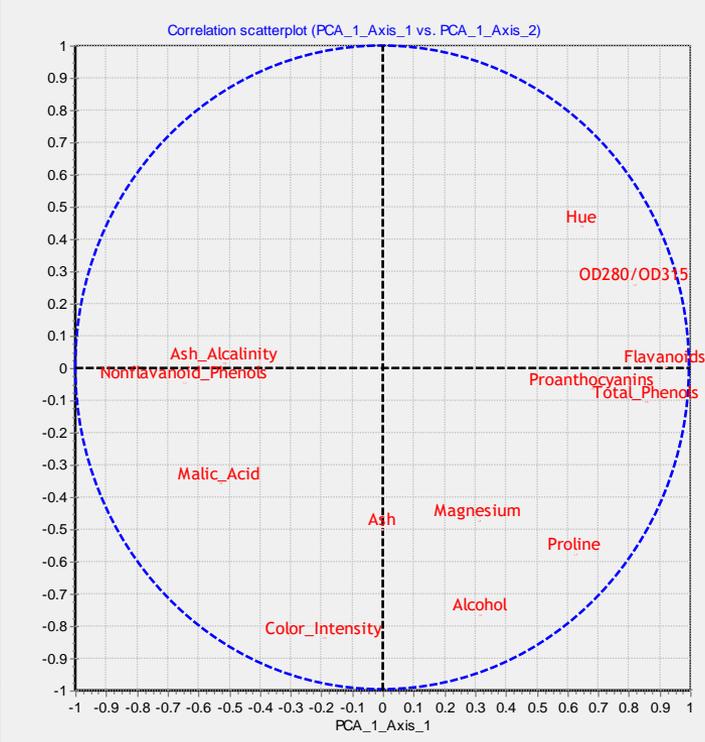
Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	4.705776	3.180134
2	2.497031	2.180134
3	1.446062	1.680134
4	0.91908	1.3468
5	0.853196	1.0968
6	0.641652	0.8968
7	0.551037	0.730134
8	0.348562	0.587277
9	0.288862	0.462277
10	0.250837	0.351166
11	0.225786	0.251166
12	0.168748	0.160256
13	0.103371	0.076923

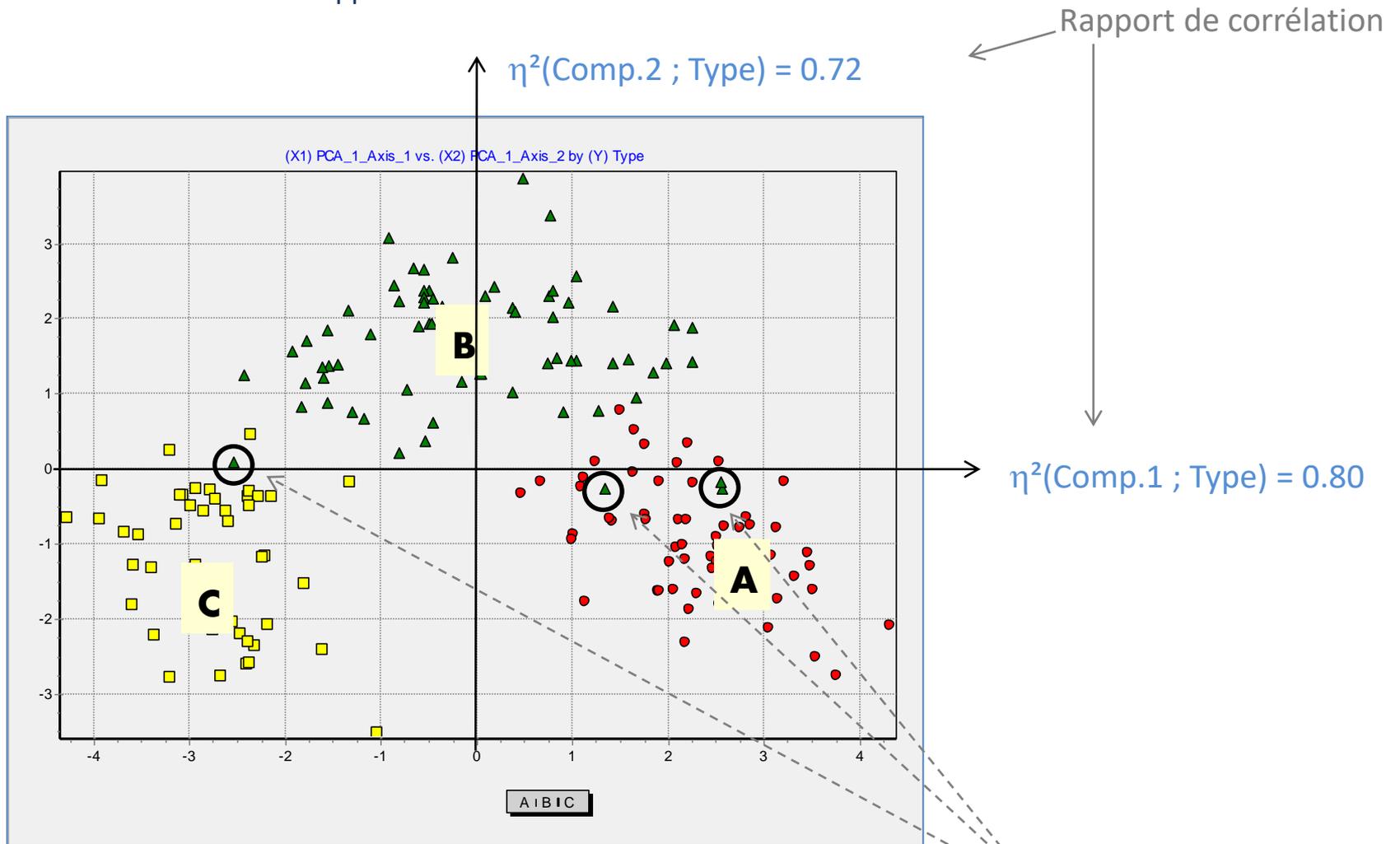
Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)
Flavanoids	0.91746	84 % (84 %)	0.00529	0 % (84 %)
Total_Phenols	0.85612	73 % (73 %)	-0.10281	1 % (74 %)
OD280/OD315	0.81601	67 % (67 %)	0.25991	7 % (73 %)
Proanthocyanins	0.67993	46 % (46 %)	-0.06211	0 % (47 %)
Nonflavanoid_Phenols	-0.64762	42 % (42 %)	-0.04547	0 % (42 %)
Hue	0.6436	41 % (41 %)	0.44132	19 % (61 %)
Proline	0.62206	39 % (39 %)	-0.57659	33 % (72 %)
Malic_Acid	-0.5319	28 % (28 %)	-0.35545	13 % (41 %)
Ash_Alcalinity	-0.51916	27 % (27 %)	0.01673	0 % (27 %)
Color_Intensity	-0.19225	4 % (4 %)	-0.8375	70 % (74 %)
Alcohol	0.31308	10 % (10 %)	-0.76426	58 % (68 %)
Ash	-0.00446	0 % (0 %)	-0.49945	25 % (25 %)
Magnesium	0.30803	9 % (9 %)	-0.47345	22 % (32 %)
Var. Expl.	4.70578	36 % (36 %)	2.49703	19 % (55 %)

2 composantes « pertinentes »
(apparemment)

Cercle des corrélations





On distingue bien les groupes, mais la séparation n'est pas nette. Voir aussi les rapports de corrélation sur chaque composante.



2 questions cruciales :

- (1) Choisir le nombre de classes
- (2) Définir les bornes (seuils) de découpage

Ex. Découpage en 4 classes de X (pourquoi 4 ?), on doit définir les coordonnées des 3 bornes (comment ?)



Pour les variables de la base WINE,
nous avons spécifié pour chaque
variable...

Source	New att	Intervals	Cut points
Alcohol	d_Alcohol	3	(12.1850 ; 12.7800)
Malic_Acid	d_Malic_Acid	3	(1.4200 ; 2.2350)
Ash	d_Ash	2	-2.03
Ash_Alcalinity	d_Ash_Alcalinity	2	-17.9
Magnesium	d_Magnesium	2	-88.5
Total_Phenols	d_Total_Phenols	3	(1.8400 ; 2.3350)
Flavanoids	d_Flavanoids	4	(0.9750 ; 1.5750 ; 2.3100)
Nonflavanoid_Phenols	d_Nonflavanoid_Phenols	2	-0.395
Proanthocyanins	d_Proanthocyanins	2	-1.27
Color_Intensity	d_Color_Intensity	3	(3.4600 ; 7.5500)
Hue	d_Hue	4	(0.7850 ; 0.9750 ; 1.2950)
OD280/OD315	d_OD280/OD315	3	(2.1150 ; 2.4750)
Proline	d_Proline	4	(468.0000 ; 755.0000 ; 987.5000)

Nombre de classes

Seuils



L'information est plus dispersée. Normal, on a démultiplié les colonnes après codage disjonctif complet (M = 37).

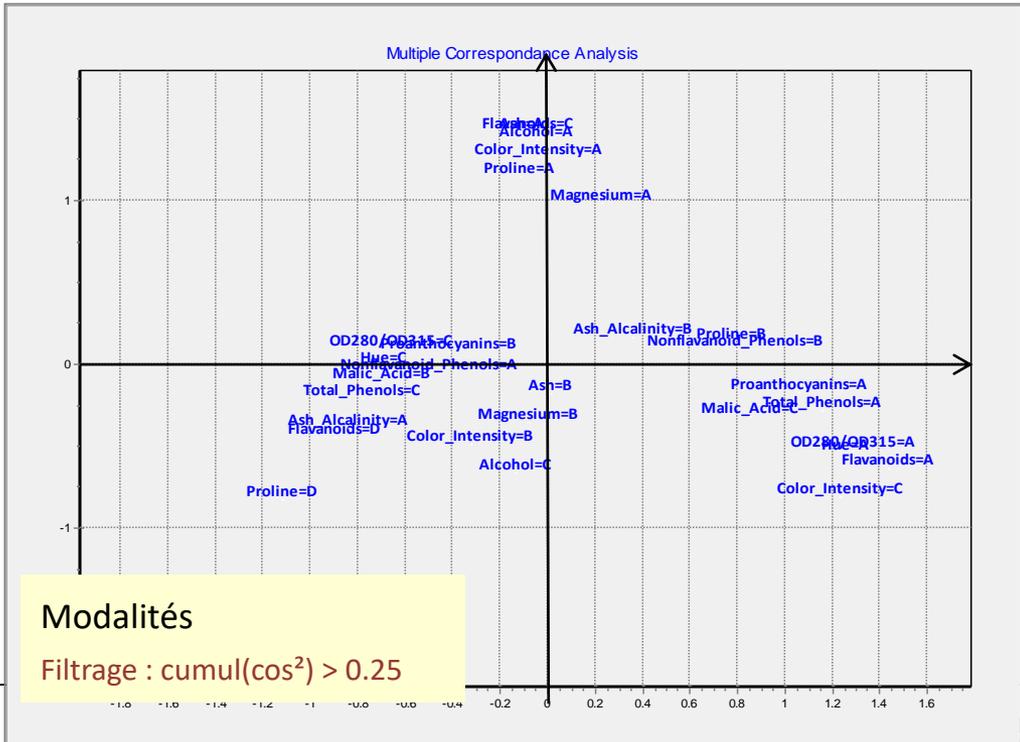
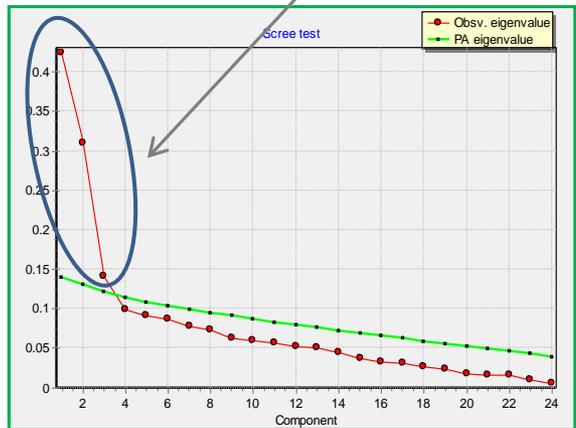
3 facteurs semblent intéressants dicit l'analyse parallèle.

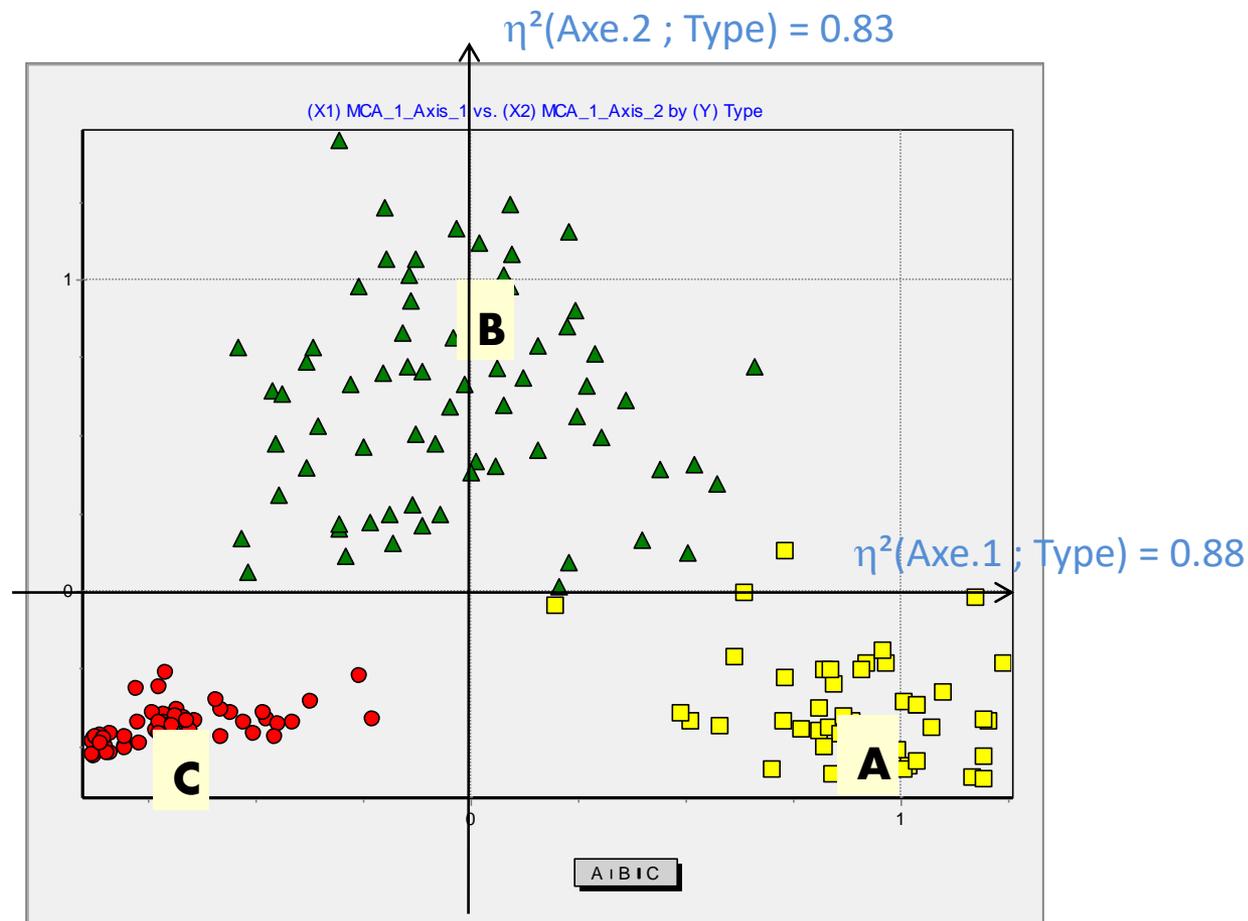
Problem statement

# of instances	178
# of variables	13
# of variable values	37
Max # of factors	24
# of factors extracted	5
Total inertia	1.846154

Eigen values

Axis	Eigen value	% explained	% cumulated
1	0.424578	23.00%	23.00%
2	0.310058	16.79%	39.79%
3	0.141764	7.68%	47.47%
4	0.099221	5.37%	52.85%
5	0.091097	4.93%	57.78%
6	0.086896	4.71%	62.49%
7	0.078214	4.24%	66.72%
8	0.073202	3.97%	70.69%
9	0.063059	3.42%	74.10%
10	0.059359	3.22%	77.32%
11	0.056832	3.08%	80.40%
12	0.051308	2.78%	83.18%
13	0.050506	2.74%	85.91%
14	0.044669	2.42%	88.33%
15	0.037383	2.02%	90.36%
16	0.032727	1.77%	92.13%
17	0.031353	1.70%	93.83%
18	0.025892	1.40%	95.23%
19	0.023828	1.29%	96.52%
20	0.016753	0.91%	97.43%
21	0.016563	0.90%	98.33%
22	0.015268	0.83%	99.15%
23	0.010328	0.56%	99.71%
24	0.005295	0.29%	100.00%

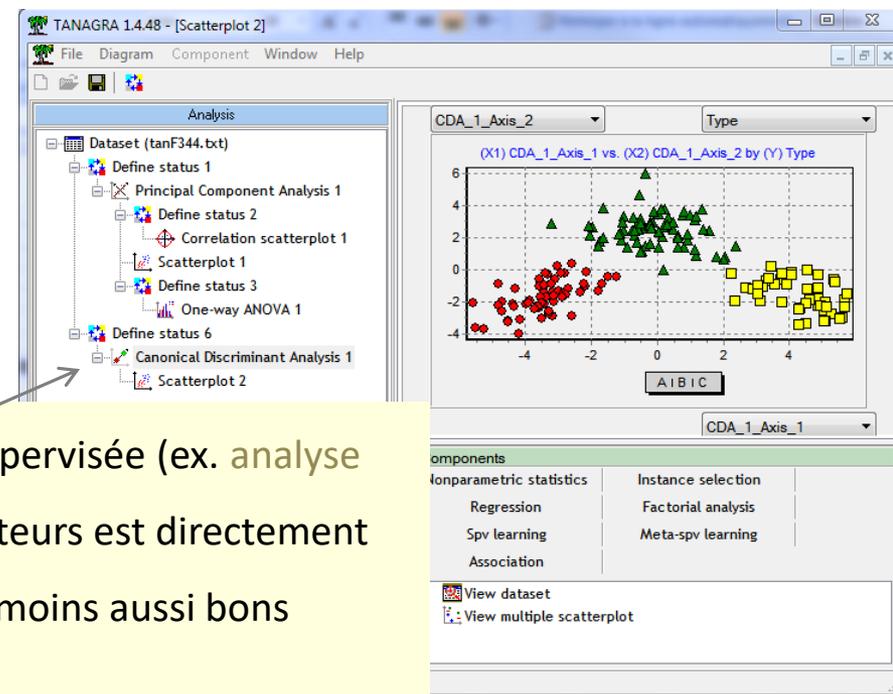




Il n'y a plus de chevauchements. Les rapports de corrélation sont améliorés simultanément sur les 2 premiers facteurs.



Mais... le découpage en classes n'est pas la panacée



1. On aurait pu partir directement sur une analyse supervisée (ex. **analyse factorielle discriminante**) – La construction des facteurs est directement guidée par la variable 'TYPE'. Les résultats sont au moins aussi bons ($\eta^2_1=0.90$; $\eta^2_2=0.81$)
2. La discrétisation engendre aussi une perte d'information. Nous perdons la variabilité à l'intérieur des classes. Ce n'est anodin.
3. Le processus de découpage reste un problème difficile : combien d'intervalles ? comment choisir les seuils ? (hum ! j'ai un peu triché dans mon exemple, j'ai découpé les variables en tenant compte de la distribution de 'type')
4. La démultiplication des colonnes après codage disjonctif complet disperse l'information. Il faut être très vigilant lors de l'interprétation des résultats.



Bibliographie



Les ouvrages incontournables sur l'analyse de données

Escofier B., Pagès J., « Analyses factorielles simples et multiples », Dunod, 2008.

Lebart L., Morineau A., Piron M., « Statistique exploratoire multidimensionnelle », Dunod, 3^{ème} édition, 2000.

Saporta G., « Probabilités, Analyse des Données et Statistique », Technip, 2006.

Tenenhaus M., « Statistique : Méthodes pour décrire, expliquer et prévoir », Dunod, 2006.

Tutoriels et supports de cours (innombrables sur le web) avec, entres autres,

Tutoriel Tanagra, <http://tutoriels-data-mining.blogspot.fr/> ; voir la section « Analyse Factorielle ».

Les plus complets (Tanagra, code source R, SAS, etc.), certains traitant le fichier « Races Canines » (version complète), sont :

- « [AFCM – Races Canines](#) » (Mars 2008)
- « [Analyse des correspondances multiples avec R](#) » (Mai 2009)
- « [Analyse des correspondances multiples – Outils](#) » (Déc. 2012)

