Analyse des Correspondances Discriminante

Analyse Factorielle Discriminante

(ou Analyse Discriminante Descriptive)

pour variables descriptives qualitatives

Ricco RAKOTOMALALA

Université Lumière Lyon 2

PLAN

Position du problème

Construire un nouveau système de représentation (facteurs)

qui permet de mettre en évidence les groupes

Objectif de l'analyse discriminante

Une population divisée en K groupes (classes), décrite par J variables qualitatives.

Wine dataset (H. Abdi, 2007)

Region	Woody	Fruity	Sweet	Alcohol	Hedonic
Loire	A	C	B	A	A
Loire	B	C	C	B	C
Loire	A	B	B	A	B
Loire	A	C	C	B	D
Rhone	A	B	A	C	C
Rhone	B	A	A	C	B
Rhone	C	B	B	B	A
Rhone	B	C	C	C	D
Beaujolais	C	A	C	A	A
Beaujolais	B	A	C	A	B
Beaujolais	C	B	B	B	D
Beaujolais	C	A	A	A	C

Groupe d'appartenance

Description

Objectifs :

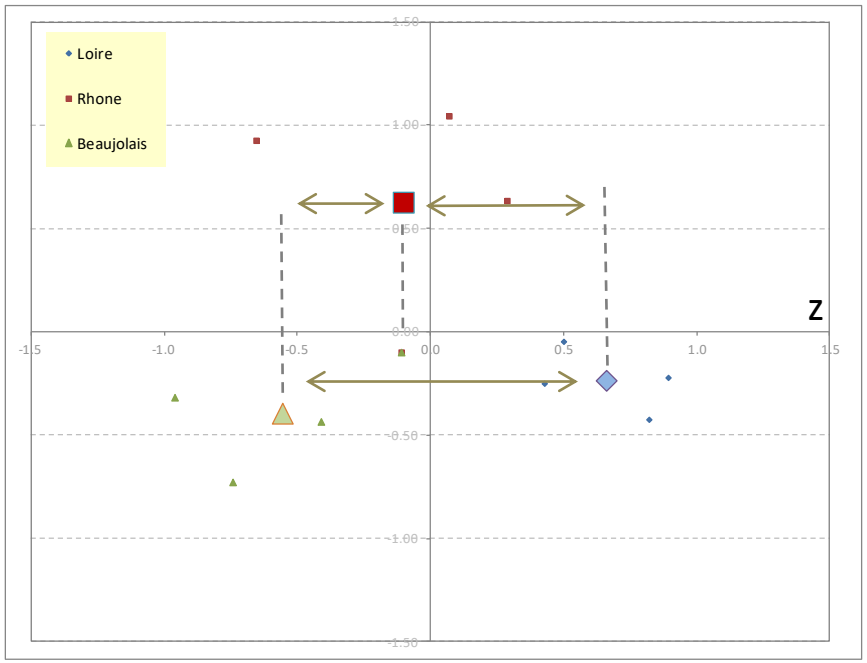
(1) Identifier les proximités entre les groupes

Ex. Les vins du Rhône ressemblent plus aux Beaujolais ou aux vins de Loire ?

(2) Caractériser l'appartenance aux groupes

Ex. Qu'est-ce qui distingue les vins du Rhône du Beaujolais ?

Principe de l'analyse discriminante



$$\sum_i (z_i - \bar{z})^2 = \sum_k n_k (\bar{z}_k - \bar{z})^2 + \sum_k \sum_i (z_{ik} - \bar{z}_k)^2$$

SC Totaux = SC Expliqués + SC Résiduels

1 Principe : Construire des « facteurs », qui sont des combinaisons linéaires des indicatrices des variables descriptives, permettant de discerner au mieux les (centres de) groupes.

1' Ou (de manière équivalente) : construire des « facteurs » sur lesquels les centres de groupes sont le plus dispersés possibles autour de la moyenne globale (ne tenant pas compte de l'appartenance aux groupes).

2 On souhaite maximiser SCE ⇔ Maximiser le rapport de corrélation

$$\eta^2 = \frac{SCE}{SCT} \quad \text{avec} \quad \boxed{0 \leq \eta^2 \leq 1}$$

3 Si la discrimination n'est pas parfaite sur le 1er facteur, on traite la partie résiduelle (non expliquée) avec le 2nd c.-à-d. maximiser les écarts (entre barycentres) non pris en compte sur le facteur précédent, etc.

Distance dans l'espace de description originel

Distance entre centres de groupes

Distance avec le barycentre global (sans tenir compte des groupes)

Comptabiliser la distance entres groupes que l'on pourra appréhender sur les axes factoriels

Tableau des indicatrices – Moyenne globale et moyennes conditionnelles – Distance du KHI-2

Region	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D	Total
Loire	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	5
Loire	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	5
Loire	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	5
Loire	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	5
Rhone	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	5
Rhone	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	5
Rhone	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	0	5
Rhone	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	5
Beaujolais	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	0	5
Beaujolais	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0	5
Beaujolais	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1	5
Beaujolais	0	0	1	1	0	0	1	0	0	1	0	0	0	0	1	0	5
Total	4	4	4	4	4	4	3	4	5	5	4	3	3	3	3	3	60
Profil.Global	0.067	0.067	0.067	0.067	0.067	0.067	0.050	0.067	0.083	0.083	0.067	0.050	0.050	0.050	0.050	0.050	
Total.Loire	3	1	0	0	1	3	0	2	2	2	2	0	1	1	1	1	20
Profil.Loire	0.150	0.050	0.000	0.000	0.050	0.150	0.000	0.100	0.100	0.100	0.100	0.000	0.050	0.050	0.050	0.050	
Total.Rhone	1	2	1	1	2	1	2	1	1	0	1	3	1	1	1	1	20
Profil.Rhone	0.050	0.100	0.050	0.050	0.100	0.050	0.100	0.050	0.050	0.000	0.050	0.150	0.050	0.050	0.050	0.050	
Total.Beaujolais	0	1	3	3	1	0	1	1	2	3	1	0	1	1	1	1	20
Profil.Beaujolais	0.000	0.050	0.150	0.150	0.050	0.000	0.050	0.050	0.100	0.150	0.050	0.000	0.050	0.050	0.050	0.050	

Nombre total de réponses

Nombre de réponses pour Loire

0.067 = 4/60

0.150 = 3/20

$$d^2(Loire, G) = \frac{1}{0.067}(0.15 - 0.067)^2 + \frac{1}{0.067}(0.050 - 0.067)^2 + \cdots + \frac{1}{0.050}(0.050 - 0.067)^2 = 0.490$$

$$d^2(Loire, Rhone) = \frac{1}{0.067}(0.15 - 0.050)^2 + \frac{1}{0.067}(0.050 - 0.100)^2 + \cdots + \frac{1}{0.050}(0.050 - 0.050)^2 = 1.325$$

Distance du KHI-2 : écart entre profils, pondéré par l'inverse de la fréquence marginale des modalités

$$d^2(k, k') = \sum_{l=1}^{L} \frac{1}{n_{.l}/n} \left(\frac{n_{kl}}{n_{k.}} - \frac{n_{k'l}}{n_{k'.}} \right)^2$$

Distances (Exemple)

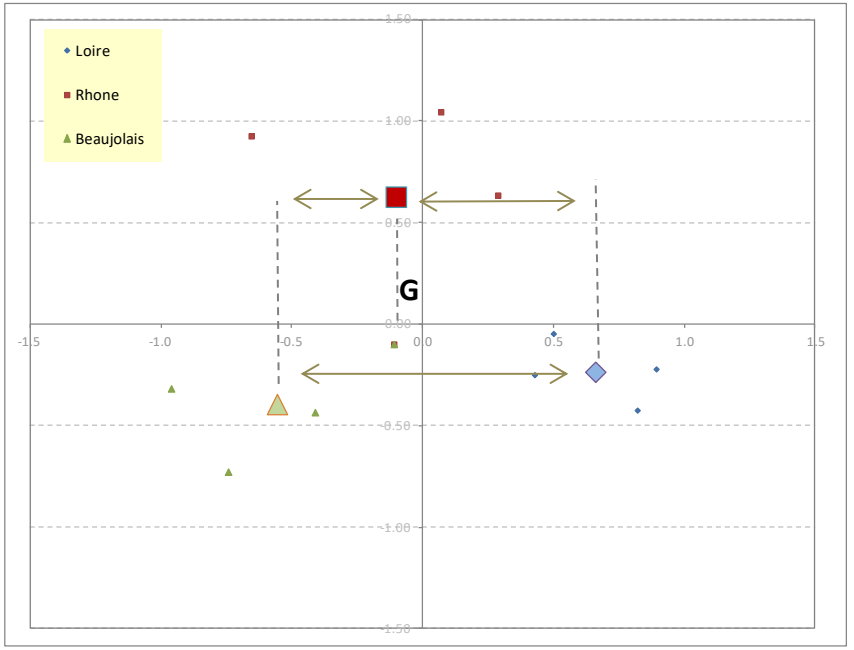
Les centres de classes (de groupes) sont tous à peu près à égale distance du barycentre global

d^2(Loire,G)	0.490
d^2(Rhone,G)	0.405
d^2(Beaujolais,G)	0.465

Les centres de classes (de groupes) sont tous à peu près à égale distance entre eux

d^2(k,k')	Loire	Rhone	Beaujolais
Loire	0	1.325	1.505
Rhone	1.325	0	1.25
Beaujolais	1.505	1.25	0

Objectif de l'analyse des correspondances discriminante : comment représenter ces proximités avec des graphiques dont on peut contrôler (quantifier) la fidélité (*) ; comment caractériser l'appartenance aux groupes à l'aide des variables descriptives (**).



(*) Ex. Si on ne prend que le premier axe, est-ce que le graphique retraduit les résultats ci-dessus ?

(**) Qu'est-ce qui oppose « Beaujolais » et « Loire » sur le 1er axe ? Qu'est-ce qui oppose « Rhône » à « Beaujolais – Loire » sur le 2nd axe ?

Analyse des Correspondances Discriminante (ACD)

Effectuer une transformation judicieuse des données

Et s'appuyer sur les résultats de l'analyse factorielle des correspondances (AFC)

Du tableau de données à un tableau de contingence un peu particulier

Region	Woody	Fruity	Sweet	Alcohol	Hedonic
Loire	A	C	B	A	A
Loire	B	C	C	B	C
Loire	A	B	B	A	B
Loire	A	C	C	B	D
Rhone	A	B	A	C	C
Rhone	B	A	A	C	B
Rhone	C	B	B	B	A
Rhone	B	C	C	C	D
Beaujolais	C	A	C	A	A
Beaujolais	B	A	C	A	B
Beaujolais	C	B	B	B	D
Beaujolais	C	A	A	A	C

Transformation des variables en indicatrices + consolidation par groupe : **la construction des facteurs se fera à partir des informations sur les centres de classes.**

N.B. Les tableaux croisés (Groupe x Descripteur) sont accolées entre eux : **les comptages sont dupliqués**. Ex. Somme(Loire) = 20 parce que 4 individus « Loire » x 5 variables (1 réponse par variable) = 20.

Region	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D	Total
Loire	3	1	0	0	1	3	0	2	2	2	2	0	1	1	1	1	20
Rhone	1	2	1	1	2	1	2	1	1	0	1	3	1	1	1	1	20
Beaujolais	0	1	3	3	1	0	1	1	2	3	1	0	1	1	1	1	20
Total	4	4	4	4	4	4	3	4	5	5	4	3	3	3	3	3	60

Notation

Y / X	x_1	x_l	x_L	Σ
y_1				
		$\vdots$		
y_k	$\cdots$	n_{kl}	$\cdots$	$n_{k.}$
		$\vdots$		
y_K				
Σ		$n_{.l}$		n

Profils

P(Loire / Woody_A) = 3 / 4 = 75% → 75% des vins présentant la propriété « Woody = A » sont des vins de Loire.

P(Woody_A / Loire) = 3 / 20 = 15% → pour les vins de Loire, la propriété « Woody = A » a été attribuée dans 15% des réponses possibles (et non pas '15%' des vins de Loire ont la propriété 'Woody = A ', cette lecture est fausse ici).

Pourquoi l'AFC est-elle applicable ?

(1) Tableau de valeurs positives

(2) On peut lire les marges

(3) Les profils sont interprétables

→ On peut appliquer l'AFC !!!

Matrice N

Region	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D	Total
Loire	3	1	0	0	1	3	0	2	2	2	2	0	1	1	1	1	20
Rhone	1	2	1	1	2	1	2	1	1	0	1	3	1	1	1	1	20
Beaujolais	0	1	3	3	1	0	1	1	2	3	1	0	1	1	1	1	20
Total	4	4	4	4	4	4	3	4	5	5	4	3	3	3	3	3	60

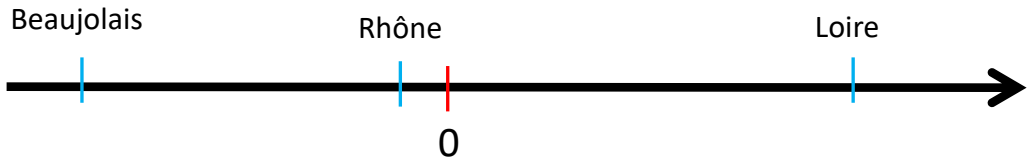
Profil Ligne

Region	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D	Total
Loire	0.15	0.05	0.00	0.00	0.05	0.15	0.00	0.10	0.10	0.10	0.10	0.00	0.05	0.05	0.05	0.05	1
Rhone	0.05	0.10	0.05	0.05	0.10	0.05	0.10	0.05	0.05	0.00	0.05	0.15	0.05	0.05	0.05	0.05	1
Beaujolais	0.00	0.05	0.15	0.15	0.05	0.00	0.05	0.05	0.10	0.15	0.05	0.00	0.05	0.05	0.05	0.05	1
Total	0.07	0.07	0.07	0.07	0.07	0.07	0.05	0.07	0.08	0.08	0.07	0.05	0.05	0.05	0.05	0.05	1

Profil colonne

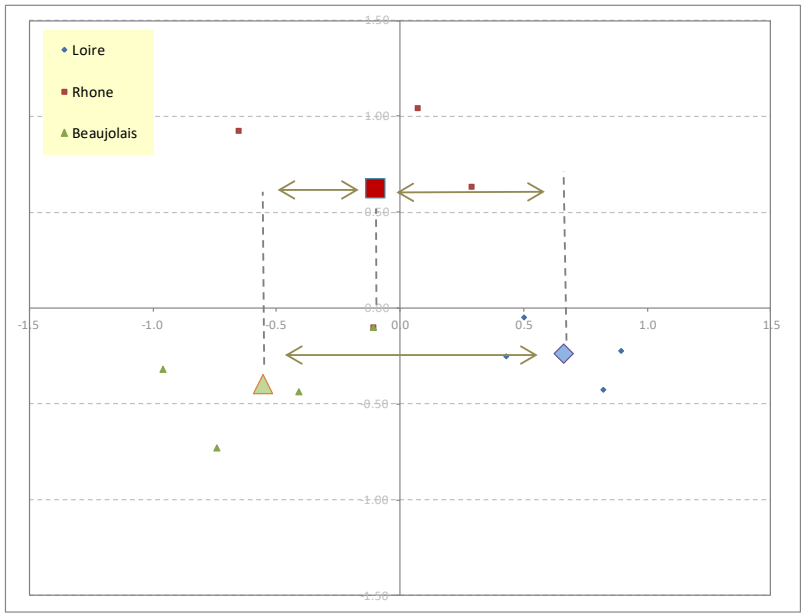
Region	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D	Total
Loire	0.750	0.250	0.000	0.000	0.250	0.750	0.000	0.500	0.400	0.400	0.500	0.000	0.333	0.333	0.333	0.333	0.333
Rhone	0.250	0.500	0.250	0.250	0.500	0.250	0.667	0.250	0.200	0.000	0.250	1.000	0.333	0.333	0.333	0.333	0.333
Beaujolais	0.000	0.250	0.750	0.750	0.250	0.000	0.333	0.250	0.400	0.600	0.250	0.000	0.333	0.333	0.333	0.333	0.333
Total	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Que fait l'AFC ?



- $\bar{z}_{k1}$ sont les « coordonnées factorielles » des modalités lignes (groupes) sur le 1er facteur
- La moyenne pondérée des points modalités est égale à 0
- λ représente la variance des points modalités → on cherche à maximiser λ
- Exactement ce que l'on cherche à faire avec l'analyse discriminante !!!
- Sauf que l'AFC effectue simultanément le même travail pour les modalités colonnes (introduit une contrainte supplémentaire dans les calculs)

$$\lambda_1 = \sum_{k=1}^{K} \frac{n_{k.}}{n} \times \bar{z}_{k1}^2$$



Inertie totale = Dispersion des centres de classes dans l'espace initial

$$\phi^2 = \sum_{k=1}^{K} \frac{n_{k.}}{n} \times d^2(G_k, G)$$

(1) Trouver les facteurs – combinaisons linéaires des indicatrices – qui maximisent les écarts entre les centres de classes

(2) Le nombre de facteurs H est égal à MIN(K-1, L-1). Souvent (K-1) puisque le nombre de groupes est plus faible que celui des indicatrices

(3) Les facteurs sont deux à deux orthogonaux

(4) Le facteur n°h explique l'écartement entre les centres de classes non pris en compte sur les (h-1) premiers facteurs

(5) Les calculs basés sur l'AFC assurent la décroissance de la variance expliquée (λ, SCE), pas celle du rapport de corrélation

(6) Le pouvoir de représentation d'un facteur est obtenu avec la part d'inertie reproduite

Analyse des Correspondances Discriminante (ACD)

Lecture des résultats

ACD – Inertie associée aux facteurs - Choix du nombre de facteurs

Trace = Inertie totale

$$\phi^2 = \sum_{k=1}^{K} \frac{n_{k.}}{n} \times d^2(G_k, G)$$

$$= \frac{4}{12} \times 0.490 + \frac{4}{12} \times 0.405 + \frac{4}{12} \times 0.465 = 0.4533$$

Trace = 0.4533

Inertie associée aux facteurs (variance inter)

$$\lambda_1 = \frac{SCE_1}{n} = \sum_{k=1}^{K} \frac{n_{k.}}{n} \bar{z}_{k1}^2 = 0.2519$$

$$\lambda_2 = 0.2014$$

avec

$$\phi^2 = \sum_{h=1}^{H} \lambda_h$$

Factor	Canonical Correlation R	Squared R	Total variation	Explained (between) variation		
				Eigen value	Proportion (%)	Cumulative (%)
1	0.8599	0.7394	0.3407	0.2519	55.56	55.56
2	0.8327	0.6934	0.2905	0.2014	44.44	100
Tot.				0.4533	100	-

Nombre max de facteurs

H_{Max} : MIN(K-1, L-1) = MIN(3-1,16-1) = 2

Variation totale sur le facteur (calculée ex post, après projection des individus)

$$\frac{SCT_h}{n} = \frac{1}{n} \sum_{i=1}^{n} (z_{ih} - \bar{z}_h)^2 = \frac{1}{n} \sum_{i=1}^{n} z_{ih}^2$$

La moyenne des points individus est nulle

Rapport de corrélation pour le facteur h

Indique le pouvoir discriminant du facteur - Il est normalisé ($0 \leq \eta^2 \leq 1$)

$$\eta_h^2 = \frac{SCE_h}{SCT_h}$$

Choix du nombre H de facteurs

1. Décroissance des valeurs propres
2. + Etude du rapport de corrélation

ACD – Coordonnées des groupes – Distances entre groupes

Distance à la moyenne globale cf. distance du KHI-2

Peut être obtenue en utilisant les coordonnées factorielles (distance euclidienne) sur l'ensembles des facteurs

Ex. d^2(Loire, G) = (0.65953 - **0**)² + (-0.23455 – **0**)² = 0.490

Coordonnées factorielles
Axe 1 : Loire vs. Beaujolais
Axe 2 : Rhône vs. Beaujolais/Loire

Inertie = Poids x Sq.Dist

4/12 = 0.333

Group centroids on canonical variables

Row Characterization				Coord.		Contributions (%)		COS	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos 1	cos 2
Loire	0.333	0.490	0.163	0.65953	-0.23455	57.56	9.1	0.888	0.112
Rhone	0.333	0.405	0.135	-0.10263	0.62807	1.39	65.27	0.026	0.974
Beaujolais	0.333	0.465	0.155	-0.55691	-0.39351	41.04	25.62	0.667	0.333

Squared distance between group centroids

-	Loire	Rhone	Beaujolais
Loire	0	1.325	1.505
Rhone	1.325	0	1.25
Beaujolais	1.505	1.25	0

Impact de la modalité dans la construction du facteur
$Somme_k(CTR) = 100\%$

Qualité de représentation de la modalité sur le facteur
$Somme_h(COS^2) = 100\%$

Aides à l'interprétation

Distance entre centres de classes cf. distance du KHI-2

Peut être obtenue en utilisant les coordonnées factorielles (distance euclidienne) sur l'ensembles des facteurs

Ex. d^2(Loire, Rhône) = [0.65953 – (-0.10263)]² + [-0.23455 – 0.62807]² = 1.325

Que se passe-t-il si on prend une solution de dimension inférieure ($H < H_{max}$) ? Ex. H = 1

Row Characterization				Coord.		Contributions (%)		COS	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos 1	cos 2
Loire	0.333	0.490	0.163	0.65953	-0.23455	57.56	9.1	0.888	0.112
Rhone	0.333	0.405	0.135	-0.10263	0.62807	1.39	65.27	0.026	0.974
Beaujolais	0.333	0.465	0.155	-0.55691	-0.39351	41.04	25.62	0.667	0.333

Distance entre les centres de classes

(selon le nombre de facteurs pris en compte)

Sq. dist. between group centroids (H = 2)

-	Loire	Rhone	Beaujolais
Loire	0	1.325	1.505
Rhone	1.325	0	1.25
Beaujolais	1.505	1.25	0

Sq. dist. between group centroids (H = 1)

-	Loire	Rhone	Beaujolais
Loire	0	0.581	1.480
Rhone	0.581	0	0.206
Beaujolais	1.480	0.206	0

Distances obtenues sur les données initiales

= Dist. obtenues si on prend tous les axes disponibles

Distances obtenues si on ne tient compte que du 1er facteur

1. L'écart entre « Loire vs. Beaujolais » est bien approximé parce que ces modalités sont bien représentées sur le premier facteur
2. Ce n'est pas le cas pour « Loire vs. Rhône » et « Beaujolais vs. Rhône »

ACD – Coordonnées des indicatrices fournies par l'AFC

Relation quasi-barycentrique

$$\bar{u}_{lh} = \frac{1}{\sqrt{\lambda_h}} \sum_{k=1}^{K} \frac{n_{kl}}{n_{.l}} \bar{z}_{kh}$$

Points modalités colonnes (indicatrices des descripteurs)

Points modalités lignes (classes)

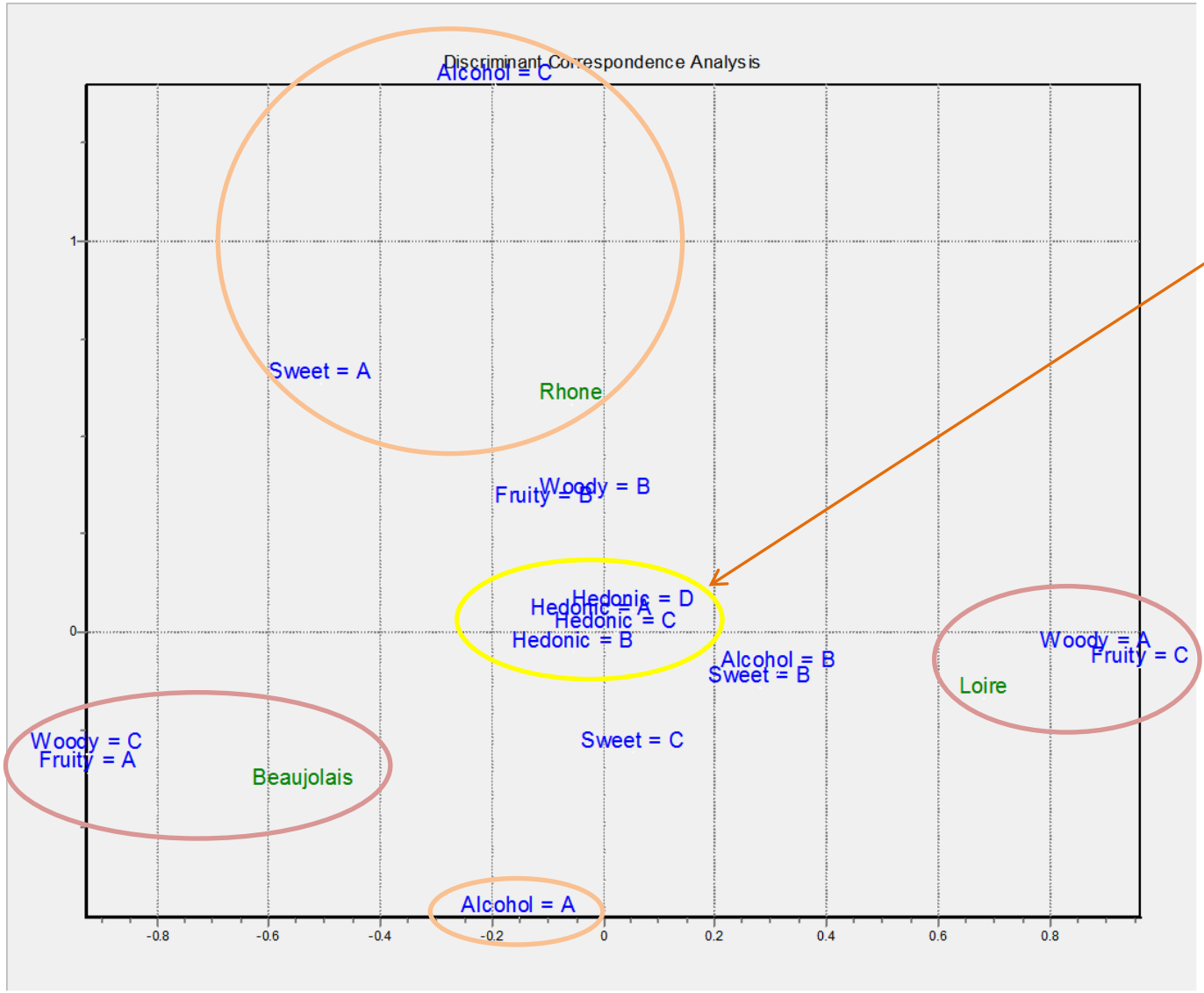
Canonical Structure

Column Characterization				Coord.		Contributions (%)		COS	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos 1	cos 2
Woody.A	0.06667	0.875	0.05833	0.93447	-0.0421	23.1	0.1	0.998	0.002
Woody.B	0.06667	0.125	0.00833	-0.05112	0.34984	0.1	4.1	0.021	0.979
Woody.C	0.06667	0.875	0.05833	-0.88335	-0.30773	20.7	3.1	0.892	0.108
Fruity.C	0.06667	0.875	0.05833	0.93447	-0.0421	23.1	0.1	0.998	0.002
Fruity.B	0.06667	0.125	0.00833	-0.05112	0.34984	0.1	4.1	0.021	0.979
Fruity.A	0.06667	0.875	0.05833	-0.88335	-0.30773	20.7	3.1	0.892	0.108
Sweet.B	0.06667	0.125	0.00833	0.32853	-0.13065	2.9	0.6	0.863	0.137
Sweet.C	0.08333	0.08	0.00667	0.0409	-0.27987	0.1	3.2	0.021	0.979
Sweet.A	0.05	0.66667	0.03333	-0.5062	0.64065	5.1	10.2	0.384	0.616
Alcohol.A	0.08333	0.56	0.04667	-0.14013	-0.73509	0.6	22.4	0.035	0.965
Alcohol.B	0.06667	0.125	0.00833	0.32853	-0.13065	2.9	0.6	0.863	0.137
Alcohol.C	0.05	2	0.1	-0.20448	1.39935	0.8	48.6	0.021	0.979
Hedonic.A	0.05	0	0	0	0	0	0	0	0
Hedonic.C	0.05	0	0	0	0	0	0	0	0
Hedonic.B	0.05	0	0	0	0	0	0	0	0
Hedonic.D	0.05	0	0	0	0	0	0	0	0

Coordonnées factorielles

Aides à l'interprétation

ACD – Représentation simultanée



Les vins ne se distinguent pas par le plaisir qu'ils procurent

Après « jittering » pour dépasser la superposition des points…

Analyse des Correspondances Discriminante (ACD)

Projection des individus supplémentaires

Obtenir les coordonnées factorielles d'un individu quelconque

Utiliser la relation quasi-barycentrique, dans l'autre sens :

1. Décrire l'individu par ses indicatrices
2. Transformer la description en profil ligne
3. Calculer la coordonnée avec la relation

$$z_{ih} = \frac{1}{\sqrt{\lambda_h}} \sum_{l=1}^{L} \frac{d_{il}}{J} \bar{u}_{ih}$$

0.2 = 1 / 5

Ex. Coordonnées du 1er individu de la base

1er individu	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D
Description	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0
Profil	**0.2**	0.0	0.0	0.0	0.0	**0.2**	0.0	**0.2**	0.0	**0.2**	0.0	0.0	**0.2**	0.0	0.0	0.0

Valeurs propres de l'analyse

Eigen value	
1	2
0.251888	0.201445

Coordonnées factorielles des modalités colonnes

-	Coord.	
Values	coord 1	coord 2
Woody_A	0.93447	-0.0421
Woody_B	-0.05112	0.34984
Woody_C	-0.88335	-0.30773
Fruity_A	-0.88335	-0.30773
Fruity_B	-0.05112	0.34984
Fruity_C	0.93447	-0.0421
Sweet_A	-0.5062	0.64065
Sweet_B	0.32853	-0.13065
Sweet_C	0.0409	-0.27987
Alcohol_A	-0.14013	-0.73509
Alcohol_B	0.32853	-0.13065
Alcohol_C	-0.20448	1.39935
Hedonic_A	0.0000	0.0000
Hedonic_B	0.0000	0.0000
Hedonic_C	0.0000	0.0000
Hedonic_D	0.0000	0.0000

Coordonnées de l'individu sur le 1er facteur

$$z_{11} = \frac{1}{\sqrt{0.251888}} \left(0.2 \times 0.93447 + 0.0 \times (-0.05112) + \cdots + 0.0 \times 0.0000\right) = 0.8198$$

Coordonnées de l'individu sur le 2nd facteur

$$z_{12} = \frac{1}{\sqrt{0.201445}} \left(0.2 \times (-0.0421) + 0.0 \times 0.34984 + \cdots + 0.0 \times 0.0000\right) = -0.4233$$

Projection des individus de la base de données dans l'espace factoriel

Tanagra fournit des coefficients applicables directement sur les indicatrices : les **coefficients canoniques**

Canonical Coefficients

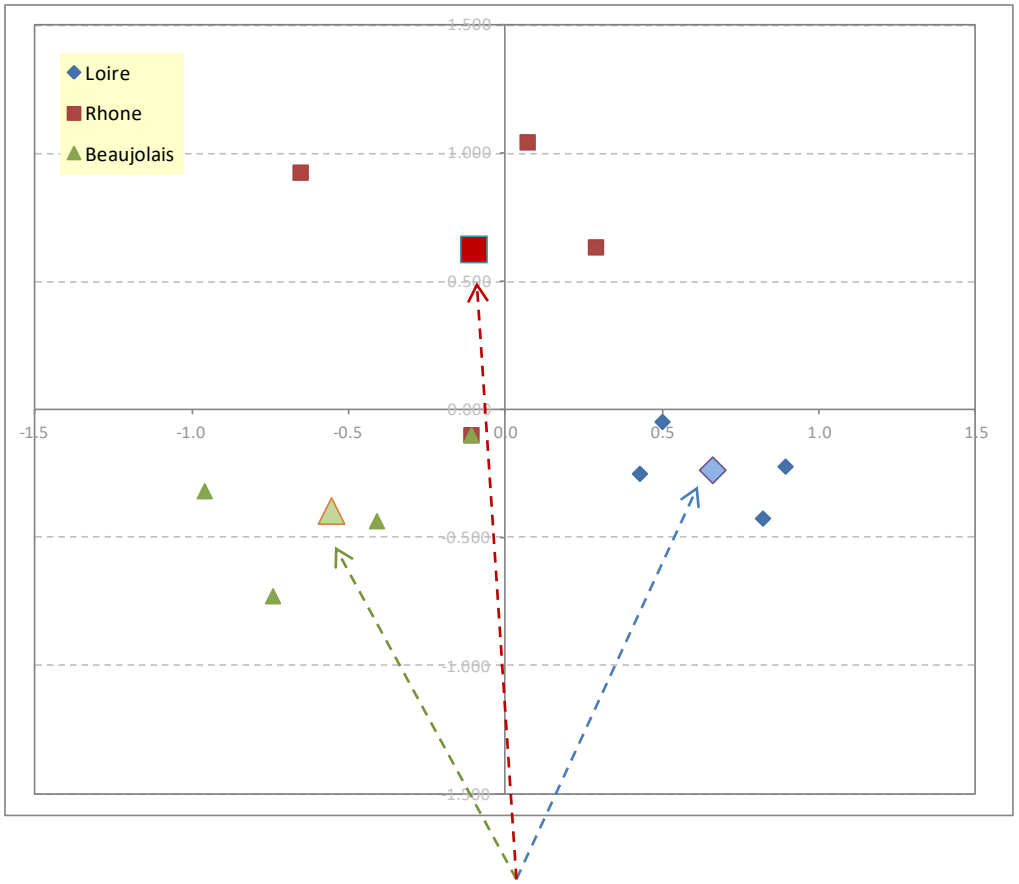
Applied to the indicator matrix i.e. columns are dumi

Attribute.Value	Factor 1	Factor 2
Woody.A	0.3724	-0.0188
Woody.B	-0.0204	0.1559
Woody.C	-0.3520	-0.1371
Fruity.C	0.3724	-0.0188
Fruity.B	-0.0204	0.1559
Fruity.A	-0.3520	-0.1371
Sweet.B	0.1309	-0.0582
Sweet.C	0.0163	-0.1247
Sweet.A	-0.2017	0.2855
Alcohol.A	-0.0558	-0.3276
Alcohol.B	0.1309	-0.0582
Alcohol.C	-0.0815	0.6236
Hedonic.A	0.0000	0.0000
Hedonic.C	0.0000	0.0000
Hedonic.B	0.0000	0.0000
Hedonic.D	0.0000	0.0000

$$a_{ilh} = \frac{\bar{u}_{lh}}{J \times \sqrt{\lambda_h}}$$

$$z_{ih} = \sum_{l=1}^{L} a_{ilh} \times d_{il}$$

Projection des 12 individus de la base initiale



Les moyennes conditionnelles calculées sur les facteurs correspondent aux coordonnées factorielles des modalités lignes (des groupes) fournies par l'ACD

A quelle région rattacher un vin avec les caractéristiques suivantes ?

Region	Woody	Fruity	Sweet	Alcohol	Hedonic
???	A	C	B	B	A

Application des coefficients canoniques sur les indicatrices

Canonical Coefficients

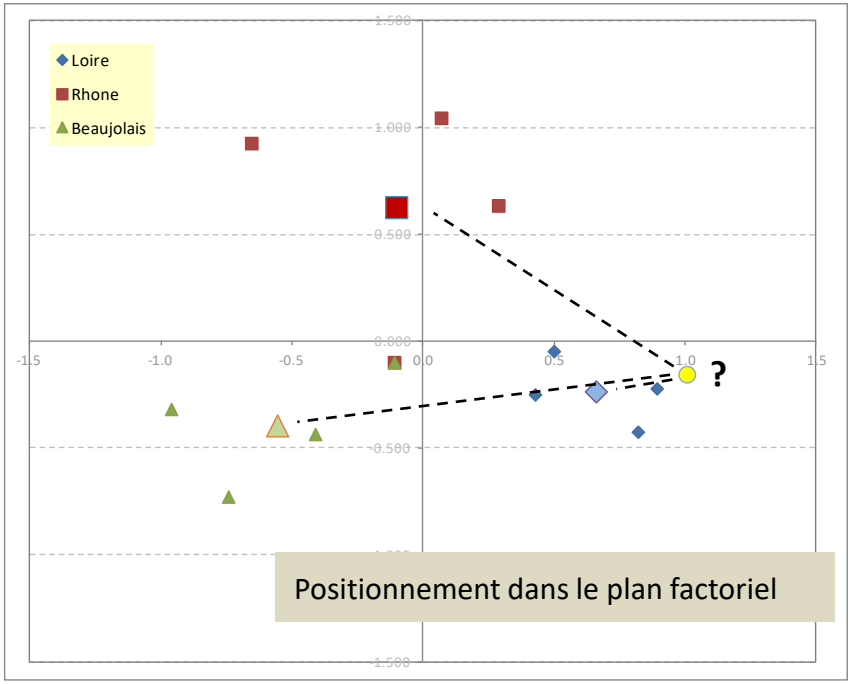
Applied to the indicator matrix i.e. columns are dummy variables

Attribute.Value	Factor 1	Factor 2	Dummy data
Woody = A	0.3724	-0.0188	1
Woody = B	-0.0204	0.1559	0
Woody = C	-0.3520	-0.1371	0
Fruity = C	0.3724	-0.0188	1
Fruity = B	-0.0204	0.1559	0
Fruity = A	-0.3520	-0.1371	0
Sweet = B	0.1309	-0.0582	1
Sweet = C	0.0163	-0.1247	0
Sweet = A	-0.2017	0.2855	0
Alcohol = A	-0.0558	-0.3276	0
Alcohol = B	0.1309	-0.0582	1
Alcohol = C	-0.0815	0.6236	0
Hedonic = A	0.0000	0.0000	1
Hedonic = C	0.0000	0.0000	0
Hedonic = B	0.0000	0.0000	0
Hedonic = D	0.0000	0.0000	0

Coordonnées sur les 2 facteurs

Coord	1.007	-0.154

$$1.007 = 0.3724 \times 1 - 0.0204 \times 0 + \cdots + 0.0000 \times 0$$



Positionnement dans le plan factoriel

$d^2(\omega,\text{Loire}) = (1.007-0.65953)^2+(-0.154-(-0.23455)^2 = 0.127$

$d^2(\omega,\text{Rhône}) = (1.007-(-0.10263))^2+(-0.154-0.62807)^2 = 1.842$

$d^2(\omega,\text{Beaujolais}) = (1.007-0.65953)^2+(-0.154-(-0.23455)^2 = 2.502$

« Loire » est le centre de classe qui lui est le plus proche (au sens de la distance euclidienne dans le repère factoriel) → c'est fort probablement un vin de Loire

N.B. Si les classes ne sont équilibrées, il faut exploiter leur prévalence : **distance généralisée**.

Les mauvaises pistes pour l'analyse factorielle discriminante pour descripteurs qualitatifs

Multiplier les analyses bi-variées, AFD sur indicatrices

Analyse bivariée – Classe vs. chaque descripteur

Mesures de l'association

Results

Row (Y)	Column (X)	Statistical indicator	
Region	Alcohol	Stat	Value
		d.f.	4
		Tschuprow's t	0.622495
		Cramer's v	0.622495
		Phi²	0.775000
		Chi² (p-value)	9.30 (0.0540)
		Lambda	0.500000
		Tau (p-value)	0.3875 (0.0741)
		U(R/C) (p-value)	0.4293 (0.0232)

Cross-tab

	A	B	C	Sum
Loire	2 (+ 1 %)	2 (+ 4 %)	0 (- 11 %)	4
Rhone	0 (- 18 %)	1 (- 1 %)	3 (+ 43 %)	4
Beaujolais	3 (+ 11 %)	1 (- 1 %)	0 (- 11 %)	4
Sum	5	4	3	12 *100%*

Row (Y)	Column (X)	Statistical indicator	
Region	Fruity	Stat	Value
		d.f.	4
		Tschuprow's t	0.559017
		Cramer's v	0.559017
		Phi²	0.625000
		Chi² (p-value)	7.50 (0.1117)
		Lambda	0.500000
		Tau (p-value)	0.3125 (0.1426)
		U(R/C) (p-value)	0.3433 (0.0598)

Cross-tab

	C	B	A	Sum
Loire	3 (+ 28 %)	1 (- 1 %)	0 (- 18 %)	4
Rhone	1 (- 1 %)	2 (+ 4 %)	1 (- 1 %)	4
Beaujolais	0 (- 18 %)	1 (- 1 %)	3 (+ 28 %)	4
Sum	4	4	4	12 *100%*

Contributions au KHI-2, attractions et répulsions entre les modalités

Ca marche, mais…

1. Très vite inextricable dès que le nombre de descripteurs augmente
2. Difficile perception du rôle conjoint des modalités des descripteurs
3. Aucune indication sur la proximité entre les classes

AFD usuelle sur les indicatrices des descripteurs

Enlever une des indicatrices pour chaque descripteur (modalité de référence)

Retirer la variable « Hedonic » (mêmes valeurs pour les 3 groupes)

Roots and Wilks' Lambda

Root	Eigenvalue	Proportion	Canonical R	Wilks Lambda	CHI-2	d.f.	p-value
1	8.25379	0.75034	0.944424	0.028846	19.5018	16	0.243503
2	2.74621	1	0.856192	0.266936	7.2641	7	0.401909

Test of H0: The canonical correlation in the current row and all that follow are zero (Bartlett's chi-square approximation)

Factor Structure Matrix - Correlations

Root	Root n1			Root n2		
Descriptors	Total	Within	Between	Total	Within	Between
Woody_A	-0.597142	-0.261732	-0.85262	0.403674	0.278082	0.522532
Woody_B	0.178481	0.060596	0.674246	0.215637	0.115065	0.738507
Fruity_A	0.418661	0.183502	0.597779	-0.619311	-0.42663	-0.801661
Fruity_B	0.178481	0.060596	0.674246	0.215637	0.115065	0.738507
Sweet_A	0.475769	0.177341	0.953168	0.16652	0.097553	0.302442
Sweet_B	-0.258541	-0.087778	-0.976689	0.062679	0.033446	0.214661
Alcohol_A	-0.264769	-0.112365	-0.39537	-0.678498	-0.452561	-0.918522
Alcohol_B	-0.258541	-0.087778	-0.976689	0.062679	0.033446	0.214661

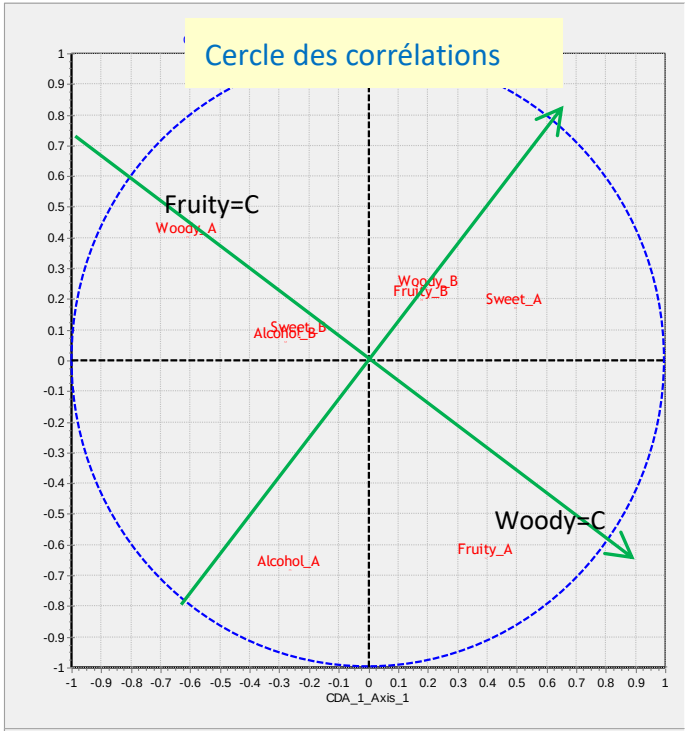
Group centroids on the canonical variables

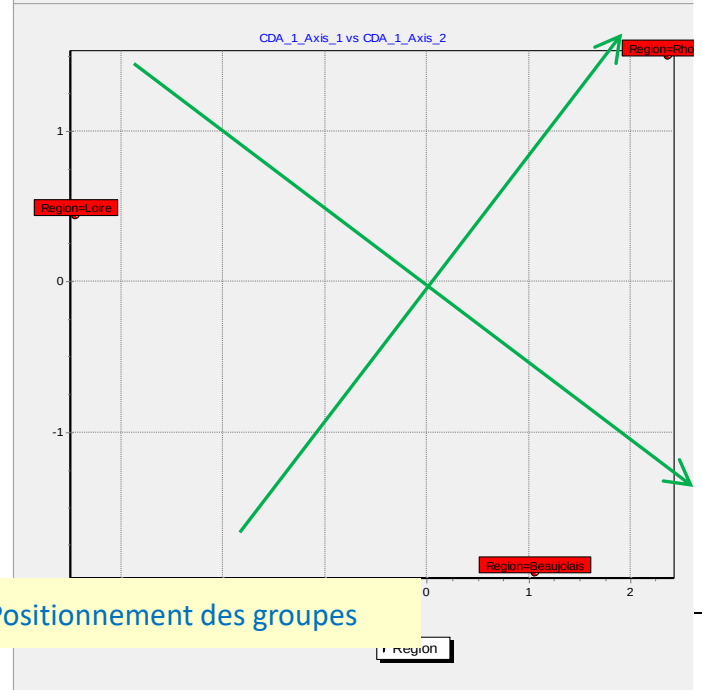
Region	Root n1	Root n2
Loire	-3.436595	0.435678
Rhone	2.372415	1.498881
Beaujolais	1.064179	-1.93456
Sq Canonical	0.891936	0.733064

Ca marche (plus ou moins, une rotation des axes est nécessaire) mais…

1. Une covariance sur les indicatrices d'une même variable (???)
2. On ne perçoit pas le rôle des modalités de référence

Cercle des corrélations

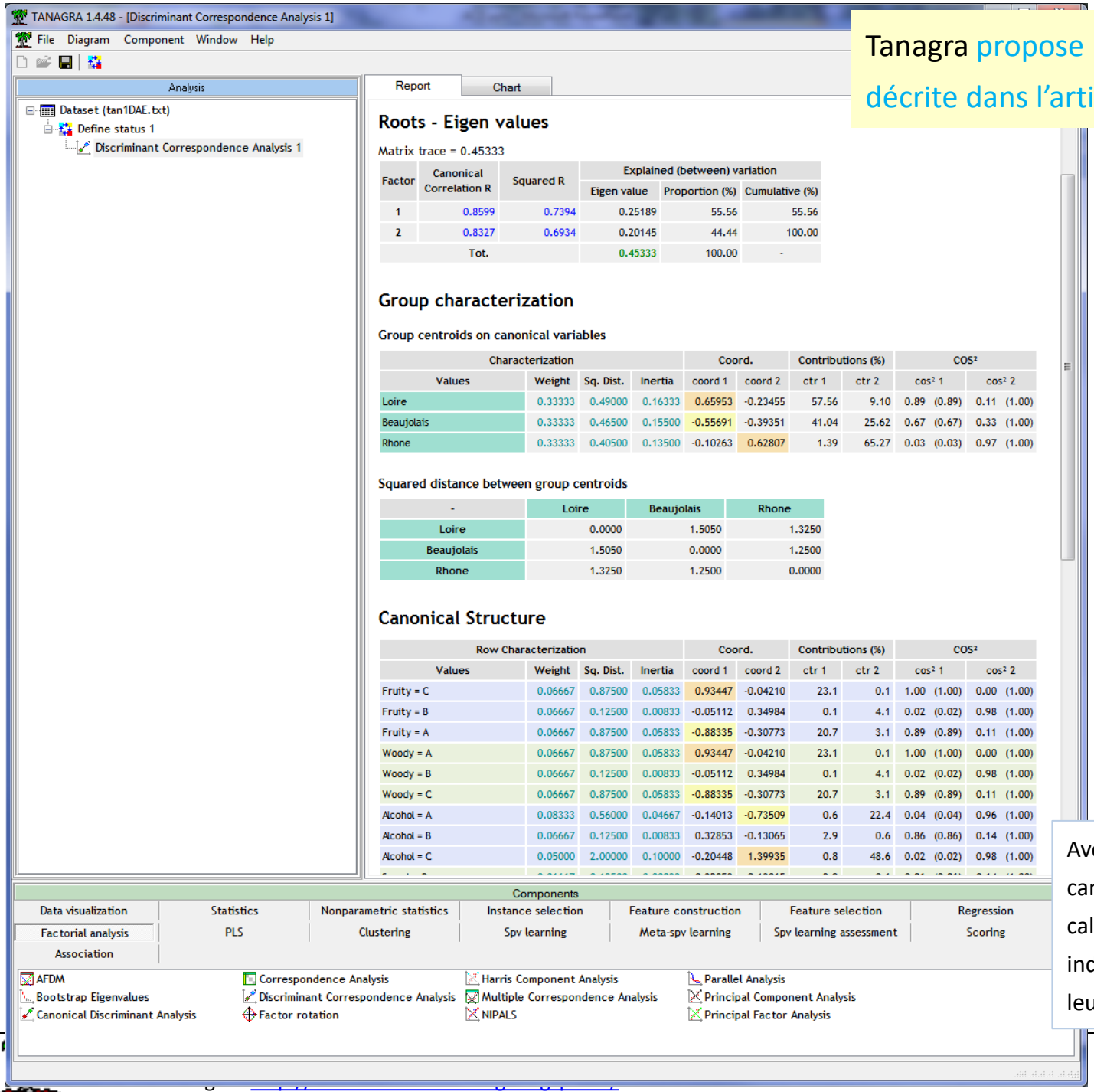




Positionnement des groupes

Les logiciels

Tanagra, R



Tanagra propose la méthode telle est décrite dans l'article de référence

Avec en prime les fonctions canoniques : les fonctions pour calculer les coordonnées des individus supplémentaires à partir de leurs indicatrices.

R – (1) Construire le tableau de contingence

```
#importer les données
library(xlsx)
wine <- read.xlsx(file="french_wine_dca.xls",sheetIndex=1,header=T)
print(summary(wine))

#sélectionner les descripteurs
descriptors <- subset(wine,select=-1)
print(summary(descriptors))

#tableau croisé entre la variable cible (référence) et un descripteur (x)
cross.tab <- function(x,ref){
 m <- table(ref,x)
 return(m)
}

#appliquer cross.tab() sur la liste des descripteurs
dataset <- lapply(descriptors,cross.tab,ref=wine$Region)

#créer le tableau de contingence global pour l'AFC
#à partir des tableaux croisés individuels
matrix.ca <- NULL
for (j in 1:ncol(descriptors)){
 m <- dataset[[j]]
 colnames(m) <- paste(colnames(descriptors)[j],colnames(m),sep=".")
 matrix.ca <- cbind(matrix.ca,m)
}
print(matrix.ca)
```

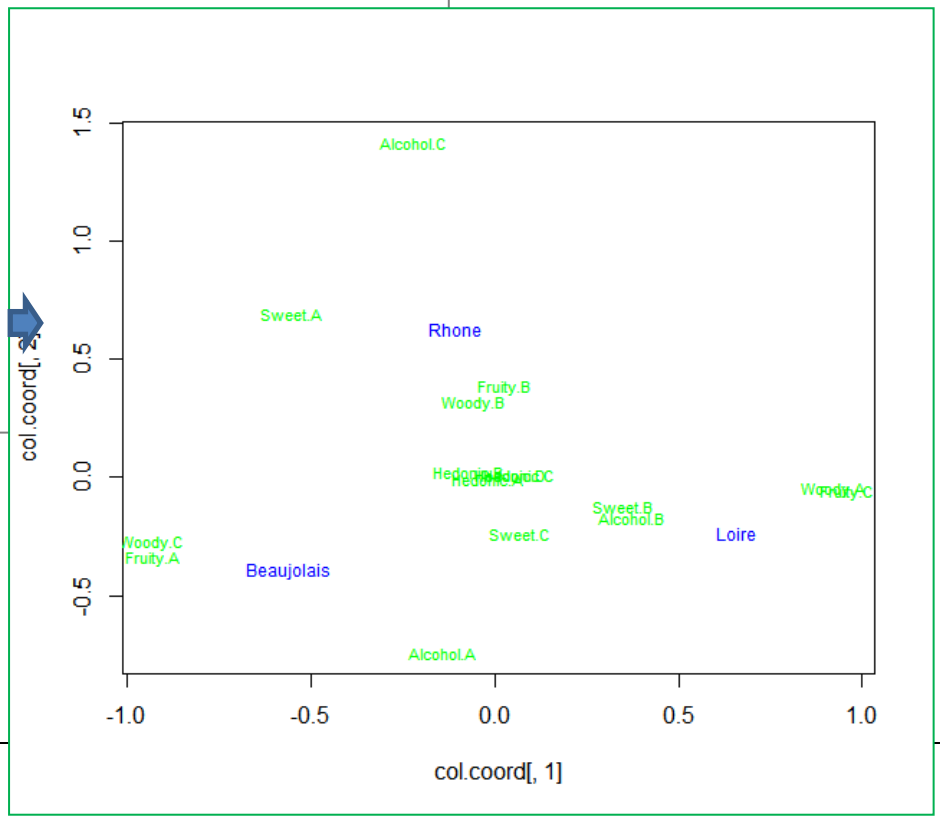
```
> print(summary(wine))
       Region  Woody Fruity Sweet Alcohol Hedonic
Beaujolais:4   A:4   A:4    A:3   A:5     A:3
Loire      :4   B:4   B:4    B:4   B:4     B:3
Rhone      :4   C:4   C:4    C:5   C:3     C:3
                                          D:3
```

```
> print(matrix.ca)
           Woody.A Woody.B Woody.C Fruity.A Fruity.B Fruity.C Sweet.A Sweet.B
Beaujolais       0       1       3        3        1        0       1       1
Loire            3       1       0        0        1        3       0       2
Rhone            1       2       1        1        2        1       2       1
           Sweet.C Alcohol.A Alcohol.B Alcohol.C Hedonic.A Hedonic.B Hedonic.C
Beaujolais       2         3         1         0         1         1         1
Loire            2         2         2         0         1         1         1
Rhone            1         0         1         3         1         1         1
           Hedonic.D
Beaujolais         1
Loire              1
Rhone              1
```

R – (2) Lancer l'AFC

```
#charger le package « ca » (http://cran.r-project.org/web/packages/ca/index.html)
library(ca)

#analyse des correspondances
fit <- ca(matrix.ca,nd=2)
print(fit)

#coordonnées des modalités lignes (identiques à celles de Tanagra)
row.coord <- cbind(fit$rowcoord[,1]*fit$sv[1],fit$rowcoord[,2]*fit$sv[2])
rownames(row.coord) <- rownames(matrix.ca)
print(row.coord)

#coordonnées des modalités colonnes
col.coord <- cbind(fit$colcoord[,1]*fit$sv[1],fit$colcoord[,2]*fit$sv[2])
#add jittering to avoid overplot
col.coord[,1] <- jitter(col.coord[,1],10)
col.coord[,2] <- jitter(col.coord[,2],10)
rownames(col.coord) <- colnames(matrix.ca)

#représentation simultanée
plot(col.coord[,1],col.coord[,2],type="n")
text(col.coord[,1],col.coord[,2],labels=rownames(col.coord), >>
>> col="green",cex=0.65)
text(row.coord[,1],row.coord[,2],labels=rownames(row.coord), >>
>> col="blue",cex=0.75)
```

```
> print(row.coord)
                [,1]       [,2]
Beaujolais -0.5569077 -0.3935147
Loire       0.6595342 -0.2345520
Rhone      -0.1026265  0.6280667
```



Etude de cas

« Les races canines »

(Tenenhaus, 2006 ; Tableau 8.1, page 254 – Saporta, 2006; Tableau 10.1, page 235)

Les races canines

Un exemple présenté dans le cadre de l'AFCM (analyse factorielle des correspondances multiples)

Chien	Taille	Poids	Velocite	Intelligence	Affection	Agressivite	Fonction
Beauceron	Taille++	Poids+	Veloc++	Intell+	Affec+	Agress+	utilite
Basset	Taille-	Poids-	Veloc-	Intell-	Affec-	Agress+	chasse
Berger All	Taille++	Poids+	Veloc++	Intell++	Affec+	Agress+	utilite
Boxer	Taille+	Poids+	Veloc+	Intell+	Affec+	Agress+	compagnie
Bull-Dog	Taille-	Poids-	Veloc-	Intell+	Affec+	Agress-	compagnie
Bull-Mastif	Taille++	Poids++	Veloc-	Intell++	Affec-	Agress+	utilite
Caniche	Taille-	Poids-	Veloc+	Intell++	Affec+	Agress-	compagnie
Chihuahua	Taille-	Poids-	Veloc-	Intell-	Affec+	Agress-	compagnie
Cocker	Taille+	Poids-	Veloc-	Intell+	Affec+	Agress+	compagnie
Colley	Taille++	Poids+	Veloc++	Intell+	Affec+	Agress-	compagnie
Dalmatien	Taille+	Poids+	Veloc+	Intell+	Affec+	Agress-	compagnie
Doberman	Taille++	Poids+	Veloc++	Intell++	Affec-	Agress+	utilite
Dogue All	Taille++	Poids++	Veloc++	Intell-	Affec-	Agress+	utilite
Epag. Breton	Taille+	Poids+	Veloc+	Intell++	Affec+	Agress-	chasse
Epag. Français	Taille++	Poids+	Veloc+	Intell+	Affec-	Agress-	chasse
Fox-Hound	Taille++	Poids+	Veloc++	Intell-	Affec-	Agress+	chasse
Fox-Terrier	Taille-	Poids-	Veloc+	Intell+	Affec+	Agress+	compagnie
Gd Bleu Gasc	Taille++	Poids+	Veloc+	Intell-	Affec-	Agress+	chasse
Labrador	Taille+	Poids+	Veloc+	Intell+	Affec+	Agress-	chasse
Levrier	Taille++	Poids+	Veloc++	Intell-	Affec-	Agress-	chasse
Mastiff	Taille++	Poids++	Veloc-	Intell-	Affec-	Agress+	utilite
Pekinois	Taille-	Poids-	Veloc-	Intell-	Affec+	Agress-	compagnie
Pointer	Taille++	Poids+	Veloc++	Intell++	Affec-	Agress-	chasse
St-Bernard	Taille++	Poids++	Veloc-	Intell+	Affec-	Agress+	utilite
Setter	Taille++	Poids+	Veloc++	Intell+	Affec-	Agress-	chasse
Teckel	Taille-	Poids-	Veloc-	Intell+	Affec+	Agress-	compagnie
Terre-Neuve	Taille++	Poids++	Veloc-	Intell+	Affec-	Agress-	utilite

En AFCM :

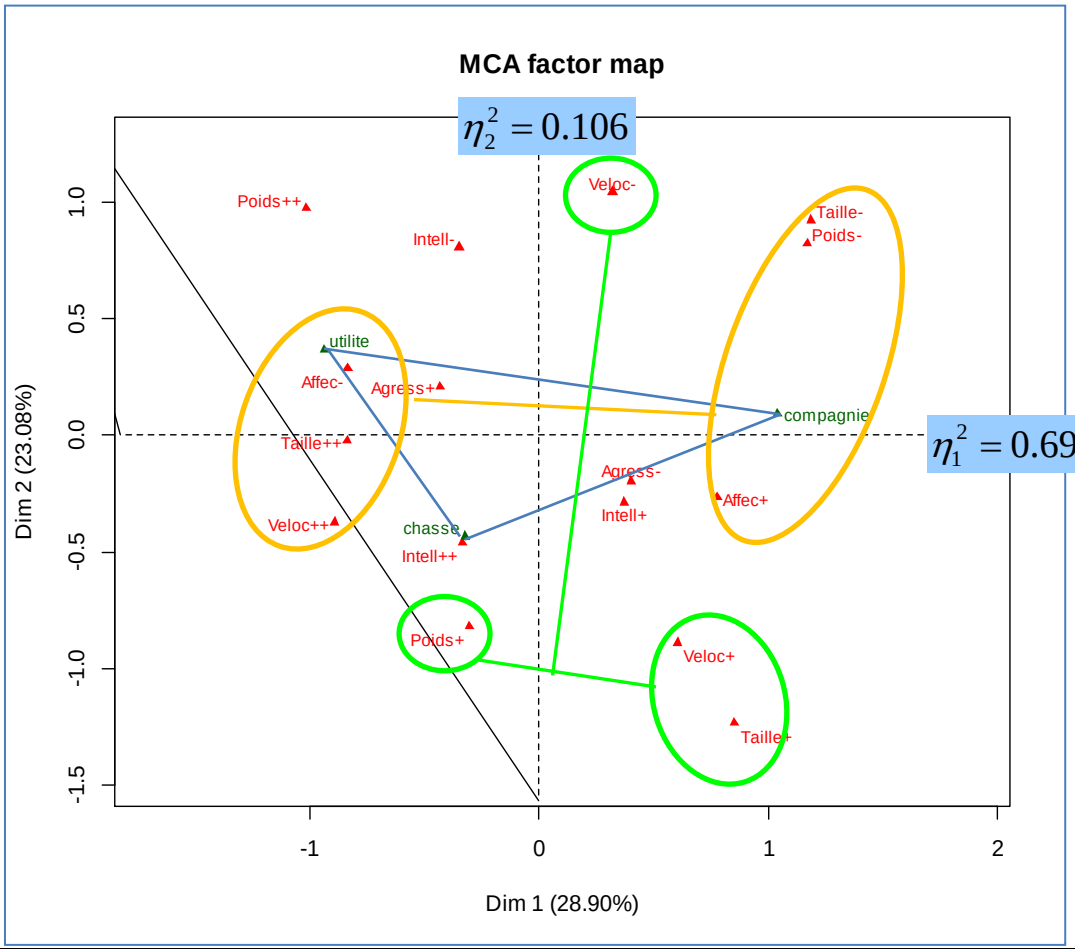
1. Dégager les principaux traits de caractères des chiens avec « Taille » … « Agressivité »
2. « Fonction » est une variable illustrative qui permet de (mieux) comprendre ces caractéristiques

```
#chargement des données
canines <- read.table(file="races_canines.txt",header=T,sep="\t",row.names=1)
summary(canines)
#charger le package
library(FactoMineR)
#lancer l'ACM
canines.acm <- MCA(canines,ncp=2,quali.sup=c(7),graph=F)
#graphique (modalités actives + modalités illustratives)
plot(canines.acm,cex=0.75,choix=« ind » ,invisible=« ind »)
```

Les races canines – Résultats de l'AFCM

1er axe : opposition « utilité » - « compagnie » forte. On perçoit clairement sur quelles propriétés est basée cette distinction.

2nd axe : on voit le pourquoi de l'axe, mais il caractérise plus ou moins bien les chiens de « chasse » par rapport aux autres. Cf. le rapport de corrélation correspondant.



Roots - Eigen values

Matrix trace = 0.46431

Factor	Canonical Correlation R	Squared R	Explained (between) variation Eigen value	Proportion (%)	Cumulative (%)
1	0.8574	0.7351	0.34586	74.49	74.49
2	0.7198	0.518	0.11845	25.51	100
Tot.			0.46431	100	-

Group characterization

Group centroids on canonical variables

Values	Characterization Weight	Sq. Dist.	Inertia	Coord. coord 1	coord 2
compagnie	0.37037	0.53718	0.19896	-0.71465	0.16265
utilite	0.2963	0.60841	0.18027	0.70531	0.33309
chasse	0.33333	0.25526	0.08509	0.16711	-0.4768

Squared distance between group centroids

-	compagnie	utilite	chasse
compagnie	0	2.0453	1.1864
utilite	2.0453	0	0.9456
chasse	1.1864	0.9456	0

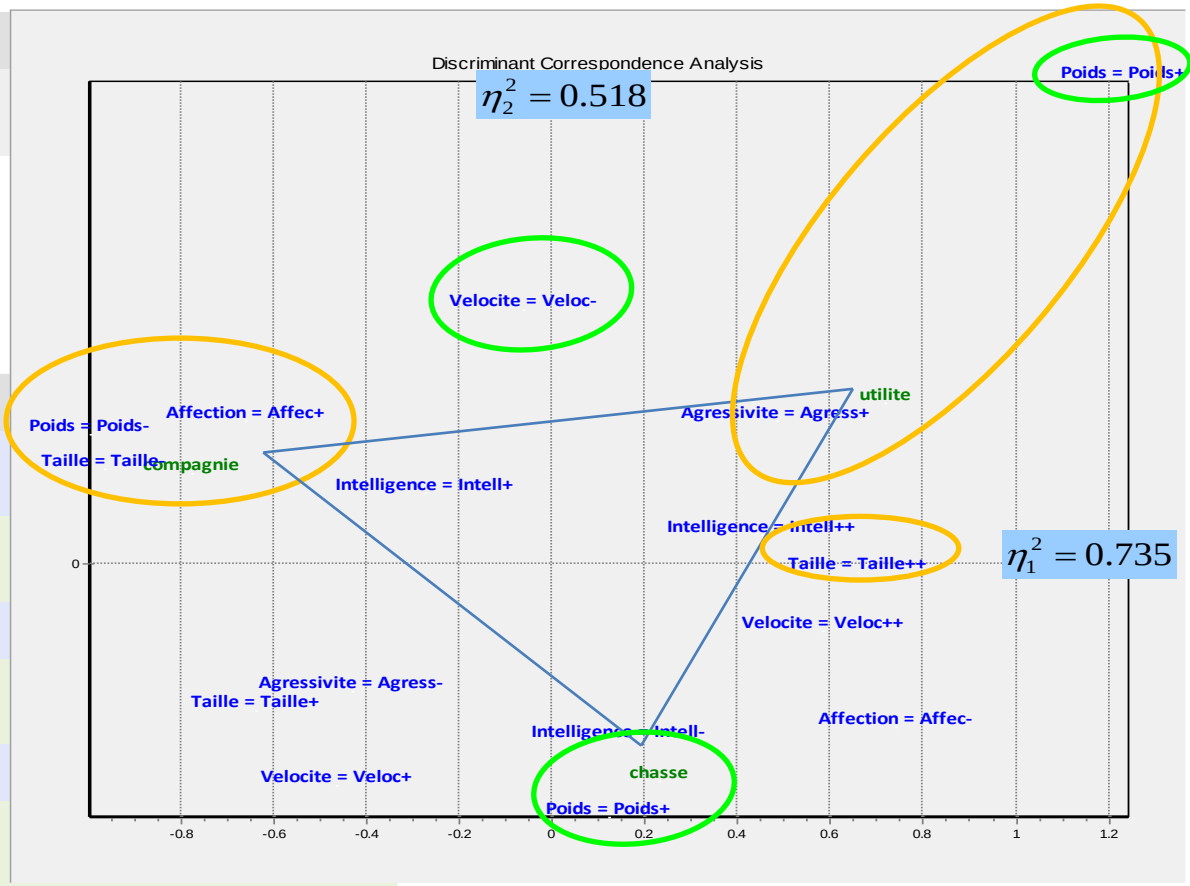
Canonical Structure

Values	Row Characterization Weight	Sq. Dist.	Inertia	Coord. coord 1	coord 2
Poids = Poids+	0.08642	0.25855	0.02234	0.15897	-0.48298
Poids = Poids-	0.04938	1.11406	0.05502	-1.02777	0.24033
Poids = Poids++	0.03086	2.375	0.0733	1.1993	0.96782
Taille = Taille++	0.09259	0.452	0.04185	0.67228	-0.00647
Taille = Taille-	0.04321	1.0449	0.04515	-1.00099	0.20716
Taille = Taille+	0.03086	0.452	0.01395	-0.61545	-0.2706
Affection = Affec+	0.08642	0.50765	0.04387	-0.65607	0.27791
Affection = Affec-	0.08025	0.58876	0.04725	0.70653	-0.29928
Velocite = Veloc++	0.05556	0.29259	0.01626	0.52429	-0.13307
Velocite = Veloc-	0.06173	0.245	0.01512	-0.09946	0.48488
Velocite = Veloc+	0.04938	0.425	0.02099	-0.46551	-0.4564
Agressivite = Agress+	0.08025	0.2821	0.02264	0.43093	0.31049
Agressivite = Agress-	0.08642	0.24324	0.02102	-0.40015	-0.28831
Intelligence = Intell+	0.08025	0.12234	0.00982	-0.31199	0.15811
Intelligence = Intell-	0.04938	0.12969	0.0064	0.13811	-0.33259
Intelligence = Intell++	0.03704	0.25208	0.00934	0.49184	0.10088

Les races canines – Résultats de l'ACD

On cherche à caractériser explicitement les fonctions des chiens.

Résultat : la différenciation des classes est meilleure, surtout sur le 2nd facteur !



Analyse des Correspondances Discriminante (ACD)

Conclusion

Analyse des correspondantes discriminante :

1. Technique factorielle descriptive.
2. Caractérisation de classes définies par une variable cible catégorielle à l'aide de variables descriptives qualitatives.
3. Le pendant de l'analyse factorielle discriminante lorsque les descripteurs sont qualitatifs.
4. La variable cible est active, elle participe aux calculs. La méthode va plus loin donc qu'une ACM (analyse des correspondances multiples) où l'éventuelle variable décrivant les classes est utilisée comme illustrative.
5. La méthode intègre toutes les modalités dans les sorties, il n'est pas nécessaire d'interpréter les modalités en fonction d'une modalité de référence qui aurait été omise.
6. La redondance entre les descripteurs n'est pas rédhibitoire.
7. Utilisable pour le classement d'individus supplémentaires.

Bibliographie

Article de référence

Abdi H., « Discriminant correspondence analysis », In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage. pp. 270-275, 2007 (http://www.utd.edu/~herve/Abdi-DCA2007-pretty.pdf).

Tutoriel (Traitement sous Tanagra et R)

Tutoriel Tanagra, « Analyse des correspondances discriminante », http://tutoriels-data-mining.blogspot.fr/2012/12/analyse-des-correspondances.html

Et les 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, 3ème édition, 2000.

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

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