1 Objectif

Analyse des correspondances discriminante.

L'analyse factorielle discriminante[1] ou analyse discriminante descriptive vise à décrire et à expliquer l'appartenance des observations à des groupes prédéfinis à partir d'un ensemble de variables explicatives (variables prédictives, descripteurs). Il s'agit d'une technique factorielle. Nous obtenons des variables latentes, combinaisons linéaires des descripteurs, qui permettent de discerner au mieux les groupes.

Etant basée sur la décomposition de la variance, l'analyse discriminante ne s'applique qu'aux variables explicatives quantitatives. La solution n'est pas directement transposable aux variables catégorielles (qualitatives). Certes, on pourrait toujours s'essayer à recoder ces dernières, en utilisant des indicatrices par exemple, mais il est à prévoir que les résultats obtenus seront pour le moins farfelus, inexploitables. En effet, que signifierait dans ce cas la covariance ou la corrélation de deux indicatrices issues d'une seule et même variable catégorielle ? Or, dans le cadre descriptif, l'objectif premier est d'interpréter les résultats fournis par l'analyse[2].

La littérature n'est pas très disserte à ce sujet. J'en étais donc resté à ce stade lorsque j'ai découvert récemment l'article d'Hervé Abdi (2007)[3]. Il propose d'exploiter les propriétés de l'analyse factorielle des correspondances (AFC) pour résoudre le problème de l'analyse discriminante descriptive sur variables catégorielles. L'approche, appelée « **discriminant correspondence analysis** » traduite librement par « **analyse des correspondances discriminante** »[4], repose sur une transformation ingénieuse des données « individus x variables » en un tableau de contingence un peu particulier. A la sortie nous obtenons des résultats qui décrivent les relations entre les modalités de la variable cible (qui définissent l'appartenance aux groupes) et celles des variables explicatives qualitatives. Il est même possible d'obtenir une représentation graphique révélant les attractions et répulsions.

Dans ce tutoriel, nous montrons la mise en œuvre de la méthode dans Tanagra 1.4.48. Nous reprendrons l'exemple de l'article de référence de la méthode. Il s'agit de caractériser la provenance des vins à partir de leurs propriétés. Notre objectif est d'expliquer pas à pas l'approche en associant les résultats de Tanagra à chaque étape de l'article.

Par la suite, nous reproduisons les calculs à l'aide d'un programme écrit pour le logiciel R.

---

[1] http://fr.wikipedia.org/wiki/Analyse_discriminante

[2] Dans le cadre de l'analyse discriminante prédictive (paramétrique) ou analyse discriminante linéaire (http://fr.wikipedia.org/wiki/Analyse_discriminante_linéaire), la situation est différente : on cherche à produire un classifieur efficace. Le codage 0/1 est moins sujet à caution même si, de manière évidente, il ne cadre pas avec les hypothèses de travail (normalité des distributions conditionnelles).

[3] H. Abdi, « Discriminant correspondence analysis », In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage. pp. 270-275, 2007.

[4] Ou bien « Analyse discriminante des correspondances » ? Je ne sais pas vraiment. Il faut traduire l'idée « utiliser l'analyse factorielle des correspondances pour résoudre un problème d'analyse discriminante sur variables qualitatives… ». Après, trouver une appellation simple, fidèle au concept, attractive, voilà un sujet pour linguistes.

2 Description des vins

2.1 Description des données

Nous disposons de n = 12 échantillons de vins en provenance de K = 3 régions différentes (Loire, Rhône, Beaujolais), décrits par p = 5 qualificatifs (Woody, Fruity, Sweet, Alcohol, Hedonic). Tous les descripteurs sont qualitatifs. Certains peuvent être issus d'une discrétisation (d'un découpage en classes). L'objectif est d'expliquer l'appartenance aux régions à partir des descripteurs.

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

Une première piste consiste à procéder à une simple analyse bivariée.

ALCOHOL

REGION	A	B	C	Sum
Loire	2	2	0	4
Rhone	0	1	3	4
Beaujolais	3	1	0	4
Sum	5	4	3	12
Cramer's V	0.622495			

FRUITY

REGION	C	B	A	Sum
Loire	3	1	0	4
Rhone	1	2	1	4
Beaujolais	0	1	3	4
Sum	4	4	4	12
Cramer's V	0.559017			

WOODY

REGION	A	B	C	Sum
Loire	3	1	0	4
Rhone	1	2	1	4
Beaujolais	0	1	3	4
Sum	4	4	4	12
Cramer's V	0.559017			

SWEET

REGION	B	C	A	Sum
Loire	2	2	0	4
Rhone	1	1	2	4
Beaujolais	1	2	1	4
Sum	4	5	3	12
Cramer's V	0.347611			

HEDONIC

REGION	A	C	B	D	Sum
Loire	1	1	1	1	4
Rhone	1	1	1	1	4
Beaujolais	1	1	1	1	4
Sum	3	3	3	3	12
Cramer's V	0				

Figure 1 - Tableaux croisés entre la cible (REGION) et les descripteurs

Nous croisons chaque descripteur avec la variable cible. Nous disposons ainsi d'une première indication sur les liaisons individuelles de chaque descripteur avec « Région ». Nous constatons qu'aucune liaison n'est significative à 5%. Ce n'est pas étonnant avec un effectif aussi faible (*n = 12)*.

Cependant, plusieurs relations sont assez fortes : ALCOHOL avec un V de Cramer de 0.62 ; FRUITY avec V = 0.56 ; WOODY avec V = 0.56. Il semble donc possible d'expliquer l'appartenance aux régions des vins à partir de leurs propriétés. Mais il faut le faire de manière multivariée c.-à-d. en tenant compte du rôle simultané de l'ensemble des descripteurs.

2.2 De l'analyse discriminante vers l'analyse des correspondances

Nous procédons à une transformation des données pour traiter la discrimination à l'aide de l'analyse des correspondances. Il s'agit d'accoler tous les tableaux de contingence individuels dans un tableau global à P colonnes (P = somme du nombre de modalités des p variables), mettant en relation la cible avec tous les descripteurs.

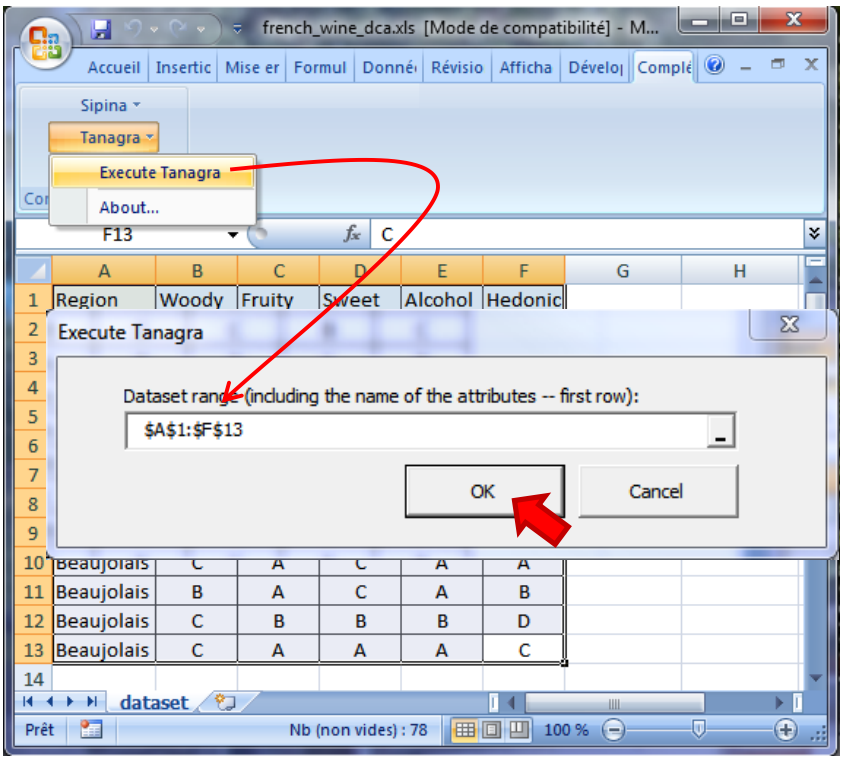
Region	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D	Total
Loire	3	1	0	0	1	3	0	2	2	2	2	0	1	1	1	1	20
Rhone	1	2	1	1	2	1	2	1	1	0	1	3	1	1	1	1	20
Beaujolais	0	1	3	3	1	0	1	1	2	3	1	0	1	1	1	1	20
Total	4	4	4	4	4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

Il y a manifestement duplication des données. L'effectif total dans cette nouvelle représentation est N = n x p = 12 x 5 = 60 « observations » (les guillemets sont importants). Ce n'est donc pas vraiment un tableau contingence au sens strict. En revanche, nous disposons d'une représentation que nous pouvons utiliser pour : évaluer les similitudes et différences entre les colonnes au regard des lignes (des valeurs de la variable cible) ; effectuer le même type d'étude pour les lignes au regard des colonnes ; appréhender les attractions et répulsions entre les modalités lignes et colonnes. C'est exactement le propos de l'analyse des correspondances.

2.3 Analyse avec Tanagra

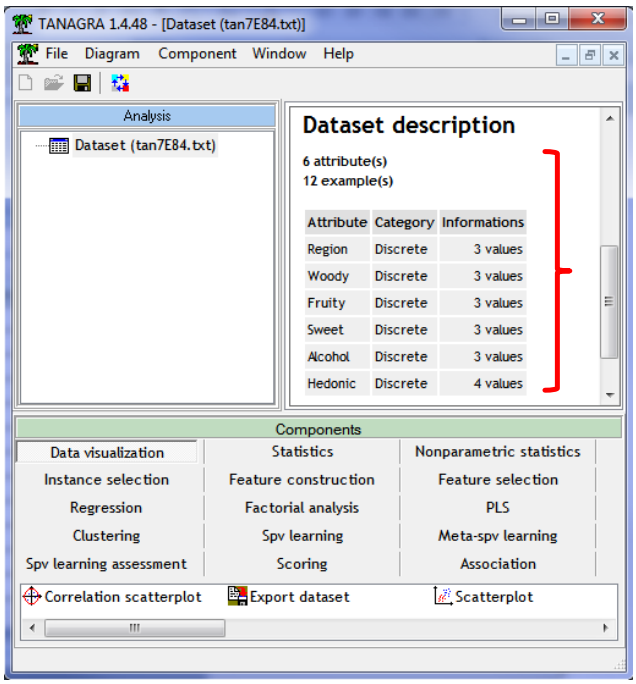
2.3.1 Importation des données

Le plus simple est de charger le fichier de données « **french_wine_dca.xls** » dans le tableur Excel.



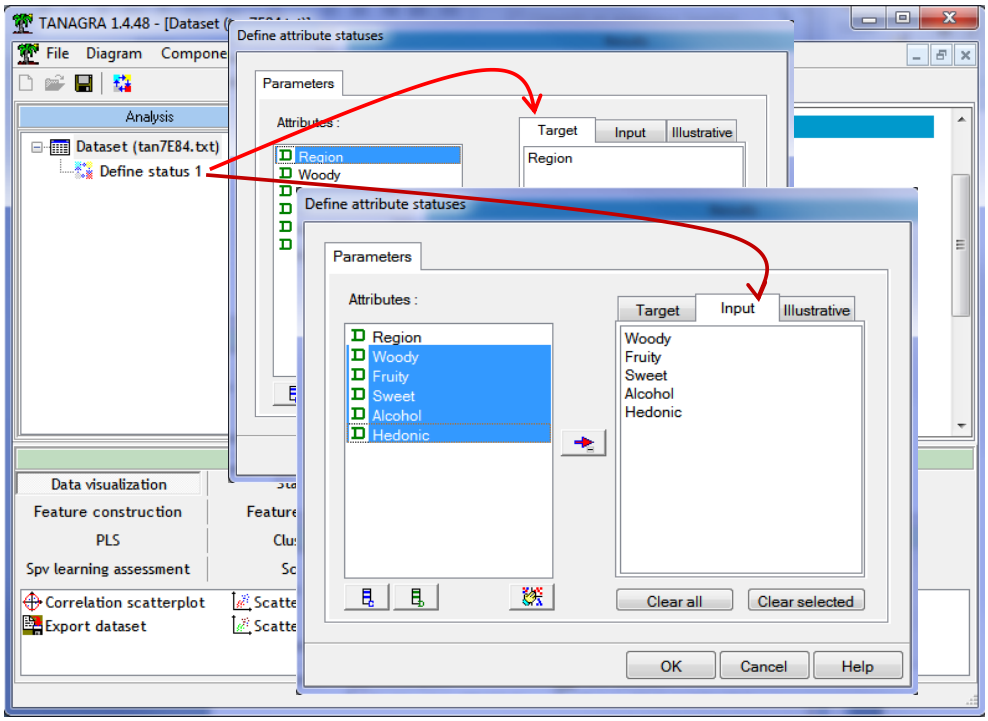
Nous le transmettons à Tanagra en actionnant le menu TANAGRA / EXECUTE TANAGRA installé dans l'onglet « Compléments » par la macro complémentaire **Tanagra.xla**[5].

Tanagra est automatiquement démarré et les données chargées.



2.3.2 Analyse discriminante pour variables qualitatives

Nous utilisons le composant DEFINE STATUS pour spécifier le rôle des variables : REGION est la cible (TARGET), le autres (WOODY…HEDONIC) sont les descripteurs (INPUT).



[5] Voir http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour Excel ; http://tutoriels-data-mining.blogspot.fr/2011/07/tanagra-addon-pour-openoffice-33.html pour Open et Libre Office.

Il n'est donc pas nécessaire d'effectuer explicitement la transformation en tableau de contingence pour l'analyse factorielle des correspondances, Tanagra s'en charge en interne.

Nous plaçons le composant DISCRIMINANT CORRESPONDENCE ANALYSIS (onglet FACTORIAL ANALYSIS) dans le diagramme. Nous actionnons le menu PARAMETERS.

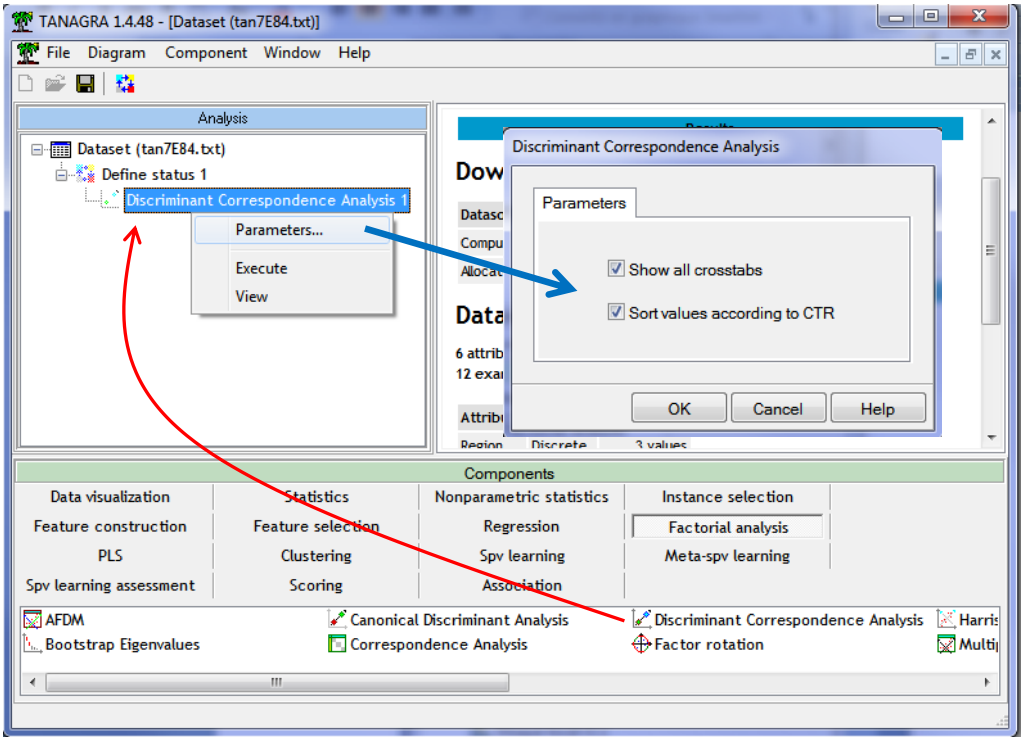
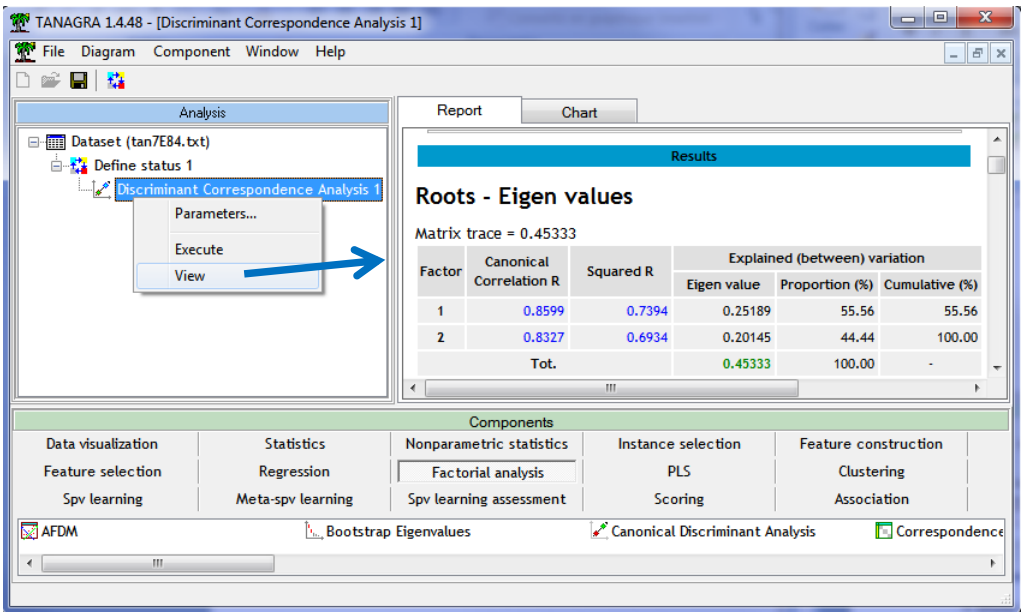


Figure 2 - Paramétrage de la méthode dans Tanagra

« Show All Crosstabs » affichera tous les tableaux croisés individuels triés selon le T de Tschuprow (voir Figure 1) ; « Sort Values according to CTR » triera les descripteurs selon leurs contributions (pour chaque variable, Tanagra s'appuie sur le maximum des contributions de ses modalités).

Nous validons et nous cliquons sur VIEW. Le tableau des valeurs propres s'affiche en premier.



2.3.3 Tableau des valeurs propres

Le tableau des valeurs propres indique la qualité de la description. Il reprend les résultats de l'AFC, mais les présente différemment.

Nombre de facteurs (F). Le nombre total de facteurs extrait est égal à **F = MIN (K-1, P- 1)**. En l'occurrence, comme le nombre de groupes sera quasiment toujours moins élevé que le nombre total d'indicatrices, nous aurons souvent (F = K-1) axes. Ce qui correspond aux résultats de l'analyse discriminante usuelle.

Trace de la matrice. La trace (Matrix Trace = 0.4533) correspond à l'inertie totale du nuage des modalités de l'AFC. Elle indique la quantité d'information que l'on peut modéliser dans la relation entre la cible REGION et les descripteurs. Elle sera décomposée sur les différents facteurs.

Valeurs propres associées aux facteurs. La valeur propre (λ) indique l'inertie (la variance) expliquée par l'appartenance aux groupes sur chaque axe. En les additionnant, nous retrouvons l'inertie expliquée par l'appartenance aux groupes dans l'espace complet c.-à-d. 0.25189 + 0.20145 = 0.45333 (matrix trace), d'où l'affichage des proportions individuelles et cumulées pour chaque axe.

Rapport de corrélation. Le rapport de corrélation (Squared R) est le ratio entre la variance expliquée par l'appartenance aux groupes (ex. λ_1 = 0.25189) et la variance totale de l'axe (elle est post-calculée par Tanagra après projection des individus, l'AFC ne la fournit pas). Il indique la qualité de discrimination des classes sur le facteur. Nous avons η^2_1=0.7394 c.-à-d. 73.94% de la variabilité des observations est expliquée par l'appartenance aux groupes sur le premier facteur. L'indicateur varie entre 0 (discrimination nulle, les sous-populations sont complètement mélangées) et 1 (discrimination parfaite, elles sont agglutinées sur les centres de classes qui sont distincts les uns des autres).

Notons un élément très important à cet stade : l'AFC assure la décroissance de la variance expliquée (λ) mais pas celle du rapport de corrélation (η^2) sur les facteurs, tout simplement parce que ce n'est pas son objectif. En passant par le tableau de contingence global, l'AFC ne travaille que sur les centres de classes pondérés par les effectifs. Il s'agit là d'une vraie divergence entre l'analyse des correspondances discriminante et l'analyse discriminante descriptive pour variables quantitatives.

Enfin, la **corrélation canonique** (Canonical Correlation R) est la racine carrée du rapport de corrélation ($\eta_1 = \sqrt{\eta_1^2} = \sqrt{0.7394} = 0.8599$).

2.3.4 Caractérisation des groupes

La caractérisation des groupes positionne les centres de classes sur les facteurs.

Group centroids on canonical variables

Row Characterization				Coord.		Contributions (%)		COS²	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos² 1	cos² 2
Loire	0.33333	0.49000	0.16333	0.65953	-0.23455	57.56	9.10	0.89 (0.89)	0.11 (1.00)
Beaujolais	0.33333	0.46500	0.15500	-0.55691	-0.39351	41.04	25.62	0.67 (0.67)	0.33 (1.00)
Rhone	0.33333	0.40500	0.13500	-0.10263	0.62807	1.39	65.27	0.03 (0.03)	0.97 (1.00)

Figure 3 - Coordonnées des centres de classes

Le premier oppose les vins de Loire à ceux du Beaujolais. Ils déterminent (**contributions** = 57.56 + 41.04) 98.61% de l'information portée par le facteur. Ils sont aussi très bien représentés puisque **cos²** = 89% (resp. 67%) de l'information véhiculée par Loire (resp. Beaujolais) est retranscrite sur cet axe.

Le second axe permet surtout de distinguer les vins du Rhône des deux premiers.

2.3.5 Distances entre centres de classes

Les distances entre centres de classes permettent de situer les proximités entre les groupes *sur l'ensemble des facteurs*.

Squared distance between group centroids

-	Loire	Beaujolais	Rhone
Loire	0.0000	1.5050	1.3250
Beaujolais	1.5050	0.0000	1.2500
Rhone	1.3250	1.2500	0.0000

Les trois types de vins sont à égale distance les uns des autres. Il est à prévoir qu'ils forment un triangle approximativement équilatéral dans le plan factoriel.

2.3.6 Structures canoniques

Les structures canoniques correspondent aux représentations des modalités colonnes du tableau de contingence – et donc des modalités des variables prédictives – dans le repère factoriel.

Canonical Structure

Row Characterization				Coord.		Contributions (%)		COS²	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos² 1	cos² 2
Fruity = C	0.06667	0.87500	0.05833	0.93447	-0.04210	23.1	0.1	1.00 (1.00)	0.00 (1.00)
Fruity = B	0.06667	0.12500	0.00833	-0.05112	0.34984	0.1	4.1	0.02 (0.02)	0.98 (1.00)
Fruity = A	0.06667	0.87500	0.05833	-0.88335	-0.30773	20.7	3.1	0.89 (0.89)	0.11 (1.00)
Woody = A	0.06667	0.87500	0.05833	0.93447	-0.04210	23.1	0.1	1.00 (1.00)	0.00 (1.00)
Woody = B	0.06667	0.12500	0.00833	-0.05112	0.34984	0.1	4.1	0.02 (0.02)	0.98 (1.00)
Woody = C	0.06667	0.87500	0.05833	-0.88335	-0.30773	20.7	3.1	0.89 (0.89)	0.11 (1.00)
Alcohol = A	0.08333	0.56000	0.04667	-0.14013	-0.73509	0.6	22.4	0.04 (0.04)	0.96 (1.00)
Alcohol = B	0.06667	0.12500	0.00833	0.32853	-0.13065	2.9	0.6	0.86 (0.86)	0.14 (1.00)
Alcohol = C	0.05000	2.00000	0.10000	-0.20448	1.39935	0.8	48.6	0.02 (0.02)	0.98 (1.00)
Sweet = B	0.06667	0.12500	0.00833	0.32853	-0.13065	2.9	0.6	0.86 (0.86)	0.14 (1.00)
Sweet = C	0.08333	0.08000	0.00667	0.04090	-0.27987	0.1	3.2	0.02 (0.02)	0.98 (1.00)
Sweet = A	0.05000	0.66667	0.03333	-0.50620	0.64065	5.1	10.2	0.38 (0.38)	0.62 (1.00)
Hedonic = A	0.05000	0.00000	0.00000	0.00000	0.00000	0.0	0.0	0.00 (0.00)	0.00 (0.00)
Hedonic = C	0.05000	0.00000	0.00000	0.00000	0.00000	0.0	0.0	0.00 (0.00)	0.00 (0.00)
Hedonic = B	0.05000	0.00000	0.00000	0.00000	0.00000	0.0	0.0	0.00 (0.00)	0.00 (0.00)
Hedonic = D	0.05000	0.00000	0.00000	0.00000	0.00000	0.0	0.0	0.00 (0.00)	0.00 (0.00)

Figure 4 - Tableau des structures canoniques

Tanagra met en surbrillance les modalités dont la contribution est supérieure à (100 / P) (moyenne non pondérée des contributions) et dont le cos² est supérieure à (1 / F).

Sans aller dans trop de détails, voici ce que nous pouvons lire dans ce tableau :

1. Le premier facteur, qui distingue les vins de Loire du Beaujolais, est caractérisé par l'opposition entre ('Fruity = A', 'Woody = C') et ('Fruity = C', 'Woody = A'). Ces modalités

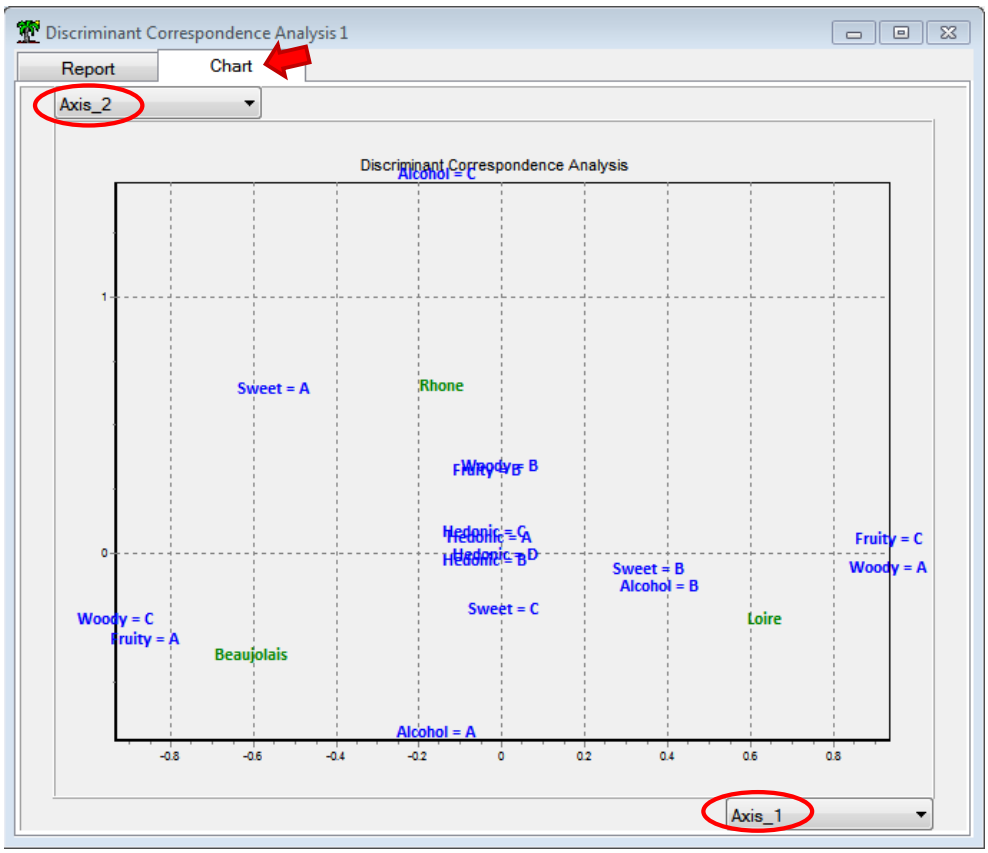
pèsent pour 87.6% dans la définition de l'axe (contributions = 20.7 + 20.7 + 23.1 + 23.1 = 87.6). On notera également que les modalités 'Alcohol = B' (cos^2 = 86%) et 'Sweet = B' (cos^2 = 86%) pèsent (relativement) peu mais sont très bien représentées.

2. Le second axe, qui distingue les vins du Rhône des deux autres, repose surtout sur l'opposition entre 'Alcohol = A' et 'Alcohol = C'.
3. Notons enfin que 'Hedonic' est absolument hors du coup dans la caractérisation des vins. On ne s'en étonnera pas. Nous avions constaté préalablement qu'elle n'était absolument pas liée avec la variable cible lors du calcul des tableaux de contingence individuels (Figure 1, V de Cramer = 0).

A ce stade commence le rôle de l'expert du domaine qui est à même de relier ces éléments numériques à la réalité œnologique. Pour ma part, mon expertise consiste principalement à ouvrir les bouteilles, avec le bruit du bouchon bien fort, et à engloutir le premier verre en prenant un air docte (ce n'est qu'un air) : « hum, il est plutôt gouleyant, je sens un goût de groseille… ».

2.3.7 Représentation simultanée

Que serait l'analyse factorielle sans les représentations graphiques ? En l'occurrence, en passant par le biais de l'AFC, nous disposons du positionnement relatif des modalités des descripteurs d'une part, de leur proximité (ou éloignement) avec les groupes définis par la variable cible d'autre part.



Nous avons eu recours au « jittering »[6] pour distinguer les points superposés. De fait, les coordonnées ne correspondent par exactement aux valeurs recensées dans les tableaux.

[6] Cf. http://www.statisticalanalysisconsulting.com/scatterplots-dealing-with-overplotting/

2.3.8 Fonctions canoniques – Projection et classement d'un individu supplémentaire

Projection. Les fonctions canoniques permettent de projeter des individus non étiquetés dans l'espace factoriel. Les coefficients s'appliquent sur les indicatrices des variables.

Canonical Coefficients

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

Attribute.Value	Factor 1	Factor 2
Woody = A	0.3723831	-0.0187617
Woody = B	-0.0203715	0.1558900
Woody = C	-0.3520116	-0.1371283
Fruity = C	0.3723831	-0.0187617
Fruity = B	-0.0203715	0.1558900
Fruity = A	-0.3520116	-0.1371283
Sweet = B	0.1309182	-0.0582172
Sweet = C	0.0162972	-0.1247120
Sweet = A	-0.2017196	0.2854764
Alcohol = A	-0.0558430	-0.3275623
Alcohol = B	0.1309182	-0.0582172
Alcohol = C	-0.0814859	0.6235602
Hedonic = A	0.0000000	0.0000000
Hedonic = C	0.0000000	0.0000000
Hedonic = B	0.0000000	0.0000000
Hedonic = D	0.0000000	0.0000000

Prenons l'individu supplémentaire (W ?) de notre document de référence (Abdi, 2007 ; page 3). Nous disposons de la description suivante :

Woody	Fruity	Sweet	Alcohol	Hedonic
A	**C**	**B**	**B**	**A**

Nous reproduisons la feuille de calcul Excel (les coefficients ont été arrondis à 3 décimales) en mettant en évidence les coefficients activés par les indicatrices non nulles de l'individu.

Attribute.Value	Factor 1	Factor 2	Dummy data
Woody = A	0.372	-0.019	1
Woody = B	-0.020	0.156	0
Woody = C	-0.352	-0.137	0
Fruity = C	0.372	-0.019	1
Fruity = B	-0.020	0.156	0
Fruity = A	-0.352	-0.137	0
Sweet = B	0.131	-0.058	1
Sweet = C	0.016	-0.125	0
Sweet = A	-0.202	0.285	0
Alcohol = A	-0.056	-0.328	0
Alcohol = B	0.131	-0.058	1
Alcohol = C	-0.081	0.624	0
Hedonic = A	0.000	0.000	1
Hedonic = C	0.000	0.000	0
Hedonic = B	0.000	0.000	0
Hedonic = D	0.000	0.000	0

Coord	1.01	-0.15

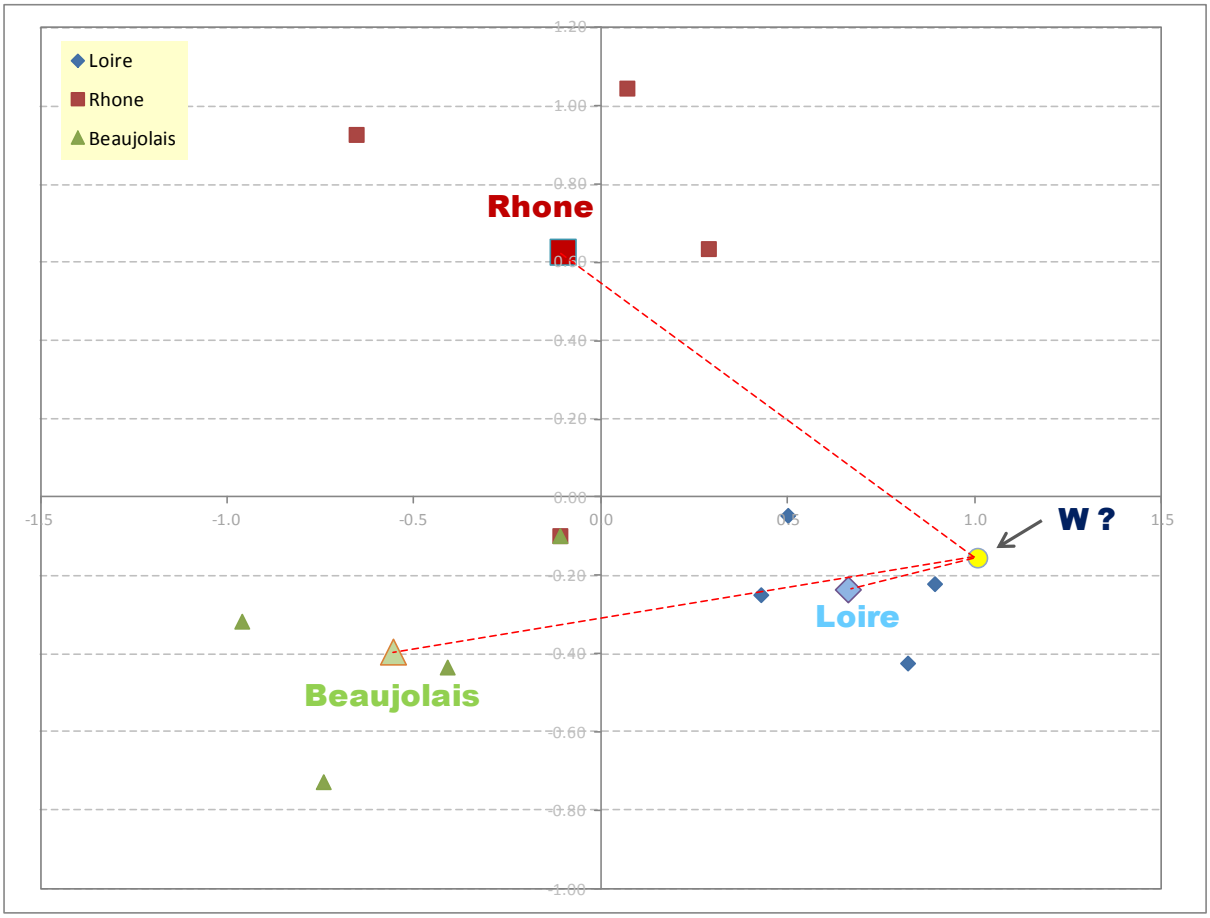
Voyons le détail pour le premier facteur (en masquant les indicatrices nulles) :

$$F_1(W) = 0.372 \times 1 + 0.372 \times 1 + 0.131 \times 1 + 0.131 \times 1 + 0.000 \times 1 = 1.01$$

Pour le second facteur, nous obtenons :

$$F_2(W) = -0.019 \times 1 - 0.019 \times 1 - 0.058 \times 1 - 0.058 \times 1 + 0.000 \times 1 = -0.15$$

Placé dans le repère factoriel avec les autres points de l'échantillon, en mettant en évidence les centres de classes, nous constatons qu'il se rapproche plutôt des vins de Loire.



Classement – Distance aux centres de classes. Comment confirmer cette impression visuelle par le calcul ? A l'instar de l'analyse discriminante usuelle, nous pouvons affecter l'individu supplémentaire à la classe dont le barycentre lui est le plus proche.

Nous calculons la distance euclidienne $[d(W,k)]$ de W avec chaque centre de classe (Figure 3)

Classes	Distance euclidienne (carré)
d²(W,Loire)	**(1.01 – 0.65953)² + (-0.15 – (-0.23455))² = 0.127**
d²(W,Rhone)	(1.01 – (-0.10263))² + (-0.15 – 0.62807)² = 1.842
d²(W,Beaujolais)	(1.01 – (-0.55691))² + (-0.15 – (-0.39351))² =2.502

Indéniablement, notre breuvage inconnu appartient aux vins de Loire.

Classement (2) – Distance généralisée. Lorsque les classes ne sont pas équilibrées c.-à-d. nous n'avons pas le même effectif pour chaque modalité de la variable cible, il est plus judicieux d'utiliser la distance généralisée[7] pour classer un individu supplémentaire

$$D^2(W,k) = -2 \times \ln \pi_k + d^2(W,k)$$

Où π_k représente la proportion des observations correspondant la modalité *k* de la variable cible.

En ce qui nous concerne, cela ne change rien puisque les groupes sont équilibrés, $\pi_k = \frac{1}{3}$, $\forall\, k$

2.3.9 Tableaux croisés

L'option « Show All Crosstabs » ayant été sélectionné lors du paramétrage du composant (Figure 2), Tanagra fournit tous les tableaux croisés entre la cible et les descripteurs. Nous obtenons un affichage équivalent à la (Figure 1) avec, de surcroît, d'autres indicateurs d'évaluation des associations (T de Tschuprow, ϕ^2, etc.). Nous n'affichons que les deux premiers tableaux ici.

Crosstabs between target and inputs

Row (Y)	Column (X)	Statistical indicator		Cross-tab				
		Stat	Value		A	B	C	Sum
		d.f.	4	Loire	2	2	0	4
		Tschuprow's t	0.622495	Rhone	0	1	3	4
		Cramer's v	0.622495	Beaujolais	3	1	0	4
Region	Alcohol	Phi²	0.775000	Sum	5	4	3	12
		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)					

Row (Y)	Column (X)	Statistical indicator		Cross-tab				
		Stat	Value		C	B	A	Sum
		d.f.	4	Loire	3	1	0	4
		Tschuprow's t	0.559017	Rhone	1	2	1	4
		Cramer's v	0.559017	Beaujolais	0	1	3	4
Region	Fruity	Phi²	0.625000	Sum	4	4	4	12
		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)					

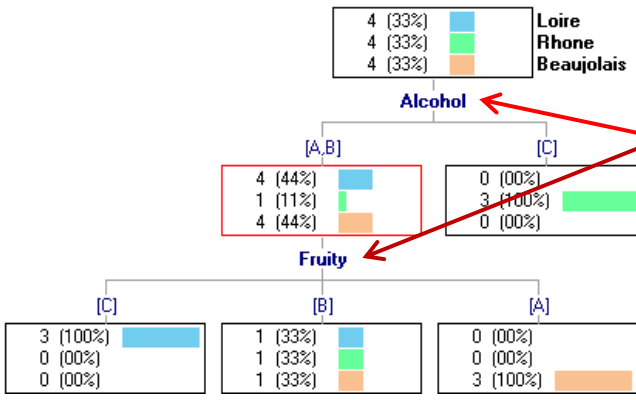
Les tableaux sont triés par ordre décroissant du T de Tschuprow. Nous remarquons qu'aucune liaison n'est significative à 5%, essentiellement parce que l'effectif (n = 12) est très faible. Cela ne nous a pas empêché de faire ressortir des relations « intéressantes » entre les variables.

2.3.10 Rapprochement avec d'autres méthodes descriptives – Les arbres de décision

Aucune méthode ne détient la vérité absolue en statistique exploratoire. Chacune à leur manière explore telle ou telle facette des informations véhiculées par les données. Il nous appartient de cerner les atouts et faiblesses des approches que nous utilisons.

[7] http://v8doc.sas.com/sashtml/stat/chap25/sect17.htm

Justement, pour confirmer notre analyse ci-dessus, j'ai construit un arbre de décision à l'aide du logiciel SIPINA (http://sipina.over-blog.fr/) pour décrire REGION à partir des descripteurs. J'ai mis en parallèle l'arbre obtenu et le tableau des structures canoniques (Figure 4).



	Coord.	
	coord 1	coord 2
Loire	0.65953	-0.23455
Rhone	-0.10263	0.62807
Beaujolais	-0.55691	-0.39351

Row	Coord.		Contributions (%)	
Values	coord 1	coord 2	ctr 1	ctr 2
Fruity = C	**0.93447**	-0.0421	23.1	0.1
Fruity = B	-0.05112	0.34984	0.1	4.1
Fruity = A	**-0.88335**	-0.30773	20.7	3.1
Woody = A	**0.93447**	-0.0421	23.1	0.1
Woody = B	-0.05112	0.34984	0.1	4.1
Woody = C	**-0.88335**	-0.30773	20.7	3.1
Alcohol = A	-0.14013	**-0.73509**	0.6	22.4
Alcohol = B	0.32853	-0.13065	2.9	0.6
Alcohol = C	-0.20448	**1.39935**	0.8	48.6
Sweet = B	0.32853	-0.13065	2.9	0.6
Sweet = C	0.0409	-0.27987	0.1	3.2
Sweet = A	-0.5062	**0.64065**	5.1	10.2
Hedonic = A	0	0	0	0
Hedonic = C	0	0	0	0
Hedonic = B	0	0	0	0
Hedonic = D	0	0	0	0

Les résultats concordent mais, de par son processus inhérent de sélection de variables, l'arbre masque certaines caractéristiques. Par exemple, le rôle de WOODY est complètement occulté. Certes, en explorant en détail l'arbre, nous pouvons retrouver ces informations[8] mais, d'une part il faut disposer de l'outil adéquat pour le faire (tous les logiciels de construction d'arbre ne proposent pas de fonctionnalités interactives) et, d'autre part, explorer en détail chaque nœud peut s'avérer rapidement très fastidieux lorsque l'arbre en comporte un grand nombre. ***Dans le contexte descriptif***, l'analyse des correspondances discriminante propose des résultats plus complets.

3 Programmation de l'approche sous R

Pour me faire une idée de l'intérêt de l'approche, je l'avais initialement implémenté sous R. Le principal enjeu est dans la construction du tableau de contingence global croisant les indicatrices des descripteurs avec la cible. Il fallait aussi le faire de la manière la plus générique possible pour que la transposition à d'autres jeux de données soit relativement facile.

3.1 Importation des données

Nous utilisons le package « xlsx »[9] pour lire directement le fichier « **french_win_dca.xls** ».

[8] Voir http://tutoriels-data-mining.blogspot.fr/2008/03/analyse-interactive-avec-sipina.html

[9] http://cran.r-project.org/web/packages/xlsx/index.html

```
#importing the data file
library(xlsx)
wine <- read.xlsx(file="french_wine_dca.xls",sheetIndex=1,header=T)
print(summary(wine))
```

La commande **summary()** permet de vérifier l'intégrité des données.

```
> 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
```

3.2 Construction du tableau de contingence pour l'AFC

Nous procédons en plusieurs étapes : (1) nous isolons les descripteurs ; (2) nous programmons une fonction permettant de croiser un descripteur avec la cible ; (3) nous appliquons cette fonction sur chaque descripteur pour obtenir les tableaux de contingence individuels ; (4) que nous accolons pour former le tableau global.

```
#(1) select the predictive attributes
descriptors <- subset(wine,select=-1)
print(summary(descriptors))

#(2) function for building crosstabs from the target attribute
#and each predictive attributes
cross.tab <- function(x,ref){
 m <- table(ref,x)
 return(m)
}

#(3) apply the function cross.tab on each predictive attribute
dataset <- lapply(descriptors,cross.tab,ref=wine$Region)

#(4) create the matrix for the correspondence analysis
#from the crosstabs
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)
```

Nous obtenons le tableau, en veillant à ce que les étiquettes de lignes et de colonnes correspondent bien aux modalités concernées.

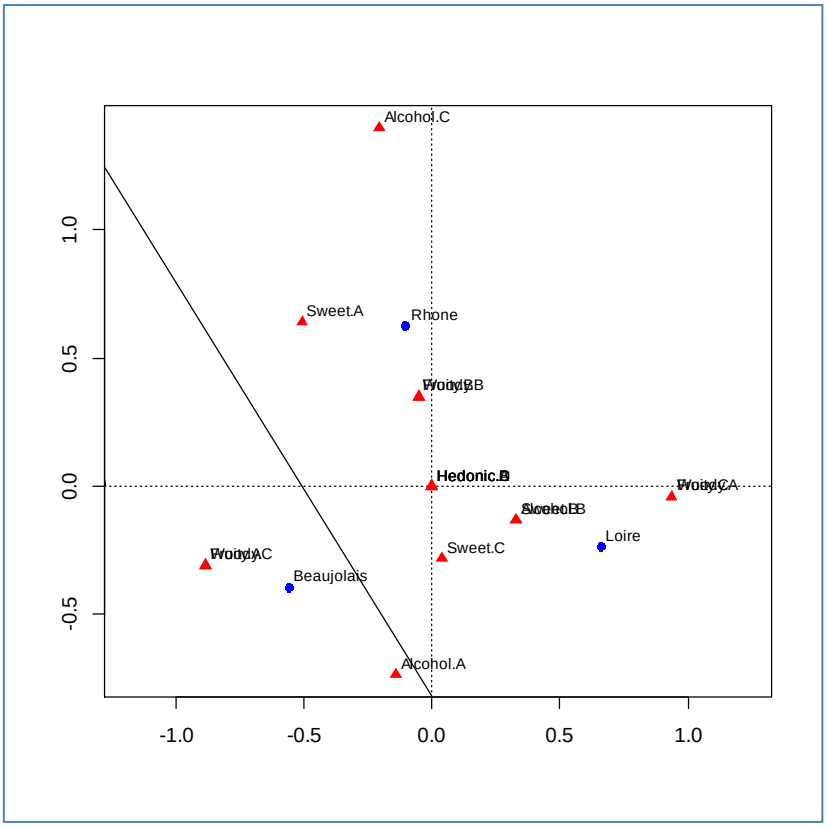
```
> print(matrix.ca)
           Woody.A Woody.B Woody.C Fruity.A Fruity.B Fruity.C Sweet.A Sweet.B
Beaujolais       0       1       3        3        1        0       1       1
Loire            3       1       0        0        1        3       0       2
Rhone            1       2       1        1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

3.3 Analyse des correspondances

Nous utilisons le package 'ca' pour l'analyse factorielle des correspondances[10].

```
library(ca)
fit <- ca(matrix.ca,nd=2)
print(fit)
#graphical representation
plot(fit)
```

La représentation simultanée permet de situer les modalités dans l'espace factoriel.



[10] http://cran.r-project.org/web/packages/ca/index.html

Les résultats concordent en tout point avec ceux de Tanagra.

```
#row coordinates (consistent with Tanagra)
row.coord <- cbind(fit$rowcoord[,1]*fit$sv[1],fit$rowcoord[,2]*fit$sv[2])
print(row.coord)
#column coordinates (consistent with Tanagra)
col.coord <- cbind(fit$colcoord[,1]*fit$sv[1],fit$colcoord[,2]*fit$sv[2])
print(col.coord)
```

Cela est confirmé par l'affichage des coordonnées factorielles des modalités lignes (variable cible) et colonnes (descripteurs) (voir le tableau des structures canoniques pour Tanagra - Figure 4).

```
> #row coordinates (consistent with 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)
                [,1]       [,2]
Beaujolais -0.5569077 -0.3935147
Loire       0.6595342 -0.2345520
Rhone      -0.1026265  0.6280667
>
> #column coordinates (consistent with Tanagra)
> col.coord <- cbind(fit$colcoord[,1]*fit$sv[1],fit$colcoord[,2]*fit$sv[2])
> rownames(col.coord) <- colnames(matrix.ca)
> print(col.coord)
                [,1]        [,2]
Woody.A     0.93446630 -0.04210386
Woody.B    -0.05112059  0.34983808
Woody.C    -0.88334571 -0.30773422
Fruity.A   -0.88334571 -0.30773422
Fruity.B   -0.05112059  0.34983808
Fruity.C    0.93446630 -0.04210386
Sweet.A    -0.50619940  0.64064720
Sweet.B     0.32852896 -0.13064732
Sweet.C     0.04089647 -0.27987047
Alcohol.A -0.14013376 -0.73509355
Alcohol.B  0.32852896 -0.13064732
Alcohol.C -0.20448234  1.39935234
Hedonic.A  0.00000000  0.00000000
Hedonic.B  0.00000000  0.00000000
Hedonic.C  0.00000000  0.00000000
Hedonic.D  0.00000000  0.00000000
```

Les variables ne sont pas triées selon les contributions dans notre programme. Mais ça doit pouvoir se faire relativement facilement.

4 Conclusion

L'analyse des correspondances discriminante permet de réaliser une analyse discriminante sur des descripteurs catégoriels. Pour les habitués de l'analyse factorielle, la lecture et l'interprétation de résultats ne pose pas de problèmes particuliers. Ce n'est pas la moindre de ses qualités.

Dans un avenir proche, j'en dériverai certainement un composant dédié à la prédiction. Je lui trouve des propriétés intéressantes dans ce contexte : il ne semble perturbé ni par les descripteurs redondants (j'ai dupliqué des descripteurs, les résultats sont certes légèrement modifiés mais restent de la même teneur), ni par les descripteurs non pertinents ('hedonic' dans notre exemple).