1 Topic

Discriminant Correspondence Analysis.

The aim of the canonical discriminant analysis is to explain the belonging to pre-defined groups of instances of a dataset. The groups are specified by a dependent categorical variable (class attribute, response variable); the explanatory variables (descriptors, predictors, independent variables) are all continuous. So, we obtain a small number of latent variables which enable to distinguish as far as possible the groups. These new features, called factors, are linear combinations of the initial descriptors. The process is a valuable dimensionality reduction technique. But its main drawback is that it cannot be directly applied when the descriptors are discrete. Even if the calculations are possible if we recode the variables using dummy variables for instance, the interpretation of the results - which is one of the main goals of the canonical discriminant analysis - is not really obvious.

In this tutorial, we present a variant of the discriminant analysis which is applicable to discrete descriptors due to Hervé Abdi (2007)[1]. The approach is based on a transformation of the raw dataset in a kind of contingency table. The rows of the table correspond to the values of the target attribute; the columns are the indicators associated to the predictors' values. Thus, the author suggests to use a correspondence analysis, on the one hand, in order to distinguish the groups, and on the other hand, to detect the relevant relationships between the values of the target attribute and those of the explanatory variables. The author called its approach "discriminant correspondence analysis" because it uses a correspondence analysis framework to solve a discriminant analysis problem.

In what follows, we detail the use of the discriminant correspondence analysis with Tanagra 1.4.48. We use the example described in the Hervé Abdi's paper. The goal is to explain the origin of 12 wines (3 possible regions) using 5 descriptors related to characteristics assessed by professional tasters. In a second part (section 3), we reproduce all the calculations with a program written for R.

2 Wines dataset

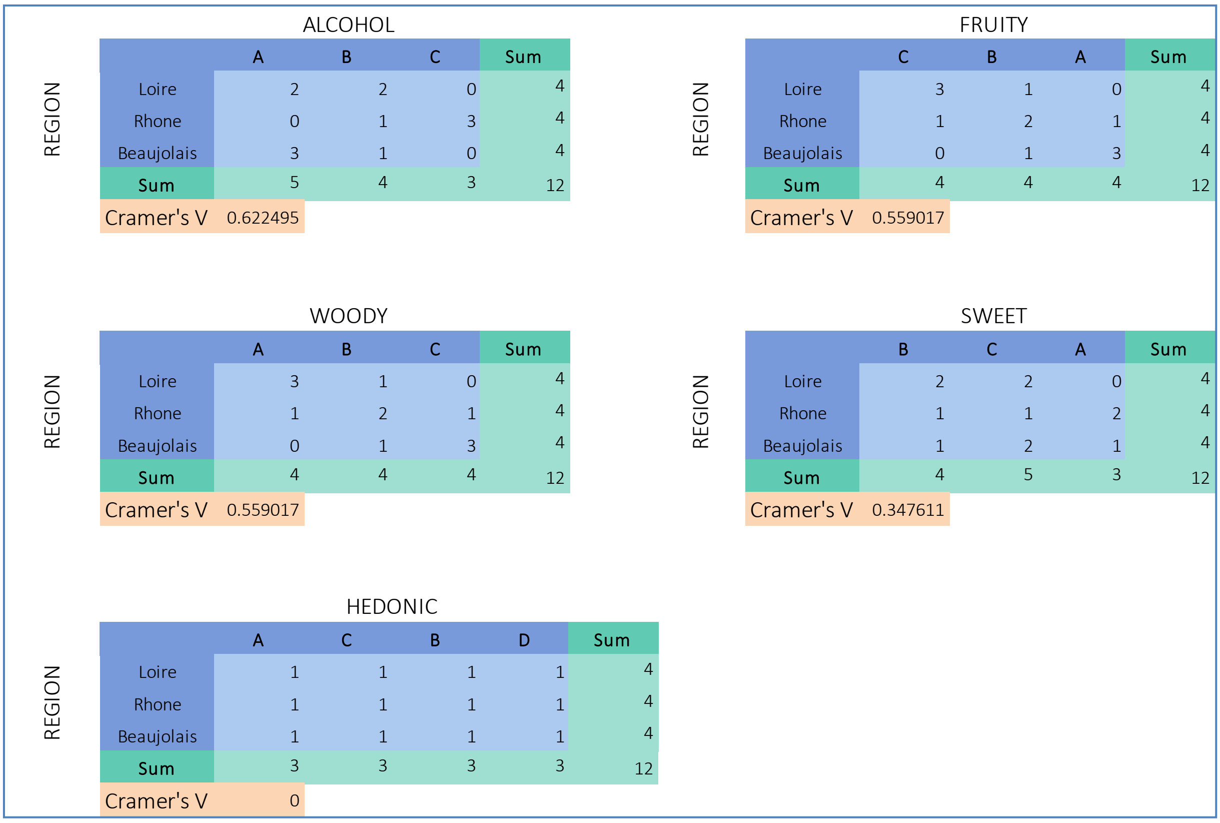
2.1 Dataset characteristics

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

There are **n = 12** instances (wines) from **K = 3** regions (Loire, Rhône, Beaujolais) and **p = 5** discrete descriptors (Woody, Fruity, Sweet, Alcohol, Hedonic) into the dataset. The goal is to determine the wine characteristics according to the regions.

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

A first solution consists in to use a bivariate analysis by calculating the crosstabs of the target variable with each descriptor.



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				

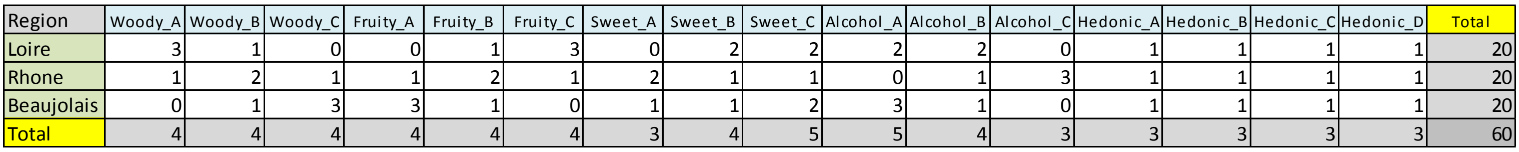
Figure 1 – Cross tabs of the target attribute (REGION) with each descriptor

No descriptor is significantly related to the class attribute according the chi-squared test of independence at the 5% level. This is not surprising considering the size of the dataset (n = 12).

However, some associations with REGION seem interesting according to the Cramer's V criterion (which measures the structure of the relationship without taking account the sample size): ALCOHOL (0.62); FRUITY (0.56); WOODY (0.56). Thus, we can expect to obtain interesting results when we perform a multivariate approach.

2.2 Data transformation for correspondence analysis

Hervé Abdi proposes to transform the raw dataset in a kind of crosstab in order to solve the discriminant analysis problem with the correspondence analysis technique. The idea is to concatenate the individual crosstabs in a unique overall table. So, we obtain a table with **K = 3** rows and **P = 16** columns. P is the number of explanatory variables values (**P =** 3 + 3 + 3 + 3 + 4 = **16**).



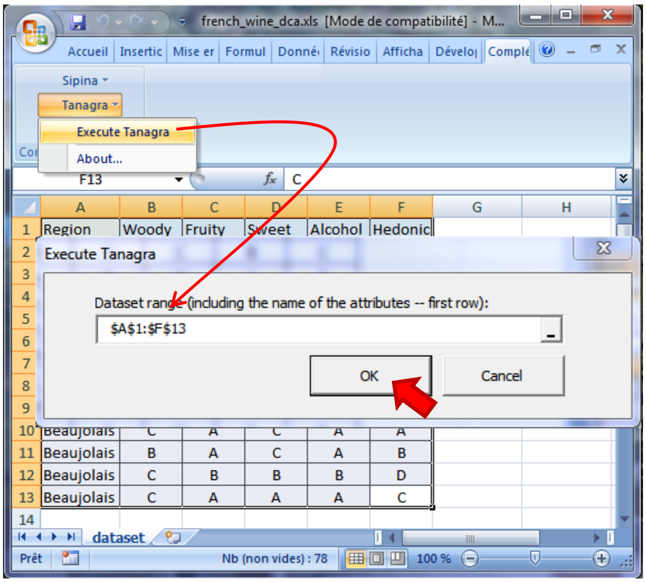
Region	Woody_A	Woody_B	Woody_C	Fruity_A	Fruity_B	Fruity_C	Sweet_A	Sweet_B	Sweet_C	Alcohol_A	Alcohol_B	Alcohol_C	Hedonic_A	Hedonic_B	Hedonic_C	Hedonic_D	Total
Loire	3	1	0	0	1	3	0	2	2	2	2	0	1	1	1	1	20
Rhone	1	2	1	1	2	1	2	1	1	0	1	3	1	1	1	1	20
Beaujolais	0	1	3	3	1	0	1	1	2	3	1	0	1	1	1	1	20
Total	4	4	4	4	4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

There is clearly a duplication of the data. The grand total in this new representation is N = n x p = 12 x 5 = 60 'observations' (the quotes are important). Therefore, it is not really a contingency table in the strict sense. But, on the other hand, we have a representation that we can use to: assess the similarities and differences between the columns considering the rows (the values of the target variable); perform the same kind of study for rows considering the columns; understand the attractions and repulsions between the rows values and the columns values. This is the purpose of the correspondence analysis.

2.3 Performing the discriminant correspondence analysis with Tanagra

2.3.1 Importing the dataset

We load the « **french_wine_dca.xls** » data file into the Excel spreadsheet.



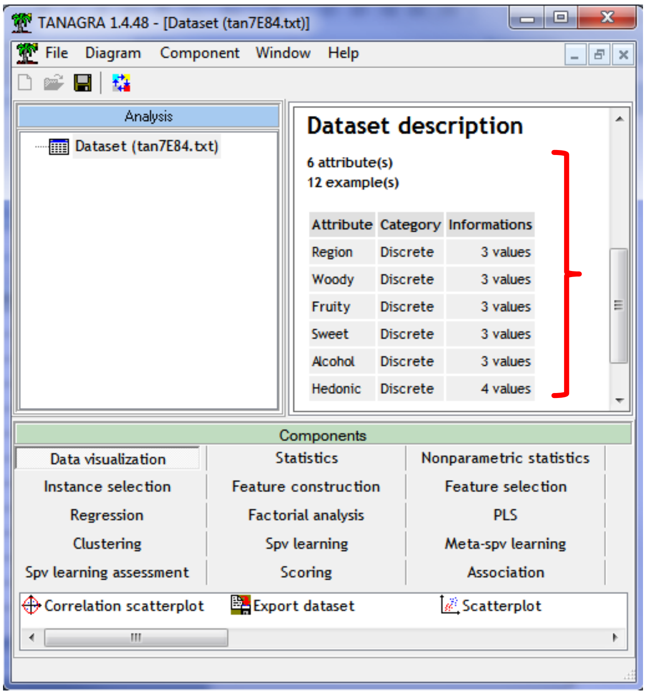
Execute Tanagra

Dataset range (including the name of the attributes -- first row): $A$1:$F$13

OK | Cancel

	A	B	C	D	E	F
1	Region	Woody	Fruity	Sweet	Alcohol	Hedonic
10	Beaujolais	C	A	C	A	A
11	Beaujolais	B	A	C	A	B
12	Beaujolais	C	B	B	B	D
13	Beaujolais	C	A	A	A	C

We select the data cells range and we click on the TANAGRA / EXECUTE TANAGRA menu, installed by the **Tanagra.xla** add-in[2]. Tanagra is launched and the dataset is automatically loaded.



Dataset description

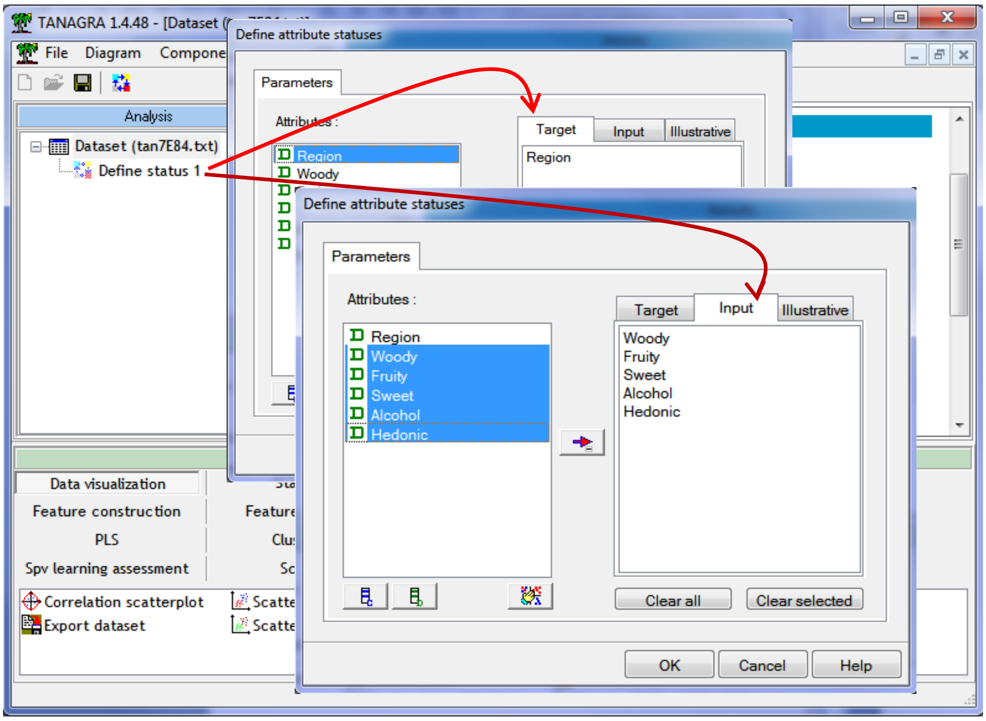
6 attribute(s)
12 example(s)

Attribute	Category	Informations
Region	Discrete	3 values
Woody	Discrete	3 values
Fruity	Discrete	3 values
Sweet	Discrete	3 values
Alcohol	Discrete	3 values
Hedonic	Discrete	4 values

[2] See http://data-mining-tutorials.blogspot.fr/2010/08/tanagra-add-in-for-office-2007-and.html

2.3.2 Discriminant analysis for discrete descriptors

We use the DEFINE STATUS component to define the role of the variables: REGION is the response variable (TARGET); the others (WOODY…HEDONIC) are the descriptors (INPUT).

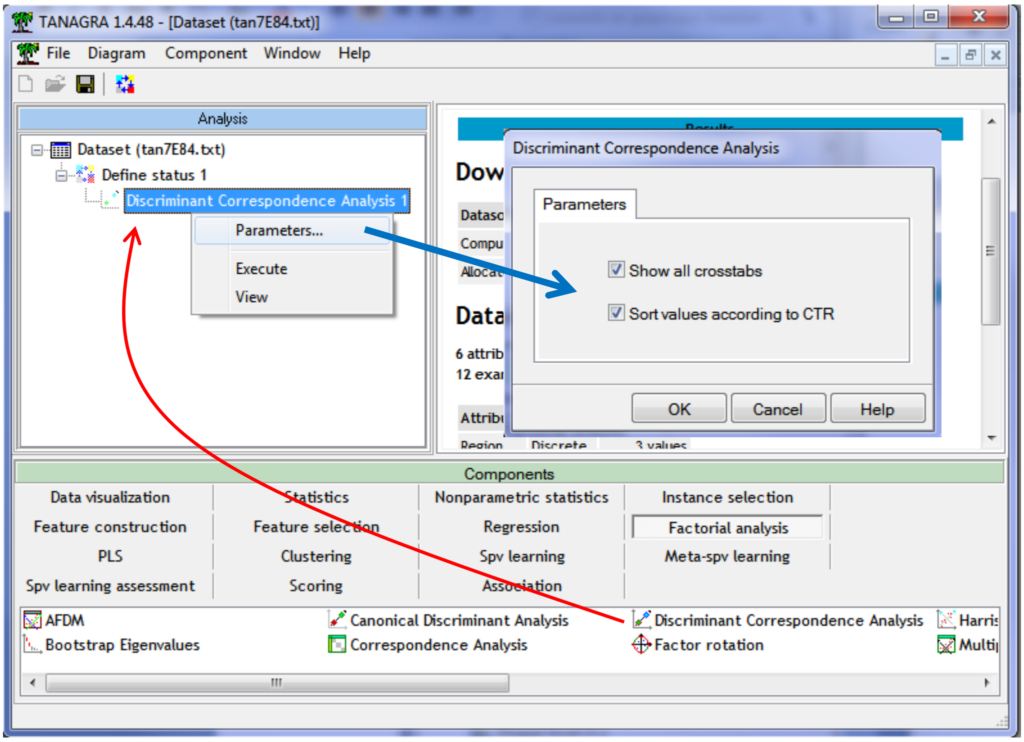


Attributes : Region, Woody, Fruity, Sweet, Alcohol, Hedonic

Target: Region

Input: Woody, Fruity, Sweet, Alcohol, Hedonic

It is not necessary to explicitly make the transformation of the raw dataset in a crosstab for the factorial correspondence analysis, Tanagra does it internally. We add the DISCRIMINANT CORRESPONDENCE ANALYSIS (FACTORIAL ANALYSIS tab) into the diagram.



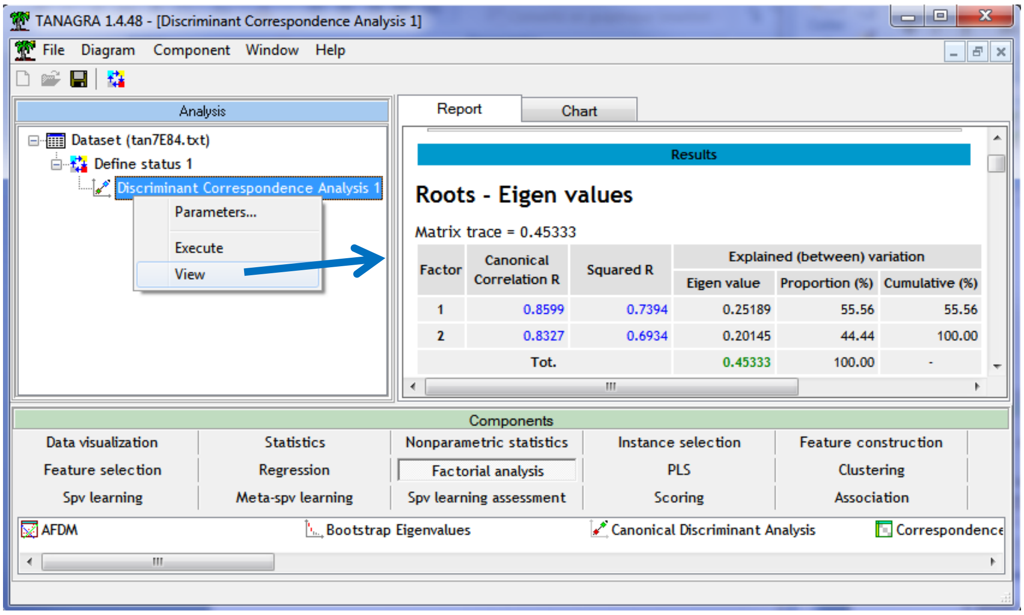
☑ Show all crosstabs

☑ Sort values according to CTR

Figure 2 – Settings of Discriminant Correspondence Analysis

We click on the PARAMETERS menu in order to specify the settings. "Show All Crosstabs" displays all the individual crosstabs between the response attribute and the descriptors (See Figure 1); "Sort Values according to CTR" sorts the variables according to their contributions to the factors (for each descriptor, Tanagra uses the maximum of the contributions of its modalities).

We validate these settings and we click on the VIEW menu. First we have the table of eigenvalues.



Results

Roots - Eigen values

Matrix trace = 0.45333

Factor	Canonical Correlation R	Squared R	Explained (between) variation		
			Eigen value	Proportion (%)	Cumulative (%)
1	0.8599	0.7394	0.25189	55.56	55.56
2	0.8327	0.6934	0.20145	44.44	100.00
	Tot.		0.45333	100.00	-

2.3.3 Table of the eigenvalues

This table shows the overall quality of the process. The output is based on the results of the correspondence analysis, but they are presented differently.

Number of dimensions (F). The maximum number of factors that we can reach is **F** = **MIN (K-1, P- 1)**. Because the number of groups (K) is usually lower than the number of variables values (P), we have (F = K-1) factors in most cases.

Matrix trace. The trace (Matrix Trace = 0.4533) is to the total inertia in the correspondence analysis. It indicates the amount of information that can be modeled in the relationship between the target REGION and descriptors. It will be decomposed on the different factors.

Eigenvalues. The eigenvalues (λ) indicates the inertia explained by each factor. By adding up them we obtain the total inertia i.e. 0.25189 + 0.20145 = 0.45333 (matrix trace). The table shows the same information with the percentage of inertia explained (and the cumulative percentage) by each factor.

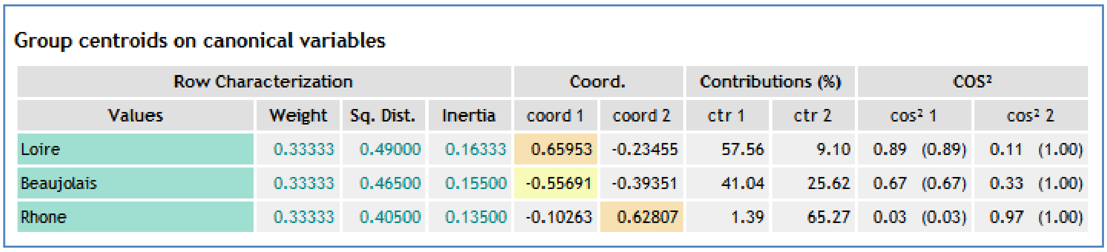
Correlation ratio. The correlation ratio (Squared R) is the ratio between the variance explained by the belonging to the groups (ex. λ_1 = 0.25189) and the total variance of the factor (it is computed from the scores of the individuals on the axis, the correspondence analysis does not provide this value). Thus, for the first factor we obtain η^2_1=0.7394 i.e. 73.94% of the dispersion is explained by group membership.

We observe that if the correspondence analysis insures the decreasing of the eigenvalues (explained variance), we do not have the same phenomenon for the correlation ratio (η^2), because this is not the purpose of the correspondence analysis performed on the overall crosstab.

Last, the **canonical correlation** is (Canonical Correlation R) is the square root of the correlation ratio (η_1 = $\sqrt{\eta_1^2} = \sqrt{0.7394} = 0.8599$).

2.3.4 Characterization of groups

This table shows the mean of groups for each factor.



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)

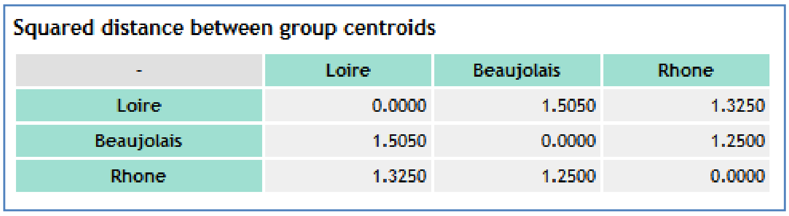
Figure 3 – Group centroids – Row coordinates

The first factor enables to distinguish the Loire wine and the Beaujolais. They determine a large part of the explained inertia (**contributions** = 57.56 + 41.04 = 98.61%). These modalities are also well represented i.e. the factor captures well the information they convey (**cos²** = 89% and 67%).

The second factor enables to distinguish the Rhone wine from the others.

2.3.5 Distance between group centroids

The distances between centroids enable to evaluate the proximities between the groups by considering all the factors. We observe here that the wines are in equal distances from each other. We can expect that the centroids form an equilateral triangle into the graphical representation.



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

2.3.6 Canonical structure

The canonical structure table shows the coordinates and the influence of the descriptors values for the determination of the factors. Thus, it allows also to characterize / explain the differences between the modalities of the target attribute.

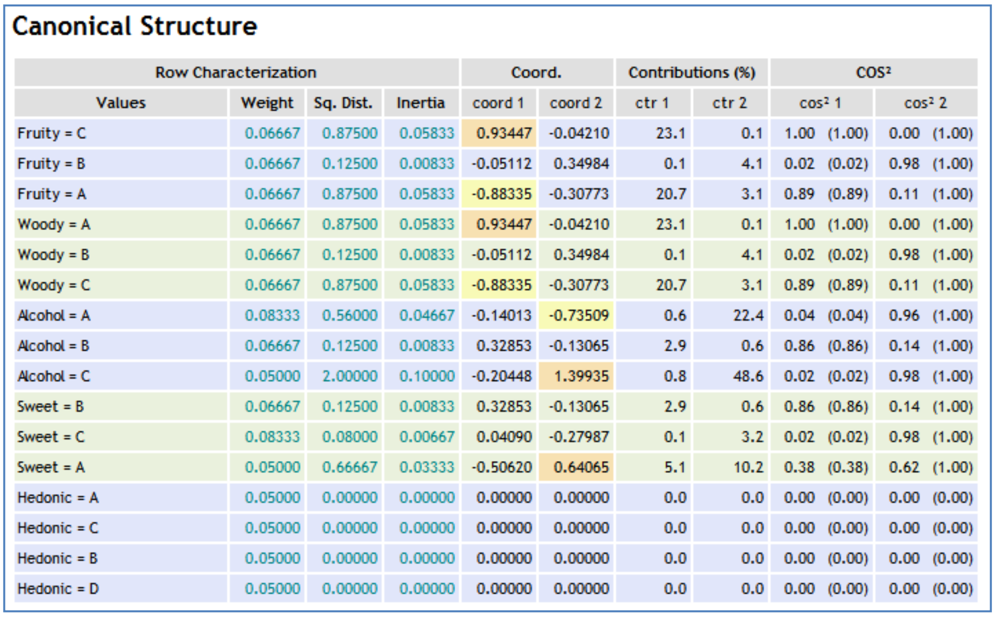
Tanagra highlights the coordinates of the modalities for which: the contribution is higher than (100/P, mean of the contributions); and the squared cosines is higher than (1/F).

Without going into too much detail, we observe into this table (Figure 4):

1. As seen previously (Figure 3), the first factor enables to separate the Loire wine from the Beaujolais. This is due to the opposition between ('Fruity = A', 'Woody = C') and ('Fruity = C', 'Woody = A'). These values influence 87.6% of the determination of the factor (contributions = 20.7 + 20.7 + 23.1 + 23.1 = 87.6). We observe also that 'Alcohol = B' (cos² = 86%) and 'Sweet = B' (cos² = 86%) are well represented.
2. The second factor which enables to distinguish the Rhone wine from the others relies mainly on the opposition between 'Alcohol = A' and 'Alcohol = C'.

3. Last, we note that Hedonic is not relevant for the explanation of the group membership. This is not surprising, we have already observed that its association to Region is null in the individual crosstab (Figure 1, Cramer's V = 0).

At this point begins the role of the domain expert who is able to connect these numerical results to the reality of the field.



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 – Canonical structure

2.3.7 Simultaneous plot

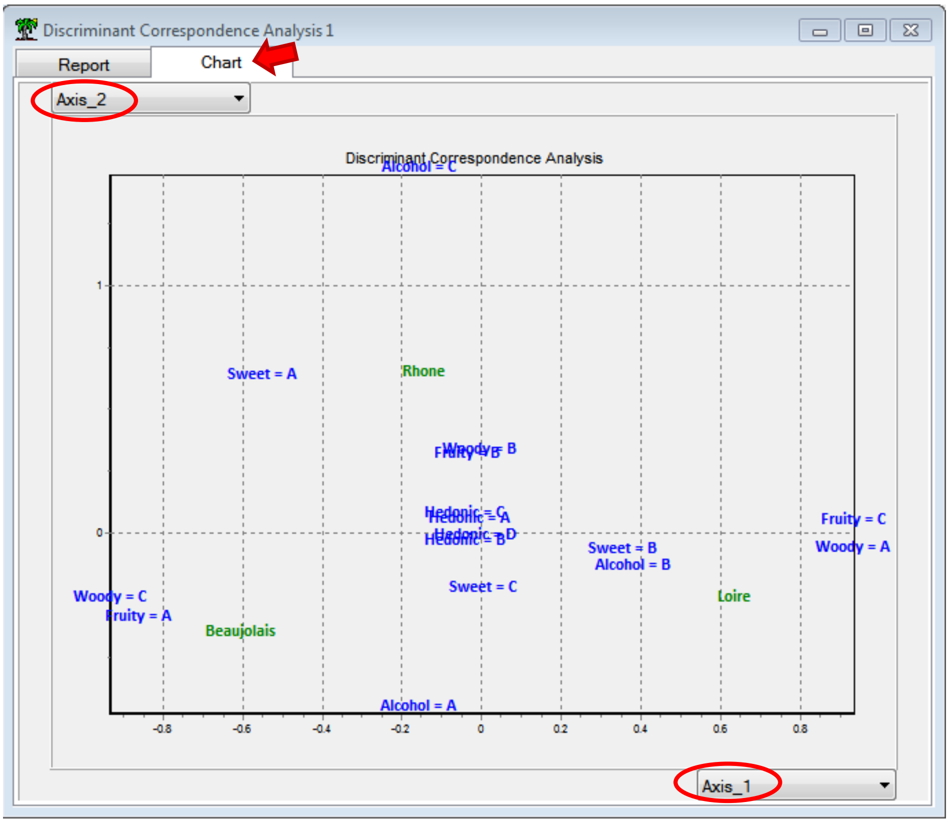


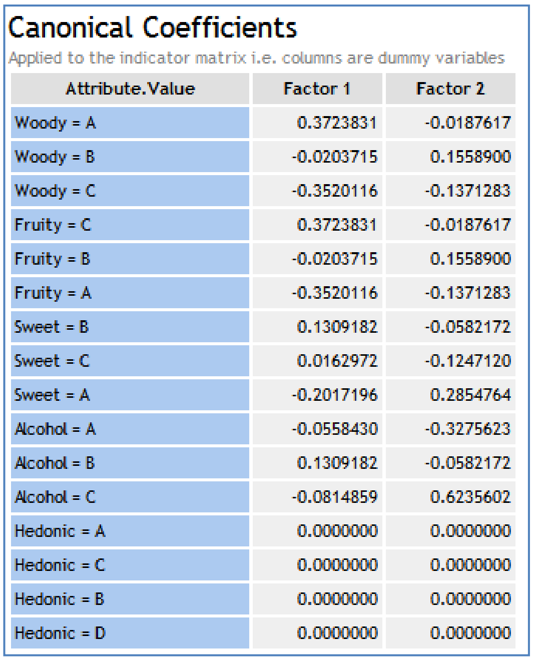
Figure 5 - Simultaneous plot of the rows and columns coordinates from the correspondence analysis

The simultaneous plot from the correspondence analysis enables to visualize the association between the values of the descriptors and those of the target variable. In our graphical representation, we use jittering[3] to overcome the overplotting problem. So we have not exactly the same values in the table (Figure 4) and in the chart (Figure 5).

2.3.8 Canonical coefficients

Calculating the scores of instances. The canonical coefficients applied on the indicator matrix enable to compute the scores of instances. Because all the coefficients are applied on indicator variables, their absolute value gives information about the magnitude of their contribution on each factor.

For instance, we note that (Woody = A, Woody = C, Fruity = C, Fruity = A) are the most relevant - *with roughly the same importance* - for the first factor. "Alcohol = C" is the most important one for the second factor.



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

We can use these functions to obtain the coordinates of the instances from the raw dataset (learning sample) into the space defined by the two factors. The graphical representation enables to visualize the relative positions of the instances labeled by their group membership (Figure 6).

Calculating the scores of unlabeled instances. Let us consider the unlabeled wine (W?) described into our reference paper (Abdi, 2007; page 3). We dispose to the following description:

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

We show below the calculations under Excel (the coefficients are rounded in 3-digit number). We highlight the coefficients activated by the non-zero indicators.

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

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

Here are the details for the first factor:

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

For the second factor, we obtain the score as follow:

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

We insert the unlabeled wine into the chart representing the instances of the learning sample. We point out the centroids for each group. The unlabeled wine seems related to the Loire wines.

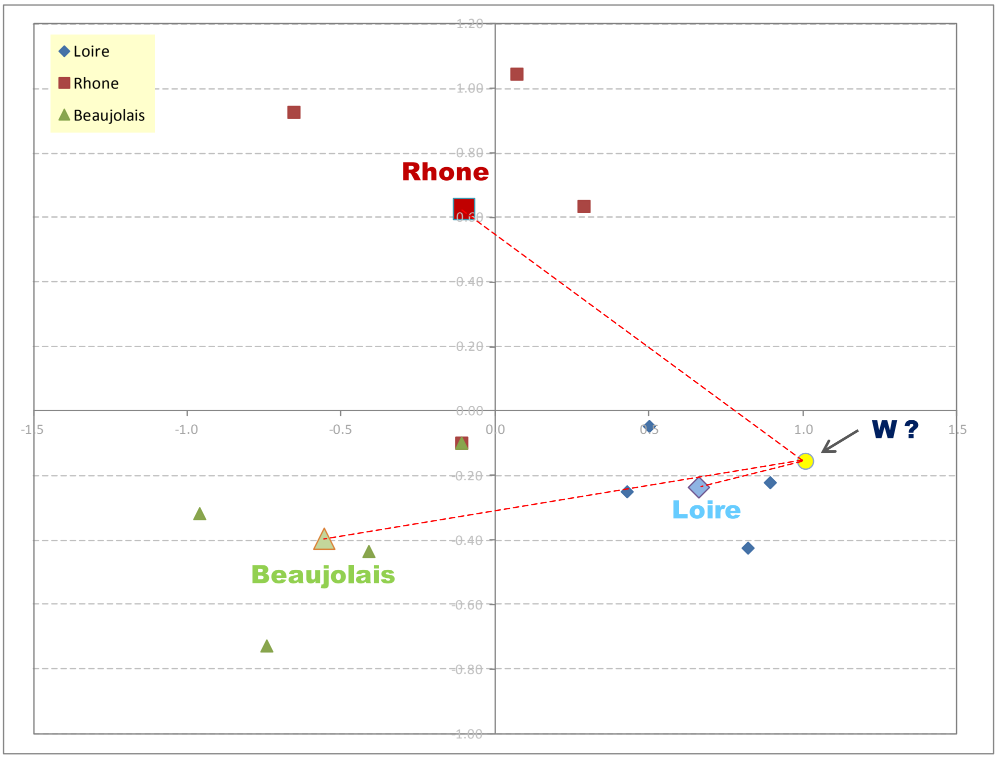


Figure 6 - Plot of instances, group centroids and the unlabeled instance W

Classification of an unlabeled instance – Euclidian distance to the group centroids. The classification of an unlabeled instance is based on the calculation of its Euclidian distance to the groups' centroids. Thus, the instance is assigned to the nearest group.

Below, we calculate the squared distance $[d^2(W,k)]$ of W with each centroid (Figure 3)

Group	Squared distance to the centroids
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

The visual impression coming from the graphical representation is clearly confirmed by the calculations. The wine W belongs to the Loire wine.

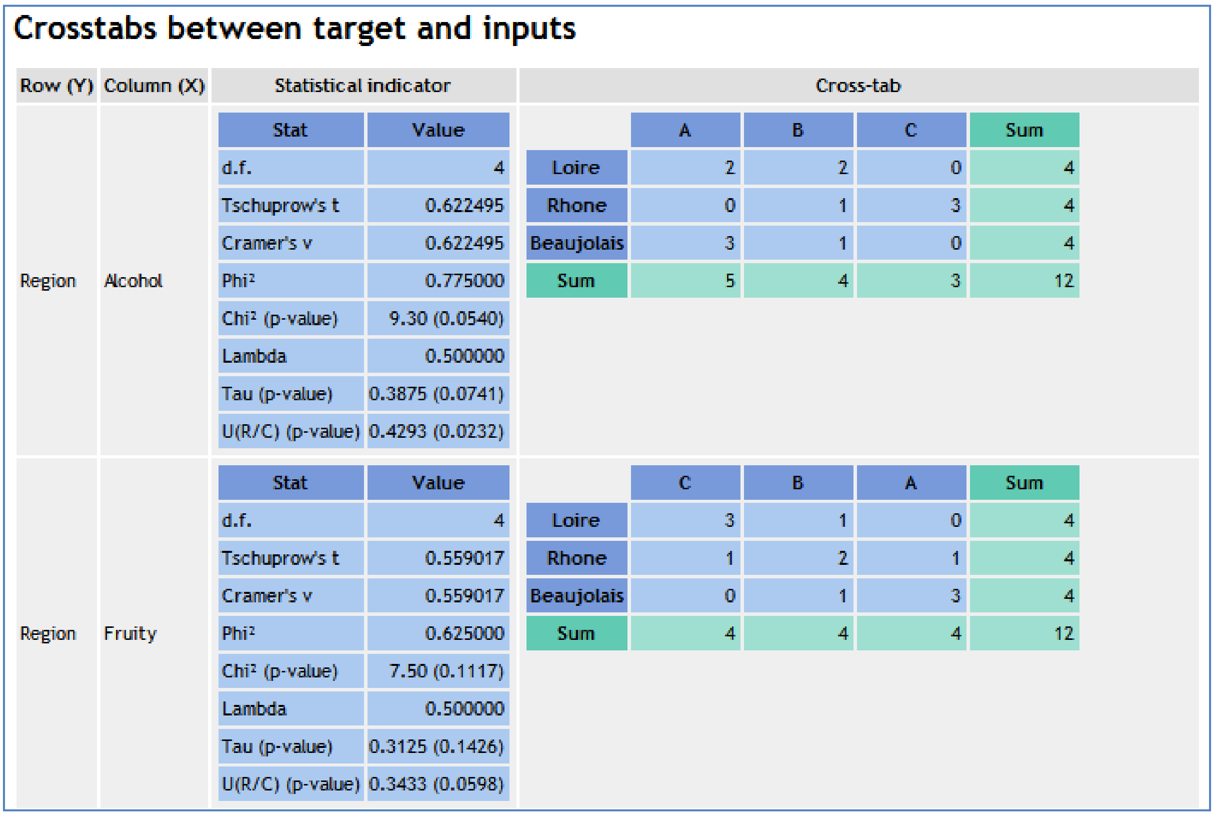
Classification process (2) – Generalized distance. When the number of instances is not equal, we must take into account the groups' size for the classification process. We use the generalized distance[4] which is defined as follow

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

Where π_k is the proportion of instances belonging to the group k. For our dataset, this no not modify anything since $\pi_k = \frac{1}{3}$,$\forall k$.

2.3.9 Crosstabs

The "Show All Crosstabs" was selected when we set the settings of the method (Figure 2).



Crosstabs between target and inputs

Row (Y)	Column (X)	Statistical indicator	Cross-tab
Region	Alcohol	see below	see below

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)

	A	B	C	Sum
Loire	2	2	0	4
Rhone	0	1	3	4
Beaujolais	3	1	0	4
Sum	5	4	3	12

Row (Y)	Column (X)
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)

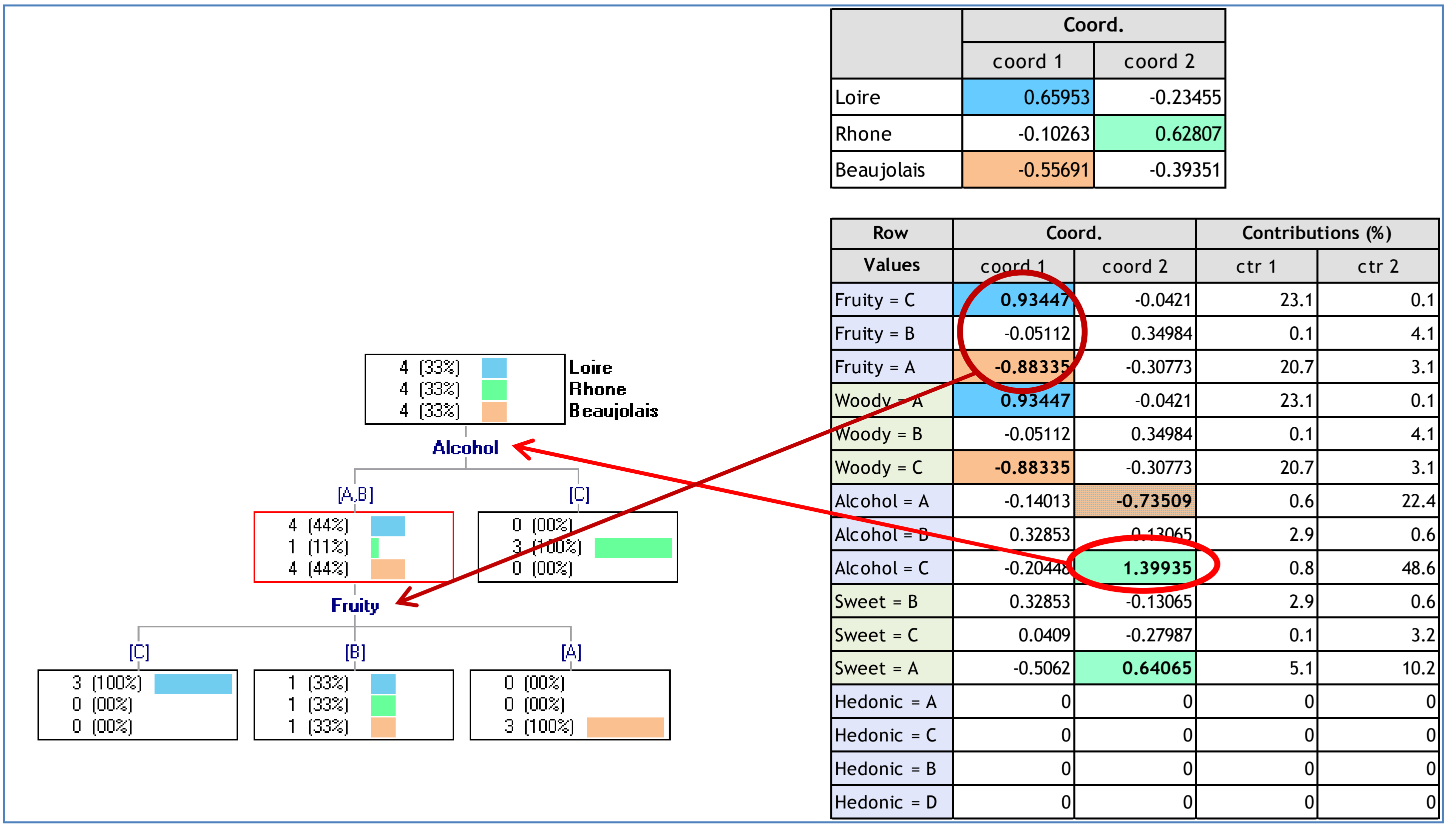
	C	B	A	Sum
Loire	3	1	0	4
Rhone	1	2	1	4
Beaujolais	0	1	3	4
Sum	4	4	4	12

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

Tanagra provides the crosstabs of each descriptor with the target attribute. The variables are sorted according to the Tschuprow's T criterion (which is closely related to the Cramer's V). We show only the two first tables here. As we highlight above, no association is significant because we have a very small learning sample (n = 12).

2.3.10 Comparison with other approaches

No method holds the truth in exploratory data analysis. Each in their own way explores such or such aspect of patterns in the data. This is our responsibility to identify the strengths and weaknesses of the approaches that we use. To confirm (or invalidate) our analysis above, we use a decision tree induction (using the SIPINA tool - http://eric.univ-lyon2.fr/~ricco/sipina.html). We compare the tree and the canonical structure provided by the discriminant correspondence analysis.



	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

The results are consistent. But because the decision tree discards redundant variables, a part of relevant information is hidden. For example, the role of WOODY is not highlighted into the tree. We could believe that it is irrelevant, in the same way as HEDONIC. We know that this is not true. WOODY is redundant with FRUITY. ***In our context***, where we want to understand the influence of all the variables, the discriminant correspondence analysis seems more suitable.

3 Programming the approach under R

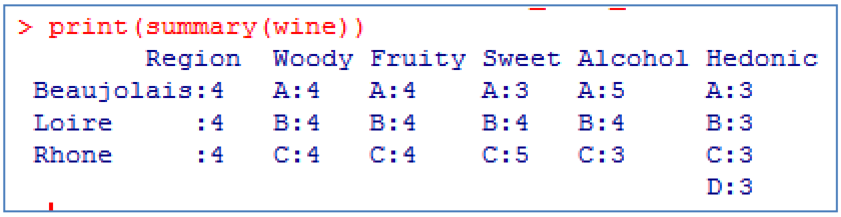
The discriminant correspondence analysis is not proposed in a package. But we can implement it easily because the correspondence analysis is available (in various packages moreover). So, the main step of our program is to create the overall contingency table from the raw dataset. In this section, we propose a small program to perform the discriminant correspondence analysis on our dataset.

3.1 Importing the dataset

We use the « xlsx »[5] package to import the "**french_win_dca.xls**" data file.

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

The **summary()** command enables to check the integrity of the dataset.



```
> 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
```

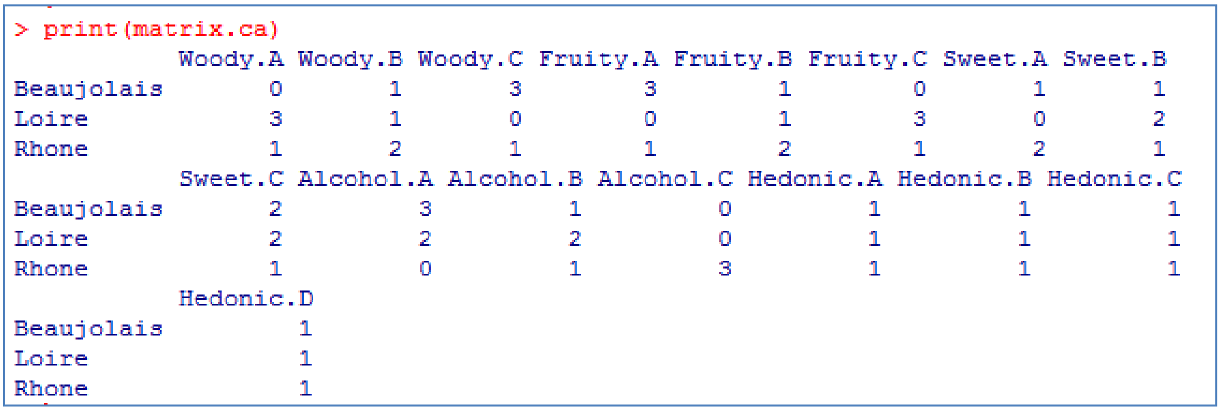
3.2 Constructing the overall crosstab for the correspondence analysis

There are several steps: (1) we select the descriptors from the whole data frame; (2) we create a call back function in order to calculate crosstab between the response variable and one descriptor; (3) we apply this function to all the descriptors; (4) we concatenate these crosstabs to obtain the overall contingency table.

```
#(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)
```

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

We set the column and the row names of the overall contingency table which are used by the correspondence analysis procedure.



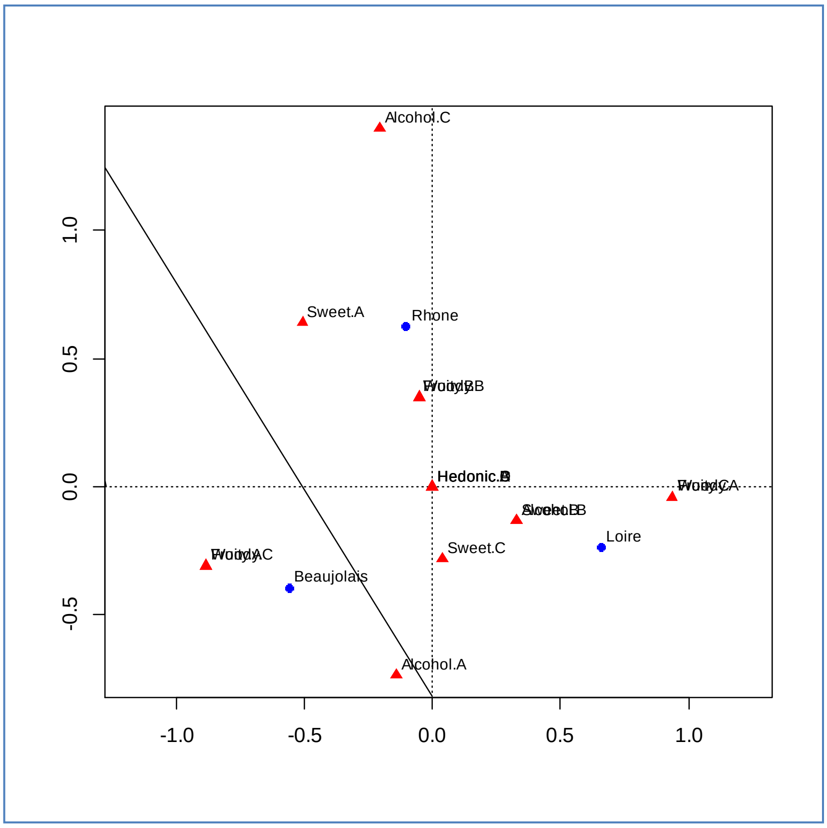
```
> print(matrix.ca)
           Woody.A Woody.B Woody.C Fruity.A Fruity.B Fruity.C Sweet.A Sweet.B
Beaujolais       0       1       3        3        1        0       1       1
Loire            3       1       0        0        1        3       0       2
Rhone            1       2       1        1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

We use the “ca” package for the correspondence analysis process[6].

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

We obtain the simultaneous plot of the row labels and the columns ones (without jittering).

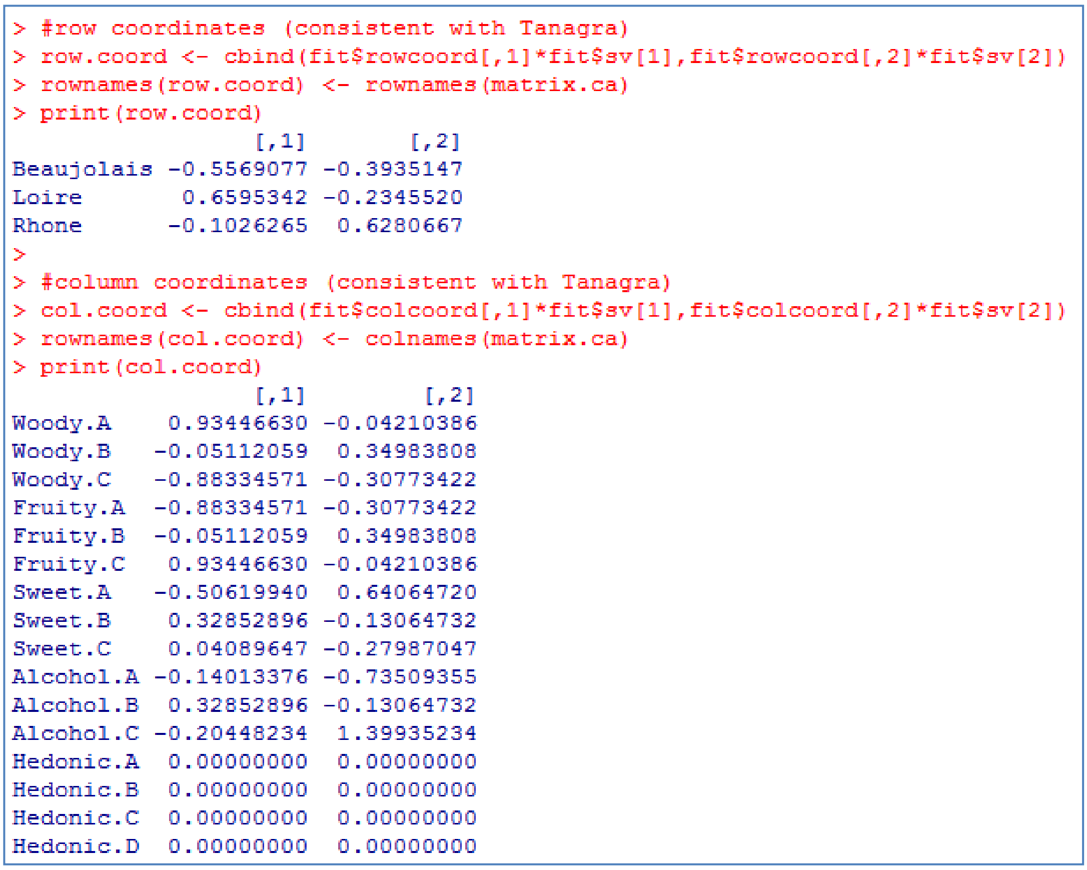


[6] http://cran.r-project.org/web/packages/ca/index.html

The results are exactly the same as those of 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)
```

We show here the coordinates of the values of the response variable and those of the descriptors (see Figure 4 for Tanagra).



```
> #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
```

The variables are not automatically sorted according to their contributions into the output tables. But we can do it easily.

4 Conclusion

The discriminant correspondence analysis is an elegant approach for performing discriminant analysis on discrete explanatory variables. We can compute the scores of the individuals and obtain a graphical representation which enables us to visually appreciate the proximities between the groups. In addition, we dispose of the tools provided by the correspondence analysis for the interpretation of the results.