Descriptive Discriminant Analysis

(Canonical Discriminant Analysis)

Multivariate characterization of differences between groups

Ricco RAKOTOMALALA

Outline

1. Problem statement
2. Determination of the latent variables (dimensions)
3. Reading the results
4. A case study
5. Classification of a new instance
6. Statistical tools (Tanagra, lda of R, proc candisc of SAS)
7. Conclusion
8. References

Issues

From a set of quantitative variables, how to compute a new representation space (dimensions) which enables to highlight the differences between groups of individuals

A population is subdivided in **K** groups (using a categorical variable, a label); the instances are described by **J** continuous descriptors.

E.g. Bordeaux wine (Tenenhaus, 2006; page 353). The rows of the dataset correspond to the year of production (**1924** to **1957**)

		Sun	Heat	Rain	Quality
Annee	**Temperature**	**Soleil**	**Chaleur**	**Pluie**	**Qualite**
1924	3064	1201	10	361	medium
1925	3000	1053	11	338	bad
1926	3155	1133	19	393	medium
1927	3085	970	4	467	bad
1928	3245	1258	36	294	good
1929	3267	1386	35	225	good

Descriptors

Group membership

Goal(s) :

(1) Descriptive (explanation): highlighting the characteristics which enable to explain the differences between groups → main objective in our context

(2) Predictive (classification): assign a group to an unseen instance → secondary objective in our context (but this is the main objective in the predictive discriminant analysis [PDA] context)

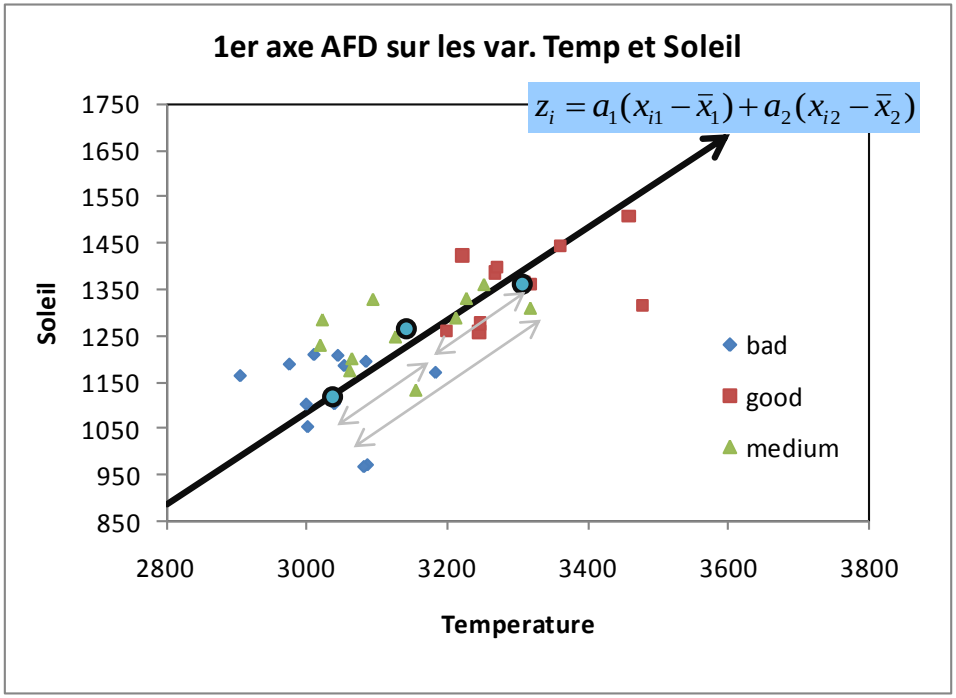
Aim: Determining the most parsimonious way to explain the differences between groups by computing a set of orthogonal linear combinations (canonical variables, factors) from the original descriptors. Canonical Discriminant Analysis.

The conditional centroids must be as widely separated as possible on the factors.

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

v = b + w

Total (variation) = **Between** class (variation) + **Within** class (variation)



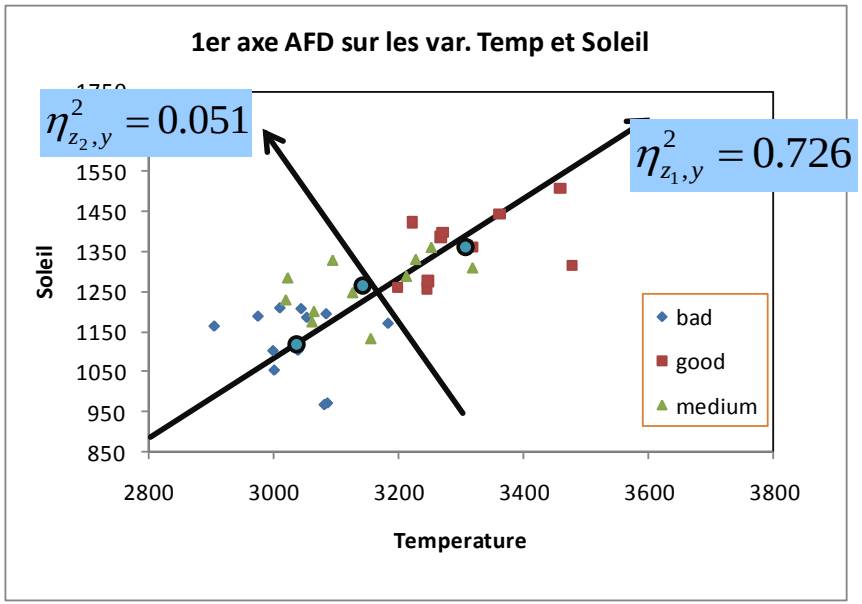
Maximizing a measure of the class separability:

the correlation ratio.

$$\eta^2_{z,y} = \frac{b}{v} \qquad \text{with} \qquad 0 \leq \eta^2_{z,y} \leq 1$$

1 → Perfect discrimination. All the points related to a groups are confounded to the corresponding centroid (W = 0)

0 → Impossible discrimination. All the centroids are confounded (B = 0)



Determining the coefficients (canonical coefficients) (a1,a2) which maximize the correlation ratio

Maximum number of "dimensions" (factors): M = min(J, K-1)

The factors are uncorrelated

A factor takes into account the differences not explained by the preceding factors

The correlation ratio measures the class separability

Solution

How to compute the canonical variables that summarize the between-class variation

Mathematical formulation

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_J \end{pmatrix}$$

« a » is the vector of coefficients which enables to define the canonical variable Z i.e. $z = a_1(x_1 - \bar{x}_1) + \cdots + a_J(x_J - \bar{x}_J)$

Huyghens' theorem → V = B + W

Total covariance matrix

$$V \to v_{lc} = \frac{1}{n}\sum_i (x_{il} - \bar{x}_l)(x_{ic} - \bar{x}_c)$$

Total sum of squares

$$TSS = a'Va$$

[ignoring a multiplication factor (1/n)]

Within groups covariance matrix

$$W \to w_{lc} = \frac{1}{n}\sum_k \sum_{i:y_i=k} (x_{il,k} - \bar{x}_{l,k})(x_{ic,k} - \bar{x}_{c,k})$$

$$RSS = a'Wa$$

Between groups covariance matrix

$$B \to b_{lc} = \sum_k \frac{n_k}{n}(\bar{x}_{l,k} - \bar{x}_l)(\bar{x}_{c,k} - \bar{x}_c)$$

$$ESS = a'Ba$$

The aim of DDA is to calculate the coefficients of the canonical variable which maximizes the correlation ratio

$$\max_a \frac{a'Ba}{a'Va} \Leftrightarrow \max_a \eta^2_{z,y}$$

Solution

$$\max_a \frac{a'Ba}{a'Va}$$ is equivalent to

$$\max_a a'Ba$$

Under the constraint $a'Va = 1$ *("**a**" is a unit vector)*

Solution: using the Lagrange function (λ is the Lagrange multiplier)

$$L(a) = a'Ba - \lambda(a'Va - 1)$$

$$\frac{\partial L(a)}{\partial a} = 0 \Rightarrow Ba = \lambda Va$$

$$\Rightarrow V^{-1}Ba = \lambda a$$

λ is the first eigenvalue of $V^{-1}B$

"a" is the corresponding eigenvector

The successive canonical variables are obtained from the eigenvalues and the eigenvectors of $V^{-1}B$.

- The number of non-zero eigenvalue is M = min(K-1, J) i.e. M canonical variables
- $\lambda = \eta^2$ The eigenvalue is equal to the square of the correlation ratio ($0 \leq \lambda \leq 1$)
- $\eta = \sqrt{\lambda}$ is the canonical correlation

Discriminant descriptive analysis

Bordeaux wine (X1 : Temperature and X2 : Sun)

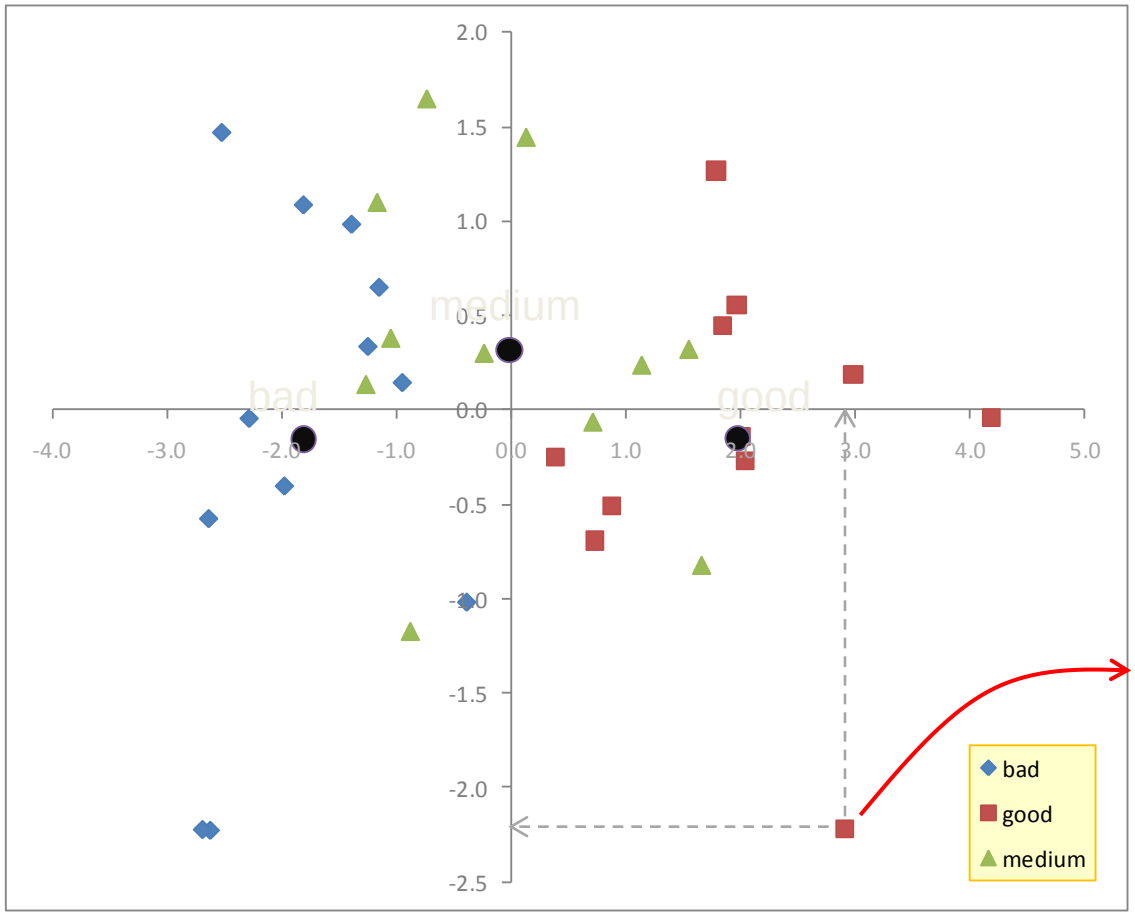
Number of factors

M = min (J = 2; K-1 = 2) = **2**

$$Z_{i2} = -0.0092(x_{i1} - \bar{x}_1) + 0.0105(x_{i2} - \bar{x}_2)$$

$$\eta_2 = \sqrt{0.051} = 0.225$$

The differences between the centroids are lesser on this factor.



$$Z_{i1} = 0.0075(x_{i1} - \bar{x}_1) + 0.0075(x_{i2} - \bar{x}_2)$$

$$\eta_1 = \sqrt{0.726} = 0.852$$

The differences between the centroids are high on this factor.

(2.91; -2.22): the coordinates of the individuals in the new representation space are called "factor scores" (SAS, SPSS, R…)

Since V = B + W, we can formulate the problem in other way:

$$\max_a \frac{a'Ba}{a'Wa}$$ is equivalent to $$\max_a a'Ba \quad \text{w.r.t.} \quad a'Wa = 1$$ (*"**a**" is a unit vector*)

- The factors are obtained from the eigenvalues and eigenvector of $W^{-1}B$.
- The eigenvectors of $W^{-1}B$ are the same as those of $V^{-1}B$ → the factors are identical.
- The eigenvalues are related with the following formula: $\rho = ESS / RSS$

$$\rho_m = \frac{\lambda_m}{1-\lambda_m}$$

E.g. Bordeaux wine

With only the variables "temperature" and "sun"

$$2.6432 = \frac{0.8518^2}{1-0.8518^2} = \frac{0.7255}{1-0.7255}$$

Root	Eigenvalue	Proportion	Canonical R
1	2.6432	0.9802	0.8518
2	0.0534	1	0.2251

E.g. The first factor explains 98% of the global between-class variation: 98% = 2.6432 / (2.6432 + 0.0534).

The two factors explain 100% of this variation [M = min(2, 3-1) = 2]

➔ The first factor is enough here!

we can state also the explained variation in percentage

Reading the results of DDA

Determining the right number of factors

Interpreting the factors

We want to check

H0: the correlation ratios of the "q" last factors are zero

$\Leftrightarrow$ H0: $\eta^2_{K-q} = \eta^2_{K-q-1} = \cdots = \eta^2_{K-1} = 0$

$\Leftrightarrow$ H0: we can ignore the "q" remaining factors

N.B. Checking a factor individually is not appropriate, because the relevance of a factor depends on the variation explained by the preceding factors.

Test statistic

$$\Lambda_q = \prod_{m=K-q}^{K-1} \left(1-\eta^2_m\right)$$

The lower is the value of LAMBDA, the more interesting are the factors.

In the case of Gaussian distribution (i.e. the data follows a multidimensional normal distribution in each group), we can use the Bartlett (chi-squared) or Rao transformation (Fisher).

Root	Eigenvalue	Proportion	Canonical R	Wilks Lambda	CHI-2	d.f.	p-value
1	2.6432	0.9802	0.8518	0.260568	41.0191	4	0
2	0.0534	1	0.2251	0.949308	1.5867	1	0.207802

Tanagra

Valeurs propres de Inv(E)*H = CanRsq/(1-CanRsq)

Test de H0 : les corrélations canoniques de la ligne en cours et suivantes sont égales à zéro

	Valeur propre	Différence	Proportion	Cumulé	Rapport de vraisemblance	Valeur de F approchée	DDL Num.	DDL Res.	Pr > F
1	2.6432	2.5898	0.9802	0.9802	0.26056848	14.39	4	60	<.0001
2	0.0534		0.0198	1.0000	0.94930784	1.66	1	31	0.2078

SAS

The two first factors are together significant at 5% level; but the last factor is not significant alone.

H0: all the correlation ratio are zero

$\Leftrightarrow$ H0: $\eta_1^2 = \cdots = \eta_{K-1}^2 = 0$

$\Leftrightarrow$ H0: we cannot distinguish the groups centroid in the global representation space

MANOVA test i.e. comparing multivariate means (centroids) of several groups

$$H_0 : \begin{pmatrix} \mu_{1,1} \\ \vdots \\ \mu_{J,1} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{1,K} \\ \vdots \\ \mu_{J,K} \end{pmatrix}$$

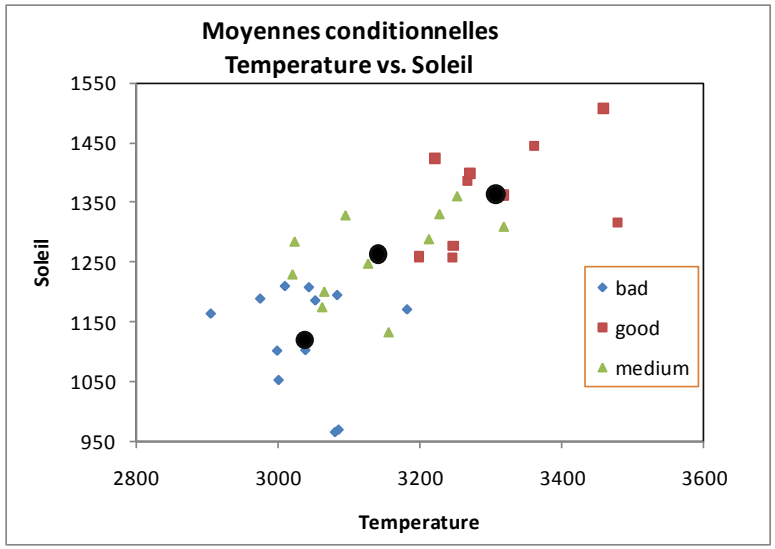
simultaneously

Test statistic:

Wilks' LAMBDA

$$\Lambda = \prod_{m=1}^{K-1} \left(1 - \eta_m^2\right)$$

The lower is the value of LAMBDA, the more different are the centroids ($0 \leq \Lambda \leq 1$).



LAMBDA de Wilks = 0.26

Bartlett transformation

CHI-2 = 41.02 ; p-value < 0.0001

Rao transformation

F = 14.39 ; p-value < 0.0001

Conclusion: At least one centroid is different to the others.

Descriptive discriminant analysis – Interpreting the canonical variables (factors)

Standardized and unstandardized canonical coefficients

Unstandardized coefficients

These coefficients enables to calculate the canonical scores of the individuals (coordinates of the individuals, discriminant scores)

$$Z = a_1(x_1 - \bar{x}_1) + \cdots + a_J(x_J - \bar{x}_J)$$
$$= a_0 + a_1 x_1 + \cdots + a_J x_J$$

The unstandardized canonical coefficients do not allow to compare the influence of the variables because they are not defined on the same unit.

Standardized coefficients

These are the coefficients of the DDA on standardized variables. We can obtain the same values by multiplying the unstandardized coefficients with the pooled within-class standard deviation of the variables. The coefficients (influence) of the variables become comparable.

$$\beta_j = a_j \times \sigma_j$$

$$\sigma_j^2 = \frac{1}{n-K} \sum_k \sum_{i:y_i=k}^{n_k} \left(x_{ij,k} - \bar{x}_{j,k}\right)^2$$

The pooled within class variance of the variable Xj

Standardized coefficients show the variable's contribution to calculating the discriminant score. Two correlated variables share their contribution, their true influence may be hidden (*W.R. Klecka, "Discriminant Analysis", 1980 ; page 33*).

We must complete this analysis by studying the structure coefficients table.

Quality = DDA (Temperature, Sun) >>

Canonical Discriminant Function

Coefficients	Unstandardized		Standardized	
Attribute	Root n°1	Root n°2	Root n°1	Root n°2
Temperature	0.007465	-0.009214	-0.653736	-0.806832
Sun	0.007479	0.010459	-0.604002	0.844707
constant	32.903185	16.049255	-	

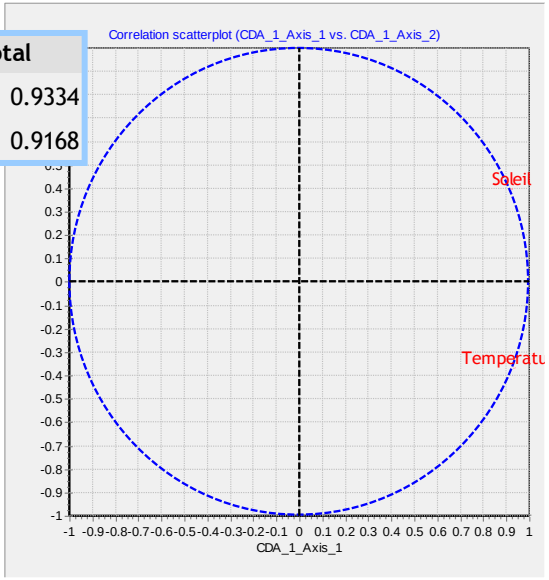
Descriptive discriminant analysis – Interpreting the canonical variables (factors)

Total structure coefficients

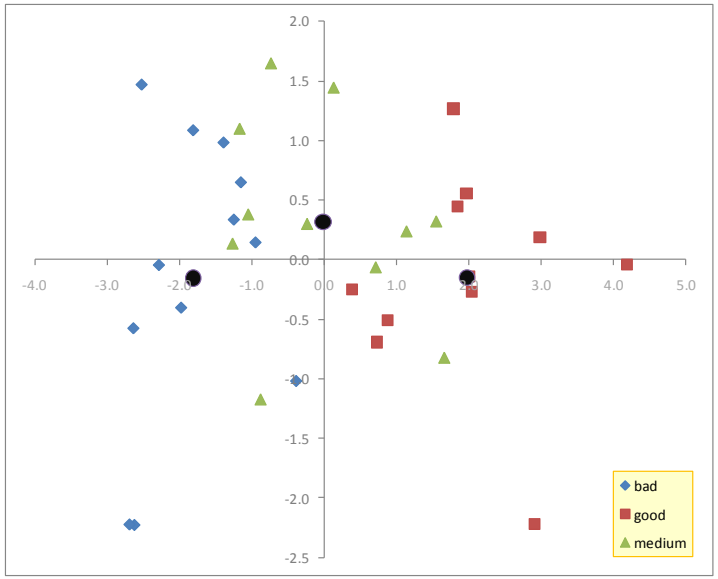
These are the bivariate correlation between the variables and the canonical variables.

We can visualize the correlation circle such as for PCA (principal component analysis).

Descriptors	Total
Temperature	0.9334
Sun	0.9168



The 1st factor corresponds to the combination of high temperature and high periods of sunshine.



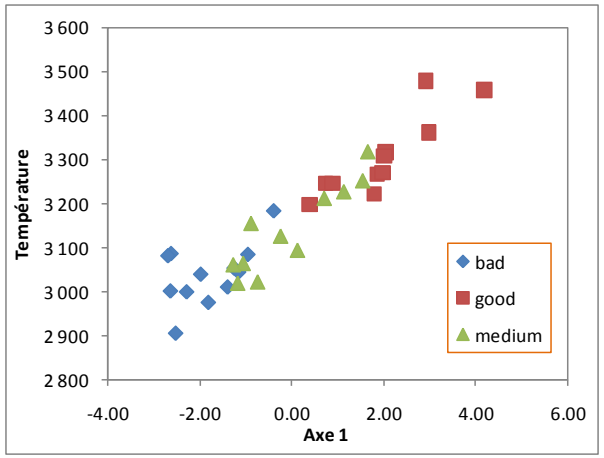
The combination of high temperature and high periods of sunshine correspond to "good" wine.

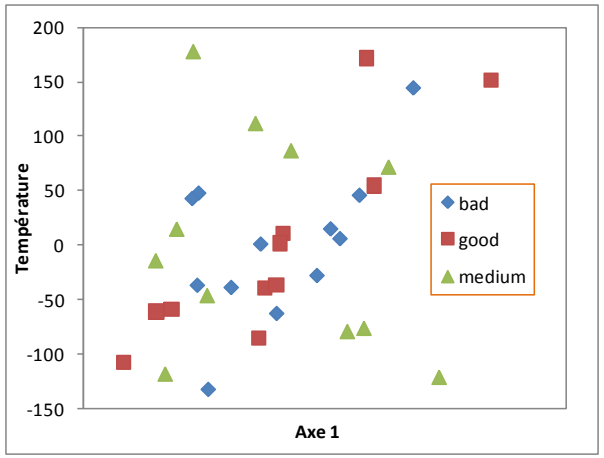
These correlation coefficients allow to interpret easily the factors.

If the sign are different to the standardized canonical coefficients → collinearity between the variables.

Within structure coefficients

These coefficients show how the variables are related to the canonical variable within the groups.





$r = 0.9334$

$r_w = 0.8134$

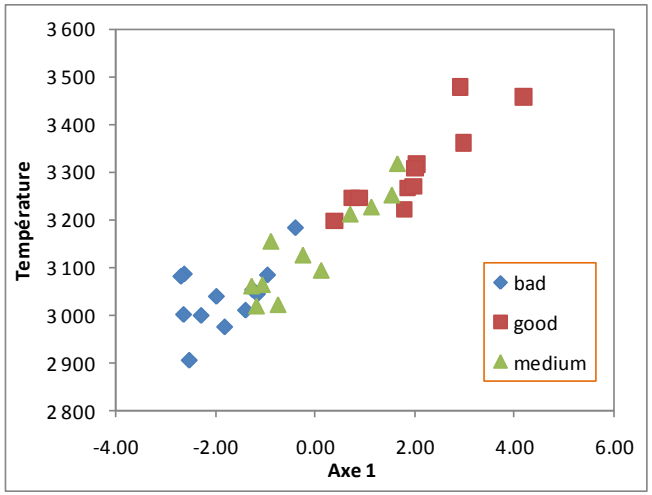
Root		Root n°1	
Descriptors	Total	Within	Between
Temperature	0.9334	0.8134	0.9949
Sun	0.9168	0.777	0.9934

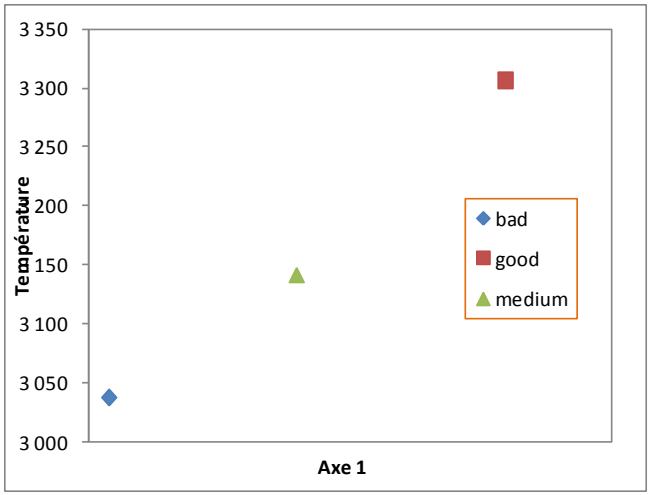
Often lower value than the total correlation (not always).

Between structure coefficients

Correlation of the variables with the factors by using only the group centroids.

Interesting but not always convenient. The value is +1 or -1 when we have only 2 groups (K = 2).





$r = 0.9334$

$r_B = 0.9949$

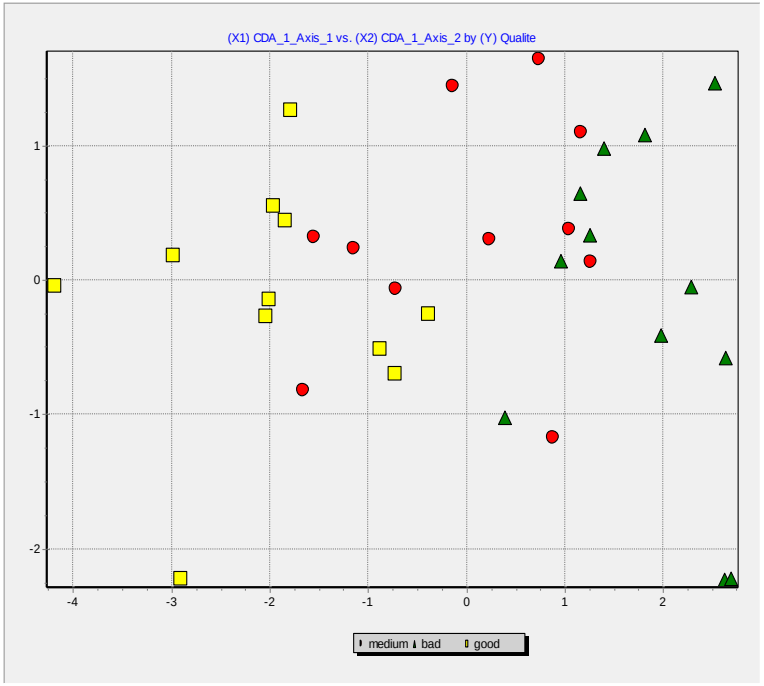
Root		Root n°1	
Descriptors	Total	Within	Between
Temperature	0.9334	0.8134	0.9949
Sun	0.9168	0.777	0.9934

Descriptive discriminant analysis – Interpreting the canonical variables (factors)

Group centroids into the discriminant representation space

Calculating the coordinates of the centroids in the new representation space.

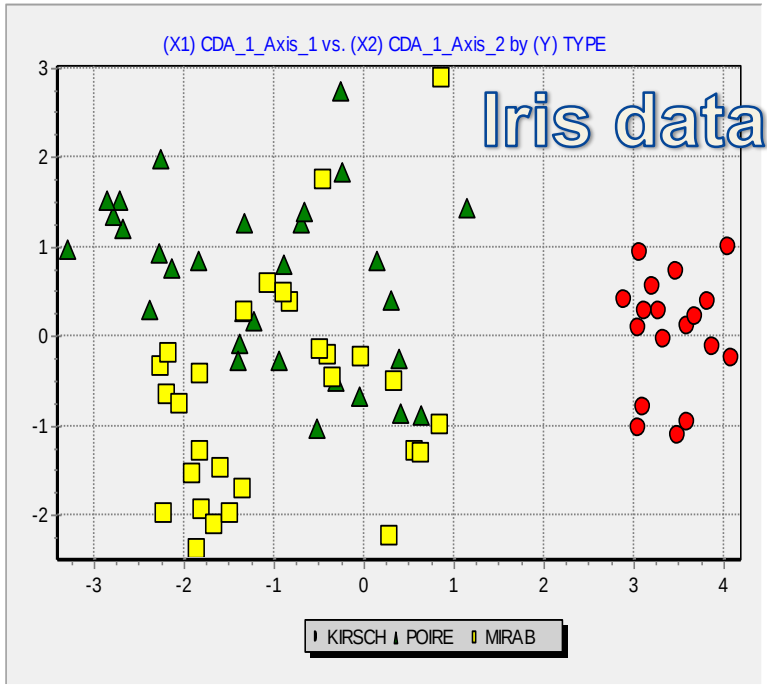
This allows to identify the groups which are well highlighted.



Qualite	Root n°1	Root n°2
bad	-1.804187	0.153917
good	1.978348	0.151489
medium	-0.01015	-0.3194
Sq Canonical corr.	0.725517	0.050692

The three groups are quite separate on the first factor

Nothing interesting on the second factor (low canonical correlation)



TYPE	Root n°1	Root n°2
KIRSCH	3.440412	0.031891
POIRE	-1.115293	0.633275
MIRAB	-0.981677	-0.674906
Sq Canonical corr.	0.789898	0.2544

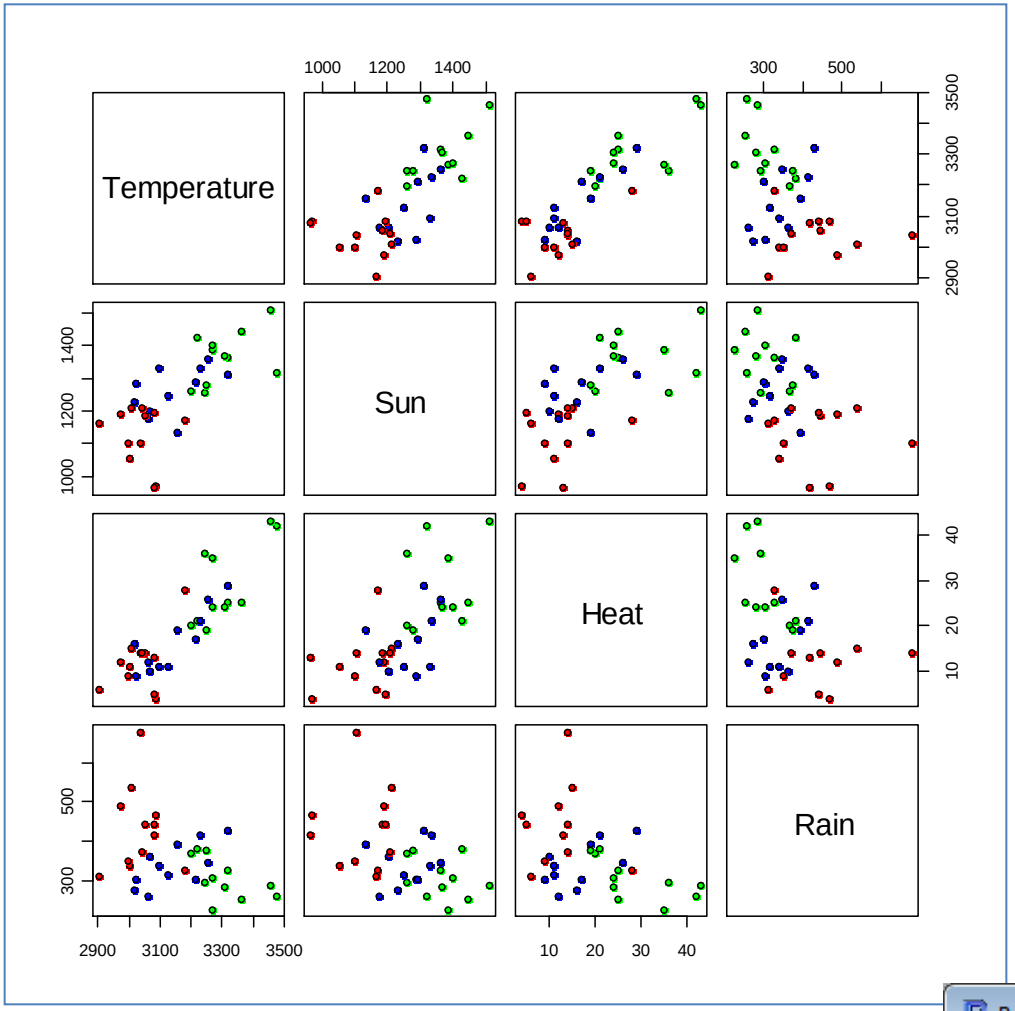
KIRSCH vs. the two other groups on the 1st factor

POIRE vs. MIRAB on the 2nd factor (significant canonical correlation)

Case study

Bordeaux wine (Tenenhaus, 2007; page 353)

Bordeaux wine - Description of the dataset



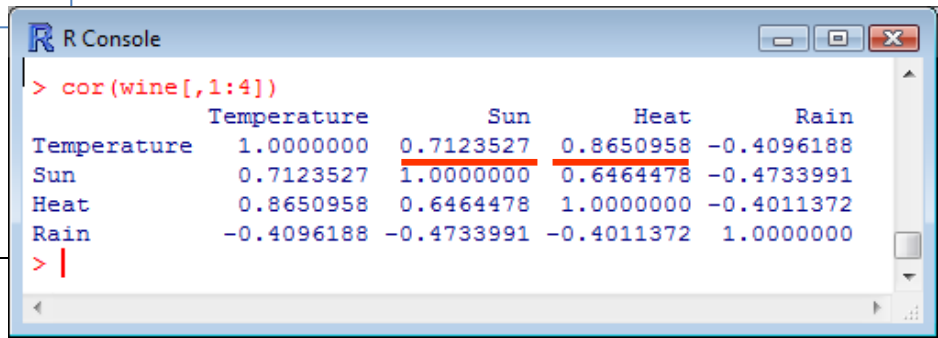
Some of the descriptors are correlated (see the correlation matrix)

(Red : Bad ; blue : Medium ; green : Good). The groups are discernible, especially for some combination of variables.

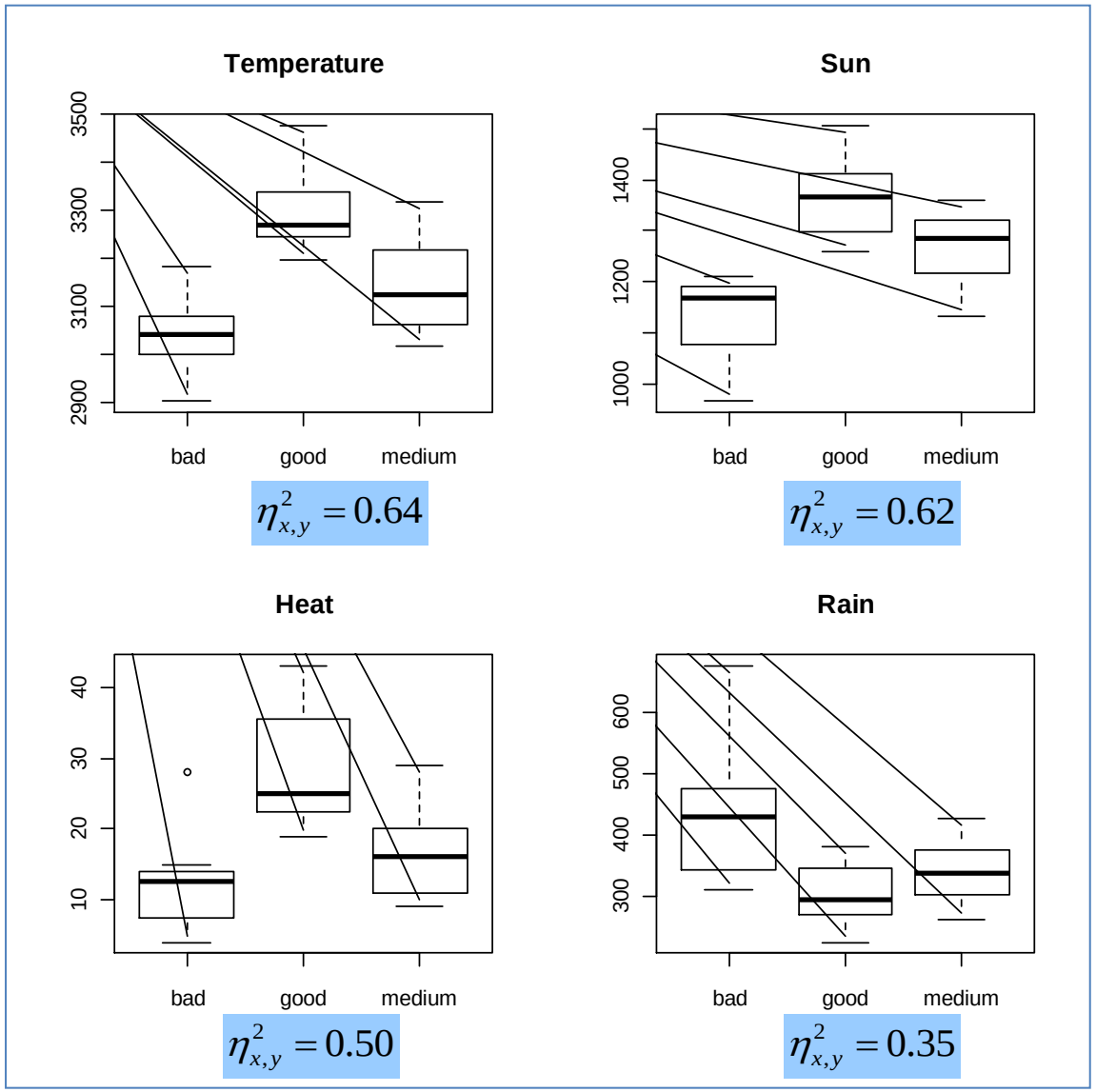
The influence on the quality is not the same according to the variables.

There are outliers...

Correlation matrix



Conditional distribution and correlation ratio



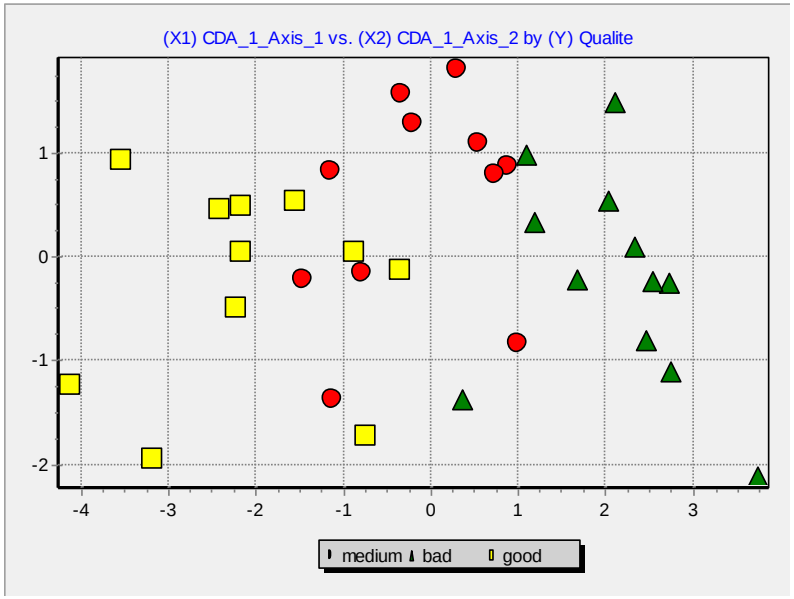
"Temperature", "Sun" and "Heat" enable to well distinguish the groups. "Rain" seems less decisive.

For all the variables, the univariate one-way ANOVA (the class means are equal or not) is significant at 5% level.

(a) The difference between groups is significant. (b) 96% of between-class variation is explained by the first factor. (c) The 2nd factor is not significant at 5% level, we can ignore it.

Roots and Wilks' Lambda

Root	Eigenvalue	Proportion	Canonical R	Wilks Lambda	CHI-2	d.f.	p-value
1	3.27886	0.95945	0.875382	0.205263	46.7122	8	0
2	0.13857	1	0.348867	0.878292	3.8284	3	0.280599



On the **first** factor, we observe the 3 groups. From the left to the right, we have the centroids of "good", "medium" and "bad".

The square of the correlation ratio for this factor is **0.766**. This is higher than any univariate correlation ratio of the variables (the higher is "temperature" with $\eta^2 = 0.64$).

Group centroids on the canonical variables

Qualite	Root n°1	Root n°2
medium	-0.146463	0.513651
bad	2.081465	-0.22142
good	-2.124227	-0.272102
Sq Canonical corr.	0.766293	0.121708

Interpreting the canonical variables

Canonical Discriminant Function

Coefficients	Unstandardized		Standardized	
Attribute	Root n°1	Root n°2	Root n°1	Root n°2
Temperature	-0.008575	0.000046	-0.750926	0.004054
Soleil	-0.006781	0.005335	-0.547648	0.430858
Chaleur	0.027083	-0.127772	0.198448	-0.936227
Pluie	0.005872	-0.006181	0.445572	-0.469036
constant	32.911354	-2.167589	-	

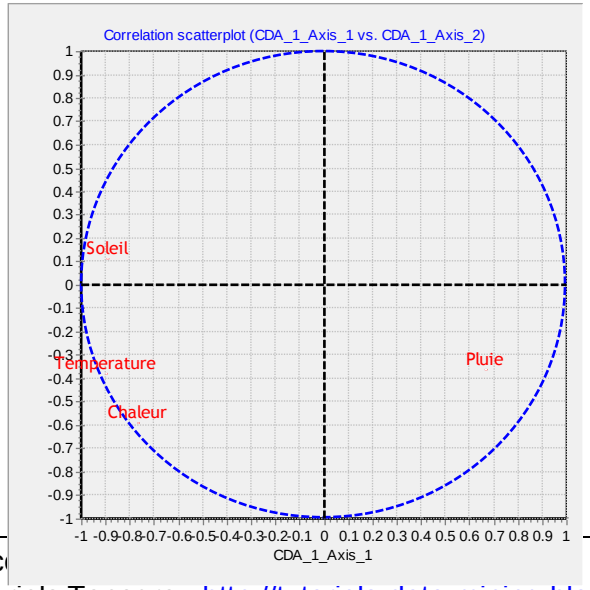
Factor Structure Matrix - Correlations

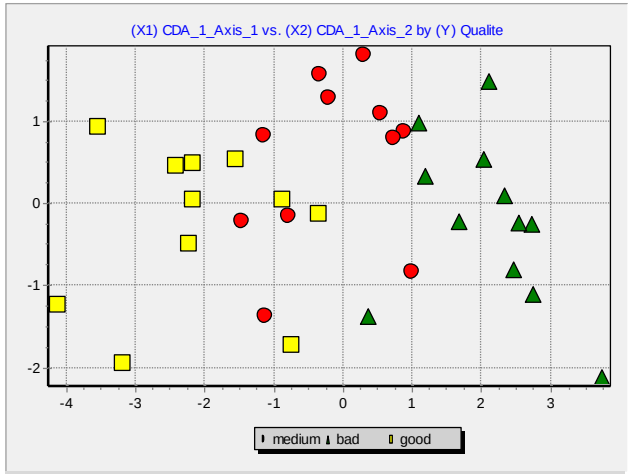
Root	Root n°1			Root n°2		
Descriptors	Total	Within	Between	Total	Within	Between
Temperature	-0.9006	-0.7242	-0.9865	-0.3748	-0.5843	-0.1636
Soleil	-0.8967	-0.7013	-0.9987	0.1162	0.1761	0.0516
Chaleur	-0.7705	-0.5254	-0.9565	-0.59	-0.7799	-0.2919
Pluie	0.6628	0.3982	0.9772	-0.3613	-0.4208	-0.2123

The first factor brings into opposition the "temperature" and the "sun" on the one side (high values: good wine), and the "rain" on the other side (high values: bad wine).

The influence of "heat" seems unclear. It has a positive influence on the first factor according to the canonical coefficients table. But it has a negative relation to the first factor according to the structure coefficients table.

Actually, this variable is highly correlated to "temperature". The partial correlation ratio of "heat" by controlling "temperature" is very low (Tenenhaus, page 376) $\eta^2_{x_3, y/x_1} = 0.0348$





Coordinates of the individuals with the group membership. Correlation circle.

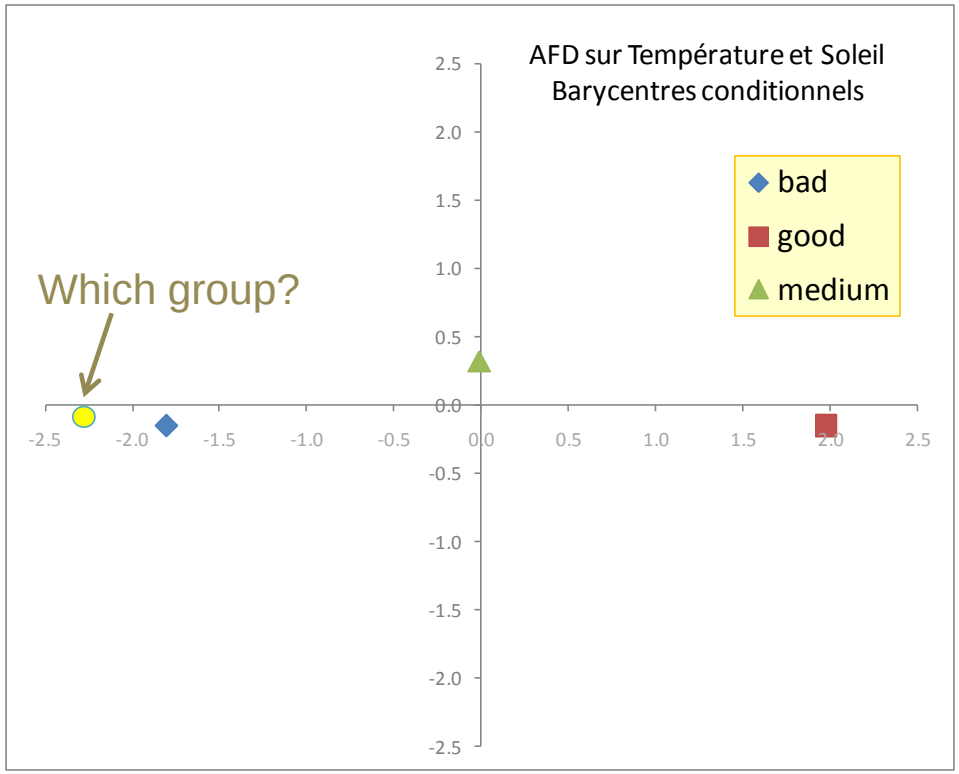
Classifying an unseen instance

Using the results of DDA to determine the group membership of a new instance

Preamble

The linear (predictive) discriminant analysis (PDA) offers a more attractive theoretical framework for prediction, with explicit probabilistic assumptions.

Nevertheless, we can use the results of the DDA to classify individuals based on geometric rules.



Steps:

1. As from the description of the individual, its coordinates in the discriminant dimensions are computed.
2. The distance to the conditional centroids is computed.
3. The instance is assigned to the group of which the centroid is the closest.

X1 = 3000 – X2 = 1100 – Year 1958 (based on the weather forecast)

1. Calculating the coordinates

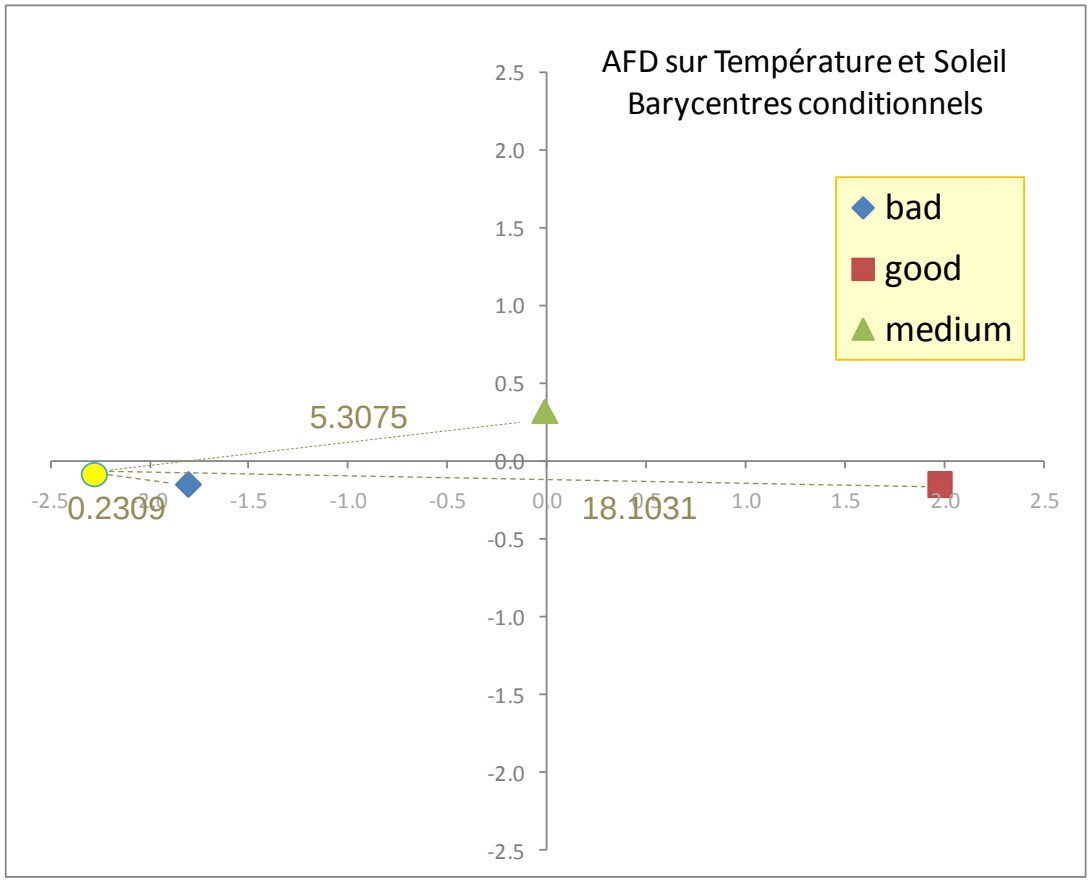
$$z_1 = 0.007457 \times x_1 + 0.007471 \times x_2 - 32.868122$$
$$= 0.007457 \times 3000 + 0.007471 \times 1100 - 32.868122$$
$$= -2.2780$$
$$z_2 = -0.009204 \times x_1 + 0.010448 \times x_2 + 16.032152$$
$$= -0.009204 \times 3000 + 0.010448 \times 1100 + 16.032152$$
$$= -0.0862$$

2. Calculating the distance to the centroids

$$d^2(bad) = (-2.2780 - (-1.8023))^2 + (-0.0832 - (-0.1538))^2$$
$$= 0.2309$$
$$d^2(good) = 18.1031$$
$$d^2(medium) = 5.3075$$

3. Conclusion

The vintage 1958 has a high probability to be “bad”. It has a very low probability to be “good”.



We can obtain the same distance as preceding in the initial representation space by using the W^{-1} metric: this is the Mahalanobis distance.

For the instance "1958", we calculate its distance to the "bad" centroid as follows…

$$d^2(bad) = (x-\mu_{bad})' W^{-1} (x-\mu_{bad})$$

$$= (3000-3037.3; 1100-1126.4) \begin{pmatrix} 7668.46 & 1880.15 \\ 1880.15 & 6522.33 \end{pmatrix}^{-1} \begin{pmatrix} 3000-3037.3 \\ 1100-1126.4 \end{pmatrix}$$

$$= (-37.33 \quad -26.42) \begin{pmatrix} 0.000140 & -0.000040 \\ -0.000040 & 0.000165 \end{pmatrix} \begin{pmatrix} -37.33 \\ -26.42 \end{pmatrix}$$

$$= 0.2309$$

$$W = \begin{pmatrix} 7668.46 & 1880.15 \\ 1880.15 & 6522.33 \end{pmatrix}$$

Is the pooled within class SSCP matrix (sum of squares and cross products) [i.e. the covariance matrix multiplied by the degree of freedom (n-K)]

Why the results of DDA are important?

1. We have in addition an explanation of the prediction. "1958" is probably "bad" because of low temperature and low sun.
2. We can use only the significant canonical variables for the prediction. This is a kind of regularization (see "reduced rank LDA", Hastie et al., 2001).

Finding the closest centroid (minimization). We can transform it in a maximization problem by multiplying with -0.5

For an instance "i", we calculate as follows its distance to the centroid of the group "k". We take into account Q canonical variables (Q = M if we treat all the factors).

$$k^* = \arg\min_k d_i^2(k) \Leftrightarrow k^* = \arg\max_k f_i(k)$$

$$d_i^2(k) = \sum_{m=1}^{Q} (z_{im} - \bar{z}_{m,k})^2$$
$$= \sum_{m=1}^{Q} z_{im}^2 + \bar{z}_{m,k}^2 - 2 z_{im} \bar{z}_{m,k}$$

$$f_i(k) = \sum_{m=1}^{Q} \left(\bar{z}_{m,k} \times z_{im} - \frac{1}{2} \bar{z}_{m,k}^2 \right)$$
$$= \sum_{m=1}^{Q} \bar{z}_{m,k} \times z_{im} - \frac{1}{2} \sum_{m=1}^{Q} \bar{z}_{m,k}^2$$

Discriminant function for the factor "m"

$$z_m = a_{0m} + a_{1m} x_1 + a_{2m} x_2 + \cdots + a_{Jm} x_J$$

We have a linear classification function.

E.g. Bordeaux wine with "temperature" (x_1) and "sun" (x_2) – Only one factor (Q = 1)

$$f(bad) = -1.8023 \times (0.007457 x_1 + 0.007471 x_2 - 32.868122) - \frac{1}{2}(-1.8023)^2$$
$$= -0.0134 x_1 - 0.0135 x_2 + 57.6129$$
$$f(good) = 0.0147 x_1 + 0.0148 x_2 - 66.9081$$
$$f(medium) = -0.0001 x_1 - 0.0001 x_2 + 0.3331$$

For the instance (x1 = 3000; x2 = 1100)

$$f(bad) = 2.4815$$
$$f(good) = -6.5447$$
$$f(medium) = 0.0230$$

Conclusion: the vintage "1958" will be probably « bad »

The parametric linear discriminant analysis makes assumptions about the distribution and the dispersion of the observations (normal distribution, homogeneity of variances/covariances)

Classification function from PDA

$$d(Y_k, X) = \ln[P(Y = y_k)] + \mu_k \Sigma^{-1} X' - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k'$$

Classification rule from the DDA when we handle all the factors (M factors)

Equivalence

In conclusion, the classification rule of DDA is equivalent to the one of PDA if we have balanced class distribution i.e.

$$P(Y = y_1) = \cdots = P(Y = y_K) = \frac{1}{K}$$

Some tools make this assumption by default (e.g. default settings for the SAS PROC DISCRIM)

Introducing the correction derived from the estimated class distribution will improve the error rate (Hastie et al., 2001 ; page 95).

Some data mining tools

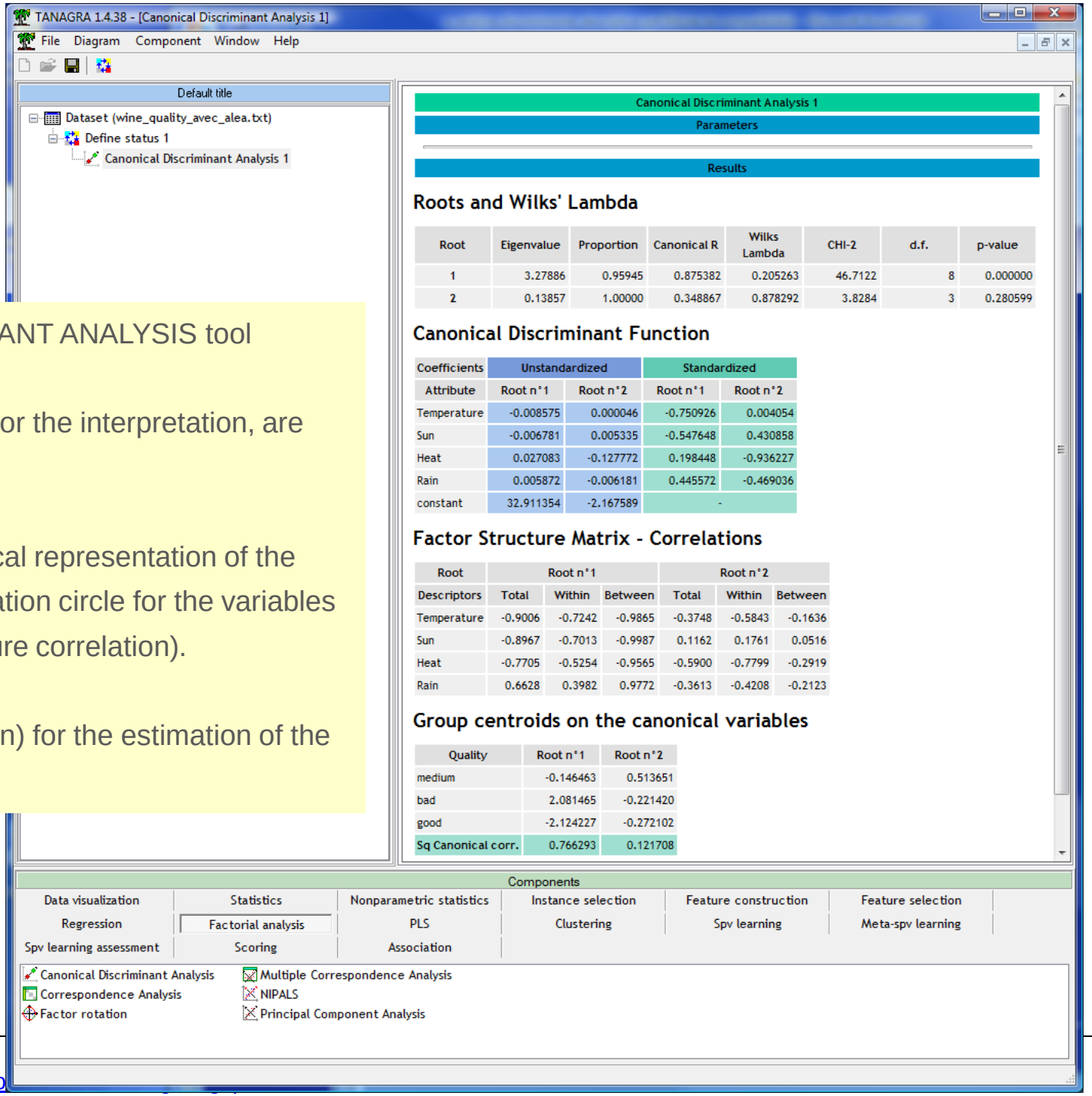
Tanagra , R and SAS

CANONICAL DISCRIMINANT ANALYSIS tool

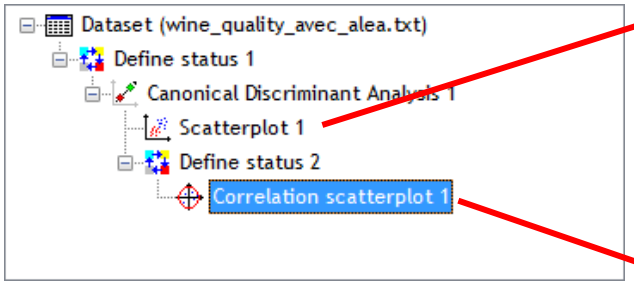
The main results, usable for the interpretation, are available.

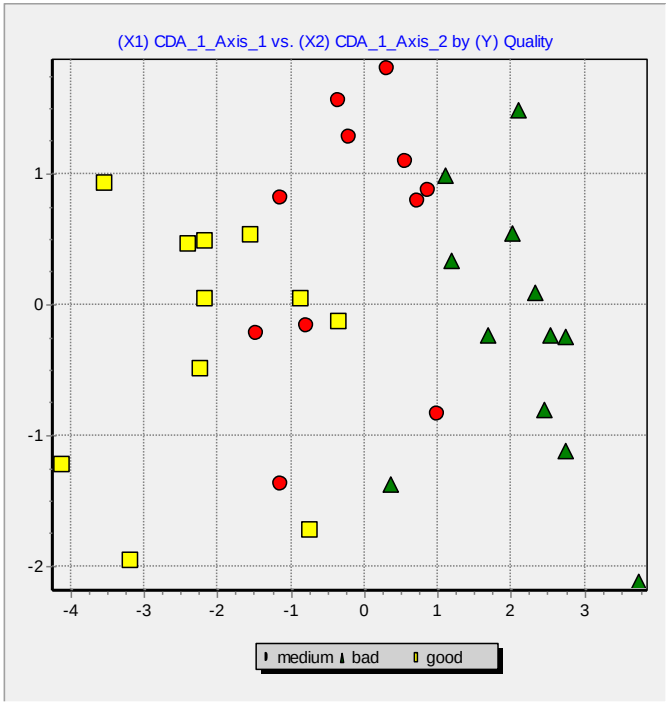
We can obtain the graphical representation of the individuals and the correlation circle for the variables (based on the total structure correlation).

French references use (1/n) for the estimation of the covariance.

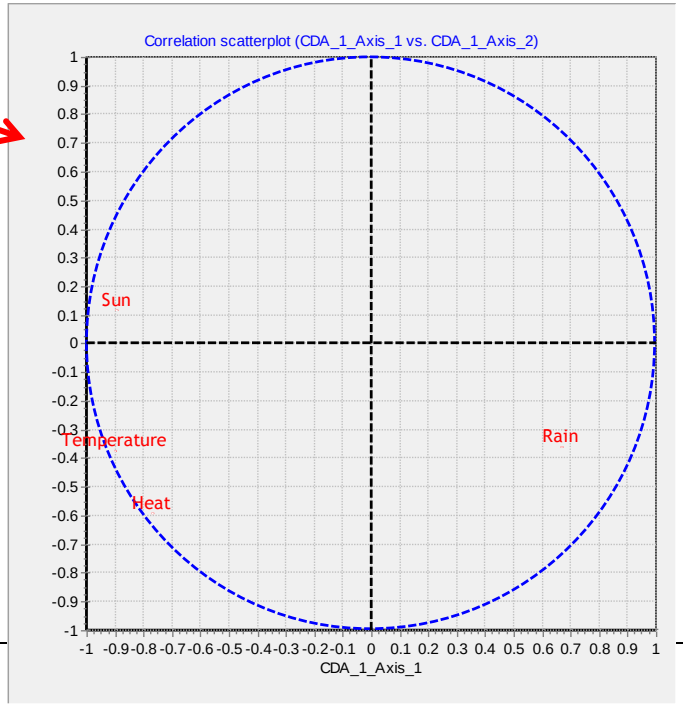


Plotting the individuals into the discriminant dimensions





Correlation circle



The "lda" procedure from the MASS package

```
rm (list=ls())

#chargement des données
library(xlsReadWrite)
setwd("D:/DataMining/Databases_for_mining/dataset_for
wine <- read.xls(file="wine_quality.xls",colNames=T)

library(MASS)
#analyse discriminante descriptive
wine.lda <- lda(Quality ~ ., data = wine)
print(wine.lda)

#carte factorielle
plot(wine.lda)
```

R Console

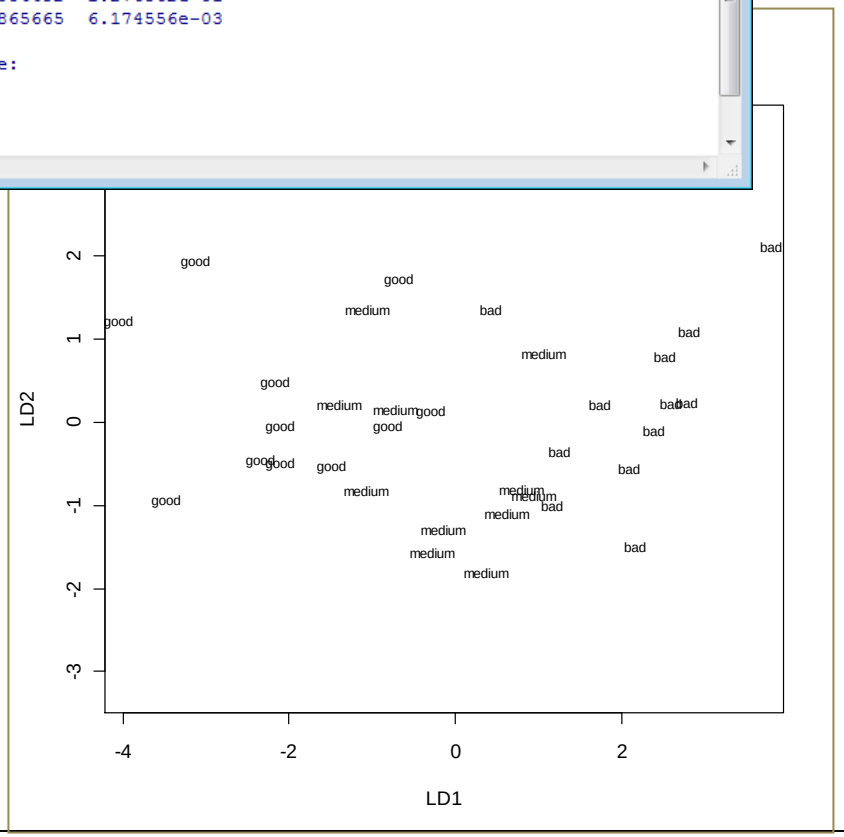
```
> print(wine.lda)
Call:
lda(Quality ~ ., data = wine)

Prior probabilities of groups:
      bad      good    medium
0.3529412 0.3235294 0.3235294

Group means:
       Temperature      Sun     Heat     Rain
bad       3037.333 1126.417 12.08333 430.3333
good      3306.364 1363.636 28.54545 305.0000
medium    3140.909 1262.909 16.45455 339.6364

Coefficients of linear discriminants:
                    LD1           LD2
Temperature -0.008566046 -4.625059e-05
Sun         -0.006773869 -5.329293e-03
Heat         0.027054492  1.276362e-01
Rain         0.005865665  6.174556e-03

Proportion of trace:
   LD1    LD2
0.9595 0.0405
>
```



The output is concise.

But with some programming instructions, we can obtain better. This is one of the main advantages of R.

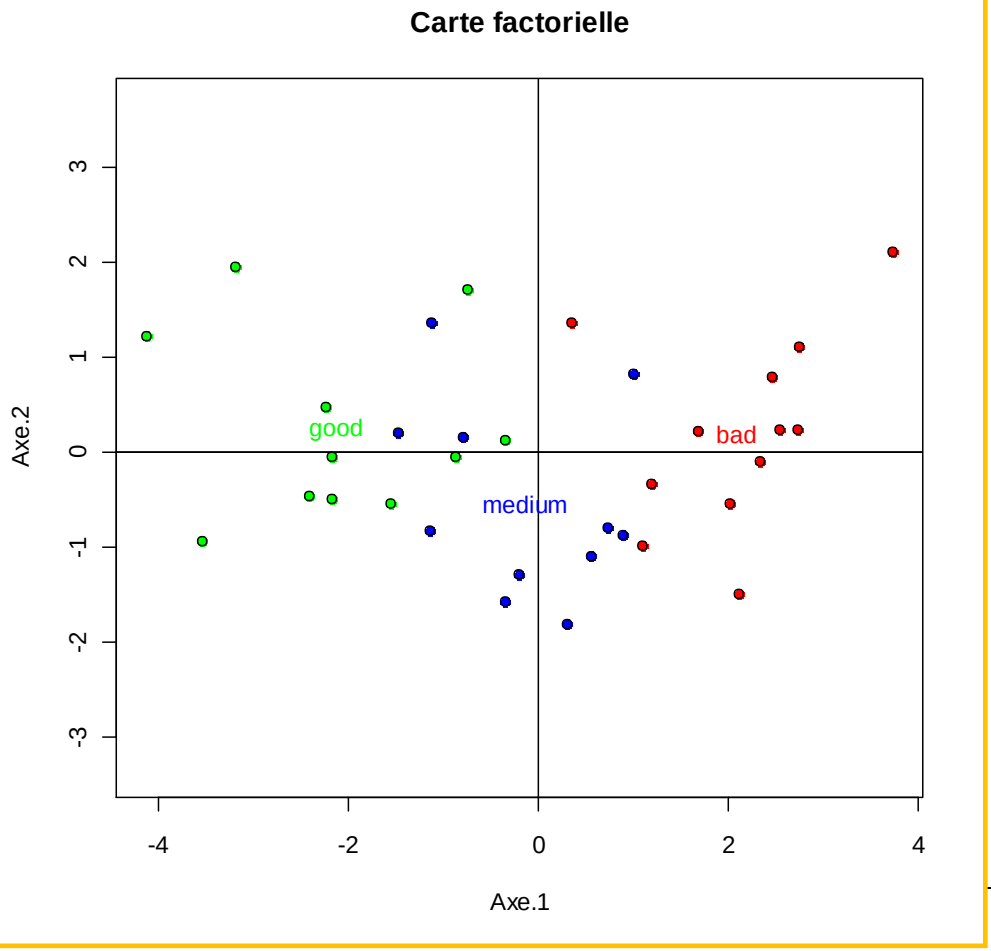
English-speaking references use [1/(n-1)] for the estimation of the covariance.

D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\discriminant_analysis\comparison\wine\wine lda pour doc.r

```
#calcul des projections sur les axes factoriels
wine.pred <- predict(wine.lda,data=wine)

#calcul des moyennes conditionnelles sur les axes
m <- matrix(rep(0,6),nrow=3,ncol=2)
for (i in 1:3){
  for (j in 1:2){
    m[i,j] <- mean(wine.pred$x[unclass(wine$Quality)==i,j])
  }
}

#graphique - carte factorielle avec les moyennes conditionnelles des groupes (asp = 1 pour que le positionnement relatif sur les axes soit respecté)
plot(wine.pred$x[,1],wine.pred$x[,2],main="Carte factorielle",xlab="Axe.1",ylab="Axe.2",pch=21,bg=c("red","green","blue")[unclass(wine$Quality)],asp=1)
abline(0,0,h=0)
abline(a=0,b=0,v=0)
text(m[,1],m[,2],labels=levels(wine$Quality),col=c("red","green","blue"))
```



DDA with R

With some programming instructions,
the result is worth it …

wine sas code.sas

```
proc candisc data = ucidata.wine_bordeaux;
class quality;
var Temperature Sun Heat Rain;
run;
```

Comprehensive results.

The “ALL” option allows to obtain all the intermediate results (matrices V, W, B ; etc.).

English-speaking references use [1/(n-1)] for the estimation of the covariance (such as R).

SAS

Fichier Édition Affichage Outils Solutions Fenêtre Aide

Résultats

- Résultats
  - Candisc: Le Système SAS
    - Counts
    - n obs
    - Class Levels
    - Statistiques multivariées
    - Canonical Analysis
      - Corrélations canoniques
      - Structure
        - Total
        - Inter
        - Combiné
      - Coefficients
        - Total
        - Combiné
        - Brut
    - Moyennes de classes

Résultats | Explorateur

Sortie - (Sans titre)

Le Système SAS 06:38 F

The CANDISC Procedure

Coefficients canoniques normalisés sur la totalité de l'échantillon

Variable	Libellé	Can1	Can2
Temperature	Temperature	1.209391427	-0.006529859
Sun	Sun	0.857727422	-0.674810955
Heat	Heat	-0.270993045	1.278475787
Rain	Rain	-0.536131215	0.564364371

Coefficients canoniques normalisés et groupés intra-classe

Variable	Libellé	Can1	Can2
Temperature	Temperature	0.7501261906	-.0040501510
Sun	Sun	0.5470642419	-.4303989053
Heat	Heat	-.1982365052	0.9352290652
Rain	Rain	-.4450968911	0.4685360966

Coefficients canoniques bruts

Variable	Libellé	Can1	Can2
Temperature	Temperature	0.0085660455	-.0000462506
Sun	Sun	0.0067738690	-.0053292933
Heat	Heat	-.0270544919	0.1276361644
Rain	Rain	-.0058656650	0.0061745562

Moyennes de classes sur les variables canoniques

Quality	Can1	Can2
bad	-2.079247035	0.221183875
good	2.121962956	0.271812018
medium	0.146306536	-0.513103518

Sortie - (Sans titre) | Journal - (Sans titre) | wine sas code.sas

C:\Users\Maison

Conclusion

DDA: multivariate method for groups' description and characterization

Tools for the interpretation of the results (test for significance of canonical variables, canonical coefficients, structure coefficients...)

Tools for the visualization of the results (individuals, variables)

The approach is related to other factorial methods (principal component analysis, canonical correlation)

The approach is in nature descriptive, but it can be implemented in a predictive framework easily.

The approach provides a white-box prediction (we can understand for which reason an unseen instance is assigned to such group).

References

M. Tenenhaus, “Statistique – Méthodes pour décrire, expliquer et prévoir” (*Statistics - Methods to describe, explain and predict*), Dunod, 2007. Chapter 10, pages 351 to 386.

W.R. Klecka, “Discriminant Analysis”, Sage University Paper series on Quantitative Applications in the Social Sciences, n°07-019, 1980.

C.J. Huberty, S. Olejnik, “Aplied MANOVA and Dscriminant Analysis”, 2nd Edition, Wiley, 2006.