

# 1 Topic

## Linear Discriminant Analysis – Data Mining Tools Comparison (Tanagra, R, SAS and SPSS).

Linear discriminant analysis is a popular method in domains of statistics, machine learning and pattern recognition. Indeed, it has interesting properties: it is rather fast on large bases; it can handle naturally multi-class problems (target attribute with more than 2 values); it generates a linear classifier linear, easy to interpret; it is robust and fairly stable, even applied on small databases; it has an embedded variable selection mechanism. Personally, I appreciate linear discriminant analysis because we can have multiple interpretations (probabilistic, geometric), and thus highlights various aspects of supervised learning.

Discriminant analysis is both a descriptive and a predictive method<sup>1</sup>. In the first case, we say Canonical Discriminant Analysis. We can consider the approach as a dimension reduction technique (a factor analysis). We want to highlight latent variables which explain the difference between the classes defined by the target attribute. In the second case, we can consider the approach as a supervised learning algorithm which intends to predict efficiently the class membership of individuals. Because we have a linear combination of the variables, we have a linear classifier. The purposes are therefore not intrinsically identical even if, when we analyze deeply the underlying formulas, we realize that the two approaches are closely related. Some bibliographic references maintain anyway the confusion by presenting them in a single framework.

Tanagra differentiates clearly the two approaches by providing two separate components: LINEAR DISCRIMINANT ANALYSIS (SPV LEARNING tab) for the prediction approach; CANONICAL DISCRIMINANT ANALYSIS (FACTORIAL ANALYSIS tab) for the descriptive (factorial) approach. It is the same for SAS software with respectively DISCRIM and CANDISC procedures<sup>2</sup>. Others combine them. This is the case for SPSS and R, mixing results which refer to different goals. For specialists who know how to distinguish important elements depending on the context, this amalgam is not a problem. For beginners, it is a bit more problematic. One can be disturbed by results which do not seem directly related to the purposes of the study.

In this tutorial, we detail in a first time with the TANAGRA outputs about Predictive Linear Discriminant Analysis. In a second time, we compare them to the results of R, SAS and SPSS. The objective is to identify important information for predictive analysis i.e. get a simple classification system, get indications on the influence (for the interpretation) and the relevance of variables (statistical significance), and dispose of a variable selection mechanism.

## 2 Dataset

We use the « [alcohol.xls](#) » data file. We want to predict the alcohol type (Kirsch, Mirabelle and Pear) from their composition (butanol, etc.; 6 descriptors). The sample contains 77 instances.

---

<sup>1</sup> [http://en.wikipedia.org/wiki/Linear\\_discriminant\\_analysis](http://en.wikipedia.org/wiki/Linear_discriminant_analysis), [http://en.wikipedia.org/wiki/Discriminant\\_function\\_analysis](http://en.wikipedia.org/wiki/Discriminant_function_analysis)

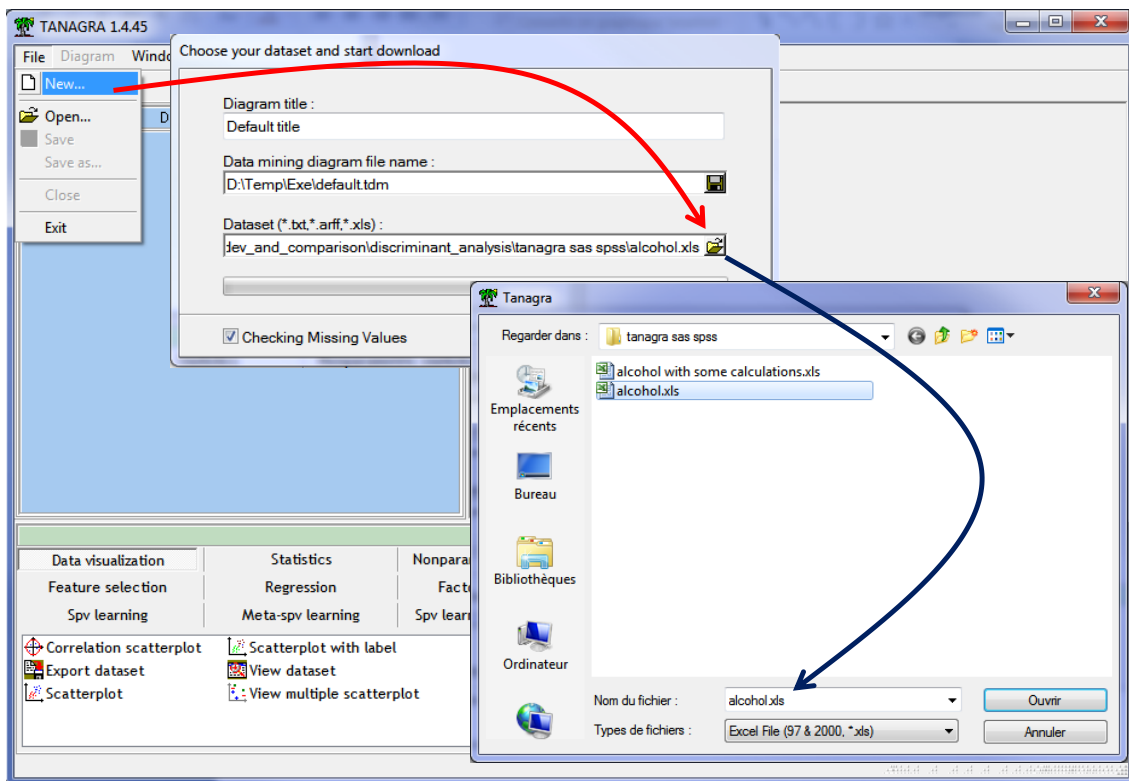
The distinction between the two approaches is clearly defined on the French version of Wikipedia ([http://fr.wikipedia.org/wiki/Analyse\\_discriminante](http://fr.wikipedia.org/wiki/Analyse_discriminante) - [http://fr.wikipedia.org/wiki/Analyse\\_discriminante\\_linéaire](http://fr.wikipedia.org/wiki/Analyse_discriminante_linéaire)).

<sup>2</sup> <http://support.sas.com/documentation/>

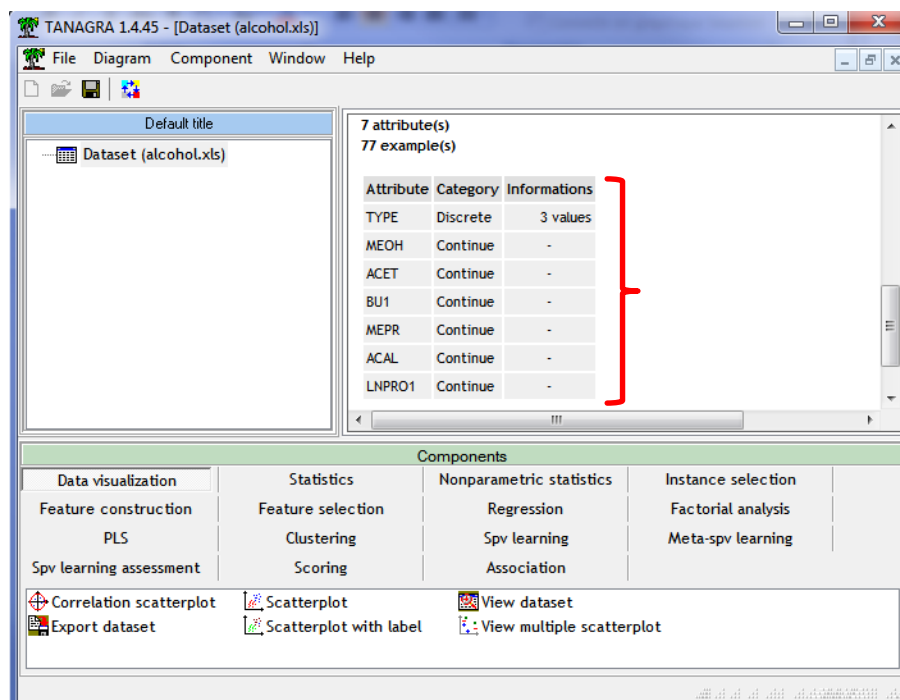
## 3 Linear discriminant analysis under Tanagra

### 3.1 Importing dataset

After we launch Tanagra, we create a new diagram by clicking on the FILE / NEW menu. We select the “alcohol.xls” data file.

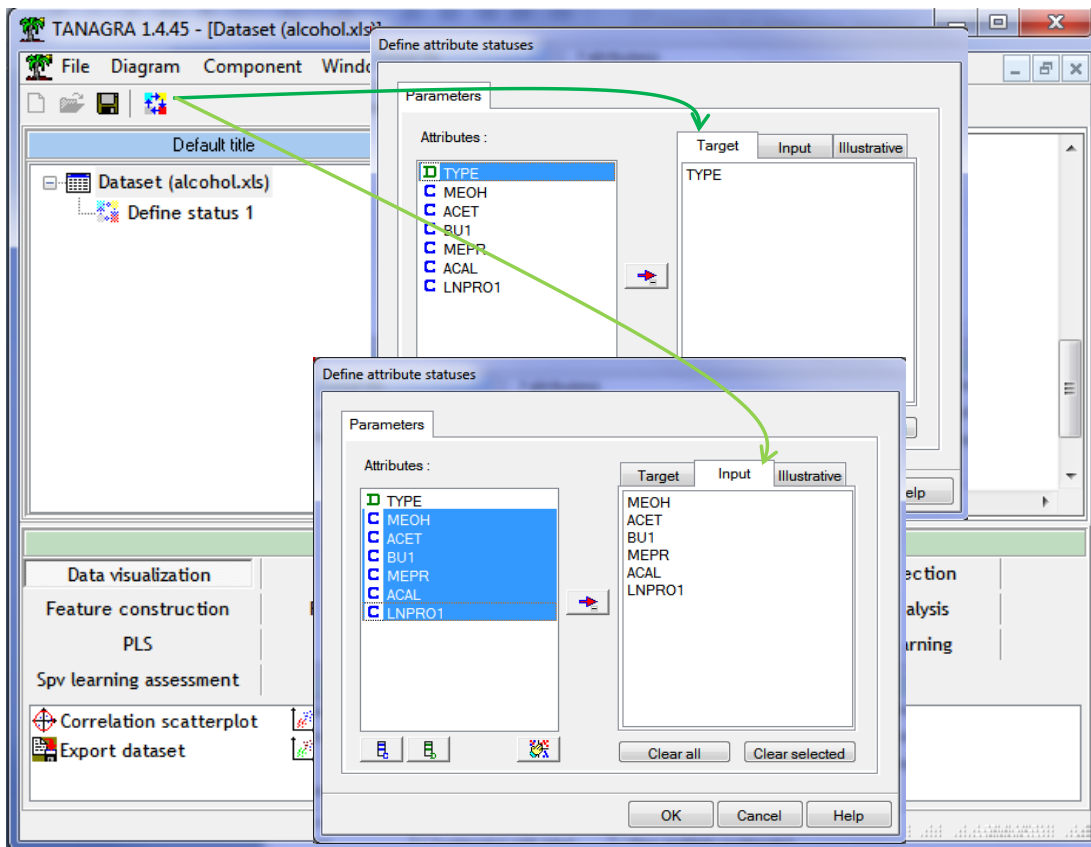


Tanagra shows the number of instances (77) and the number of variables (7, including the class attribute TYPE) which are loaded.

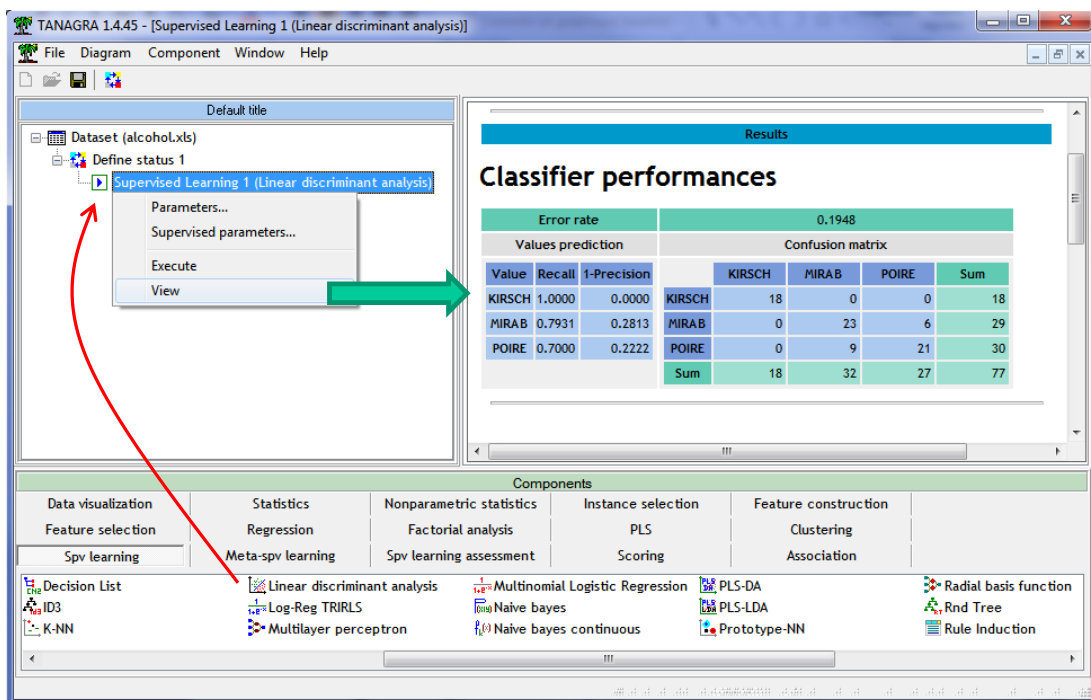


### 3.2 Linear discriminant analysis

The DEFINE STATUS enables to specify the status of the variables in the analysis: TYPE is the target attribute (TARGET); the others are the descriptors (INPUT).



We add the LINEAR DISCRIMINANT ANALYSIS (SPV LEARNING tab) into the diagram. We click on the contextual menu VIEW to obtain the results.



### 3.2.1 Confusion matrix and resubstitution error rate

By applying the classifier to the learning sample, we obtain the confusion matrix:  $(9 + 6) = 15$  instances are misclassified. The resubstitution error rate is  $19.48\% = 15 / 77$ .

Classifier performances							
Error rate			0.1948				
Values prediction			Confusion matrix				
Value	Recall	1-Precision		KIRSCH	MIRAB	POIRE	Sum
KIRSCH	1.0000	0.0000	KIRSCH	18	0	0	18
MIRAB	0.7931	0.2813	MIRAB	0	23	6	29
POIRE	0.7000	0.2222	POIRE	0	9	21	30
			Sum	18	32	27	77

Because we use the same dataset for the learning and the evaluation of the classifier, it is well known that the resubstitution error rate is *often* optimistic.

### 3.2.2 Statistical evaluation of the classifier

The Wilks' lambda statistic is used for the overall evaluation of the model. It evaluates the gap between the group centroids (the groups defined by the target attribute). When its value is close to 0, we expect that we get an efficient classifier. This point of view is closely related to the MANOVA<sup>3</sup>.

#### MANOVA

Stat	Value	p-value
Wilks' Lambda	0.1567	-
Bartlett -- C(12)	132.5414	0
Rao -- F(12, 138)	17.5556	0

In our case,  $\Lambda = 0.1567$ , this is a rather a good value (the worst value is 1). To evaluate the significance of the gap between the centroids, we use the Bartlett's C or the Rao's F transformations (C = 132.5414, d.f. = 12;  $\chi^2$  distribution; F = 17.5556, d.f. 1 = 12 and d.f. 2 = 138; Fisher distribution). At the 5% level, we reject the null hypothesis: the group centroids are significantly different.

By coupling this statistical test with the analysis of the confusion matrix, we understand that the good behavior of the model relies primarily on the situation of KIRSCH which we can detect perfectly. The descriptive analysis will confirm this result.

### 3.2.3 Classification functions

The following table gives the classification functions (Figure 1). They are used when we want classifying instances (KIRSCH, MIRAB or POIRE).

<sup>3</sup> [http://en.wikipedia.org/wiki/Multivariate\\_analysis\\_of\\_variance](http://en.wikipedia.org/wiki/Multivariate_analysis_of_variance)

Attribute	Classification functions		
	KIRSCH	MIRAB	POIRE
MEOH	0.000659	0.015208	0.016407
ACET	0.000445	0.004944	-0.00377
BU1	-0.039342	0.556654	0.553048
MEPR	0.186892	0.035901	0.102271
ACAL	0.039174	-0.160881	-0.127328
LNPRO1	6.378935	4.86937	5.34143
constant	-24.686164	-25.141755	-29.631644

Figure 1 – Classification functions - Model without variable selection

Let us consider a new unlabeled instance  $\omega$  to classify.

MEOH	ACET	BU1	MEPR	ACAL	LNPRO1
707.0	131.0	15.0	28.0	9.0	4.89

We apply the classification function for each class value. For instance, for KIRSCH we have:

$$S(\omega, \text{KIRSCH}) = 0.000659 \times 707 + 0.000445 \times 131 - 0.039342 \times 15 + 0.186892 \times 28 + 0.039174 \times 9 + 6.378935 \times 4.89 - 24.68616 = \mathbf{12.0264}$$

We apply the function for each class, we obtain:

S(.)	KIRSCH	MIRAB	POIRE
	12.0264	17.9763	17.6072

MIRAB has the highest score [ $S(\text{MIRAB}) = 17.9763$ ]. We assign the value MIRAB to this instance.

This classification operation is one of the main goals of the predictive analytics process.

### 3.2.4 Assessing the influence of the variables in the model

The influence of the predictive variables is not the same. The discriminant analysis can measure their contribution in the model. Tanagra shows these effects in the Statistical Evaluation part of the coefficients table (green).

Attribute	Classification functions			Statistical Evaluation			
	KIRSCH	MIRAB	POIRE	Wilks L.	Partial L.	F(2,69)	p-value
MEOH	0.000659	0.015208	0.016407	0.205214	0.763359	10.69497	0.00009
ACET	0.000445	0.004944	-0.00377	0.176705	0.88652	4.41622	0.015677
BU1	-0.039342	0.556654	0.553048	0.213115	0.73506	12.43496	0.000024
MEPR	0.186892	0.035901	0.102271	0.192667	0.813074	7.93155	0.000793
ACAL	0.039174	-0.160881	-0.127328	0.161541	0.969735	1.07671	0.346369
LNPRO1	6.378935	4.86937	5.34143	0.171693	0.912399	3.31241	0.042303
constant	-24.686164	-25.141755	-29.631644				

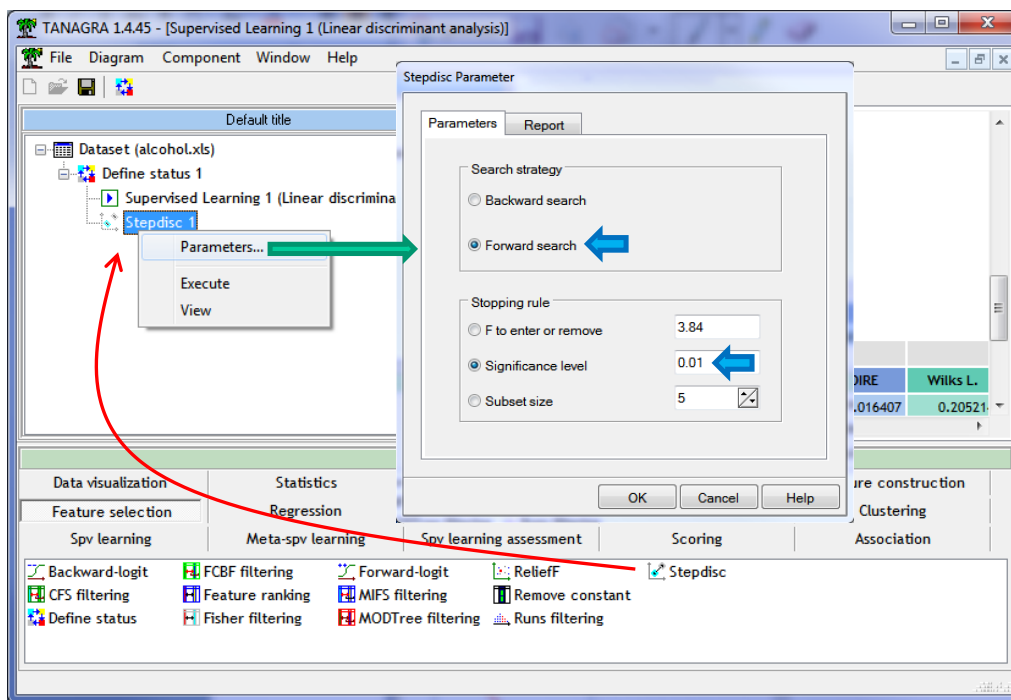
Remember that the overall lambda of the model is  $\Lambda = 0.1567$  (Section 3.2.2).

The first column (“Wilks L.”) indicates the lambda of the model if we remove the variable. For instance, if we remove MEOH from the classifier, the new value of the lambda for the model containing all the predictive variables except MEOH is  $\Lambda_{\{-\text{MEOH}\}} = 0.205214$ . The higher is the difference with the initial value of lambda, the most significant is the variable. “Partial lambda” is the ratio between these values. For instance,  $\text{Partial L.}_{\{-\text{MEOH}\}} = 0.1567 / 0.205214 = 0.763$ .

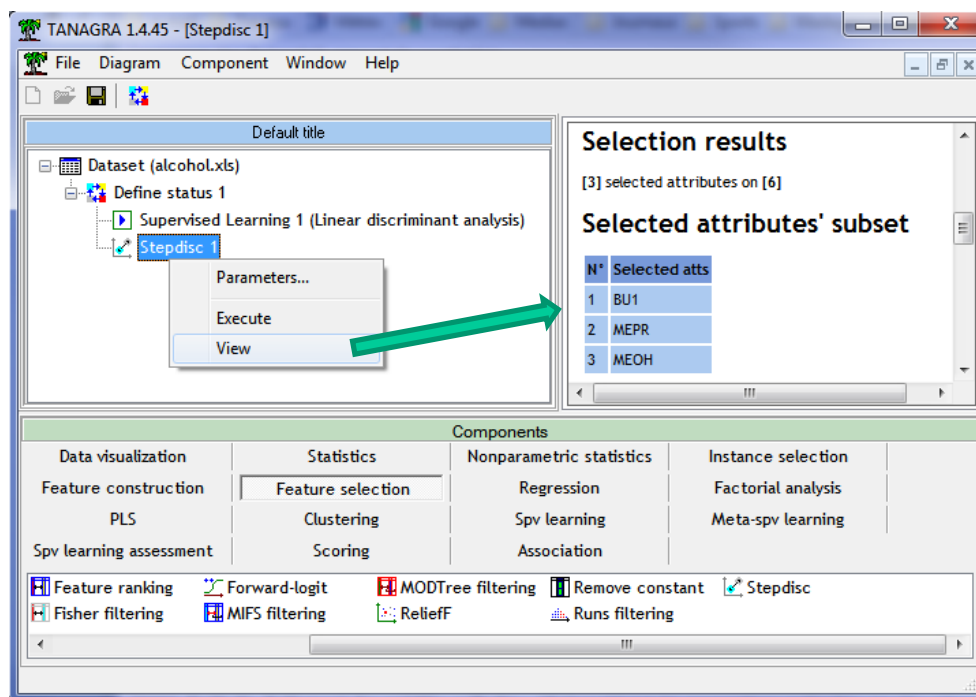
The last two columns are dedicated to the checking of the contribution of each variable. They are based on the comparison of the lambda with and without the variable that we intend to evaluate. For our dataset, only ACAL is not significant at the 5% level ( $p\text{-value}_{\{-ACAL\}} = 0.346369 > 5\%$ ).

### 3.3 Variable selection

We can select the relevant variables by using a stepwise approach.



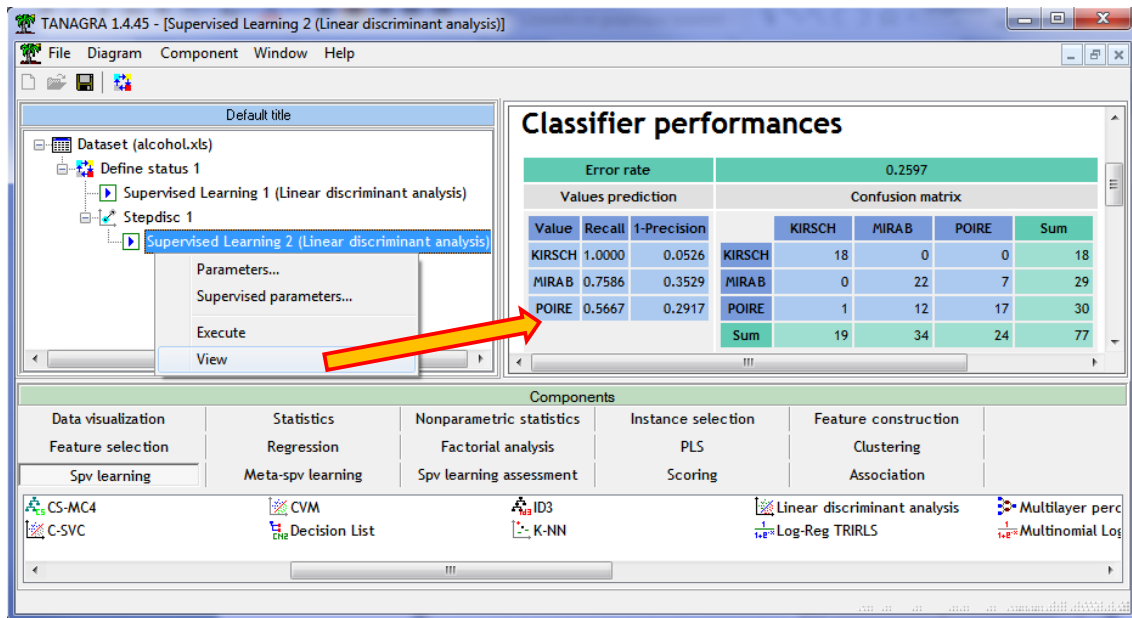
For this, we use the STEPDISC (FEATURE SELECTION tab) component. Tanagra can implement forward and backward strategies. We choose the forward strategy for our dataset. We start with the null model. We add sequentially the most significant variable as long as it is significant at the 1% level.



Because we add STEPDISC after DEFINE STATUS 1 into the diagram, it searches the relevant variables among those defined as INPUT into the preceding component. We click on the VIEW menu to obtain the results, 3 variables are selected: BU1, MEPR, MEOH

*Tanagra provides the details of the search processing. We will describe them later (cf. SAS outputs).*

All we need to do is to add again the LINEAR DISCRIMINANT ANALYSIS component **after** STEPDISC into the diagram. Tanagra performs the learning process on the selected variables only.



The resubstitution error rate is 25.97%. It seems worse than the model with all the predictive attributes (19.48%). But we know that the resubstitution error rate is not a good indicator of the performance of the models. It often favors the complex model incorporating a large number of predictive variables. We must use resampling approach (e.g. cross validation, bootstrap) to obtain a reliable evaluation of the error rate enabling to compare the models.

Here are the coefficients of the new classification functions.

Attribute	Classification functions			Statistical Evaluation			
	KIRSCH	MIRAB	POIRE	Wilks L.	Partial L.	F(2,72)	p-value
BU1	-0.19402	0.432081	0.43508	0.303303	0.66354	18.25443	0
MEPR	0.158883	0.016375	0.080243	0.251168	0.801272	8.92855	0.000344
MEOH	0.007296	0.018626	0.019405	0.248931	0.808472	8.52844	0.000474
constant	-5.235679	-13.841347	-16.982045				

Figure 2 – Classification functions – Model after variable selection

We apply the new model on the instance to classify (section 3.2.3, the unused variables are grayed):

MEOH	ACET	BU1	MEPR	ACAL	LNPRO1
707.0	131.0	15.0	28.0	9.0	4.89

We obtain the following scores. Here also, we assign the instance to the MIRAB class.

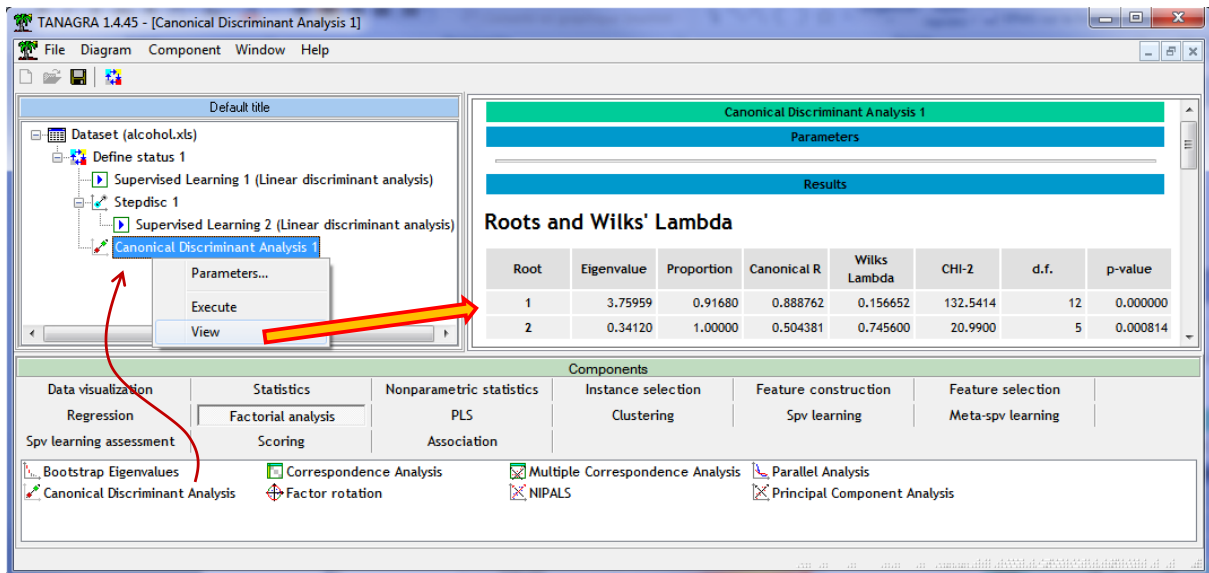
S(.)

KIRSCH	MIRAB	POIRE
1.4610	6.2670	5.5103

### 3.4 Canonical discriminant analysis (factorial discriminant analysis)

Descriptive discriminant analysis is not the main topic of this tutorial. But we present nevertheless the outputs of Tanagra to better understand the results provided of the other tools.

We add the CANONICAL DISCRIMINANT ANALYSIS (FACTORIAL ANALYSIS tab) component into the diagram. Then we click on the VIEW menu to obtain the results.



#### 3.4.1 Eigenvalues table

This table provides the eigenvalues for each factor. We have also the proportion of explained variance. The significance test of each factor is provided.

#### Roots and Wilks' Lambda

Root	Eigenvalue	Proportion	Canonical R	Wilks Lambda	CHI-2	d.f.	p-value
1	3.75959	0.9168	0.888762	0.156652	132.5414	12	0
2	0.3412	1	0.504381	0.7456	20.99	5	0.000814

#### 3.4.2 Canonical coefficients

The raw canonical coefficients enable to calculate the coordinate of the individuals on the factors. The unstandardized coefficients can be applied on the untransformed values of the variables.

Canonical Discriminant Function				
Coefficients	Unstandardized		Standardized	
	Root n1	Root n2	Root n1	Root n2
MEOH	-0.0033821	-0.000571	-0.6811478	-0.1150081
ACET	0.0000465	0.0066574	0.0051822	0.7420478
BU1	-0.1322048	0.0162599	-0.6469978	0.0795743
MEPR	0.0256226	-0.053361	0.3454686	-0.7194653
ACAL	0.0404876	-0.0297884	0.2160076	-0.1589256
LNPRO1	0.2791911	-0.3894401	0.2753705	-0.3841107
constant	1.89673943	3.17877051	-	-

Figure 3 – Raw canonical coefficients

For the instance ω with the following description:

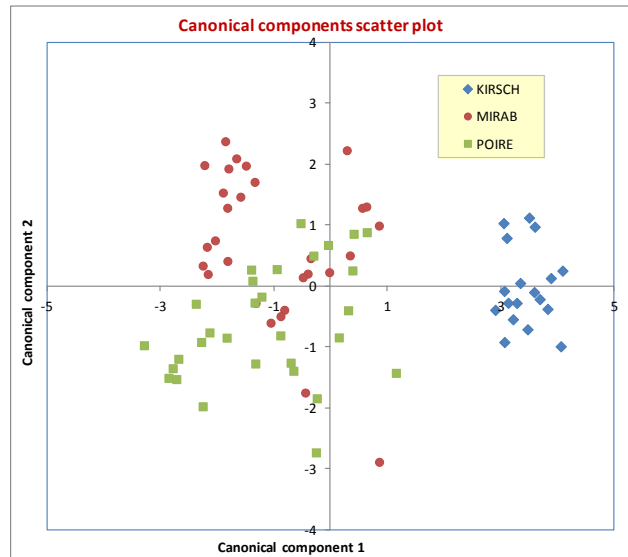


MEOH	ACET	BU1	MEPR	ACAL	LNPRO1
707.0	131.0	15.0	28.0	9.0	4.89

The coordinate on the first factor is,

$$\text{Axe 1} = -0.0033821 \times 707 + 0.0000465 \times 131 - 0.1322048 \times 15 + 0.0256226 \times 28 + 0.0404876 \times 9 + 0.2791911 \times 4.89 + 1.89673943 = \mathbf{-0.0243}$$

On the second factor,  $\text{Axe 2} = -0.000571 \times 707 + \dots - 0.3894401 \times 4.89 + 3.17877051 = \mathbf{0.2245}$



By applying these functions on the individuals of the learning sample, we obtain a representation of the instances into the two first axes. By coloring the points according to its group membership, we can evaluate the class separability. We note that KIRSCH does not overlap of the others. This confirms the results of the confusion matrix obtained earlier (section 3.2.1).

### 3.4.3 Canonical structure

The canonical structure corresponds to the correlation between the variables and the factors. There are different ways to compute this correlation: ignoring the class membership (TOTAL); controlling the class membership (WITHIN); highlighting the class membership (BETWEEN).

## Factor Structure Matrix - Correlations

Root	Root n1			Root n2		
	Total	Within	Between	Total	Within	Between
MEOH	<b>-0.890038</b>	-0.676771	<b>-0.991413</b>	-0.206864	-0.296316	-0.130768
ACET	0.044423	0.021247	0.138243	<b>0.560792</b>	0.505278	<b>0.990398</b>
BU1	<b>-0.939650</b>	-0.787683	<b>-0.997447</b>	-0.118534	-0.187184	-0.071407
MEPR	-0.171747	-0.086008	-0.378992	<b>-0.738951</b>	-0.697111	<b>-0.925400</b>
ACAL	-0.159216	-0.073841	-0.929377	-0.111430	-0.097353	-0.369131
LNPRO1	0.521401	0.272151	0.968743	-0.235265	-0.231331	-0.248066

At a first glance, we observe that: (1) the distinction between KIRSCH and the others relies mainly on MEOH and BU1 on the first axis; (2) the distinction between MIRAB and POIRE on the second axis is due to ACET and MEPR.

These results are consistent with the predictive approach where BU1, MEPR and MEOH were the selected variables after the STEPDISC FORWARD process at 1% level.

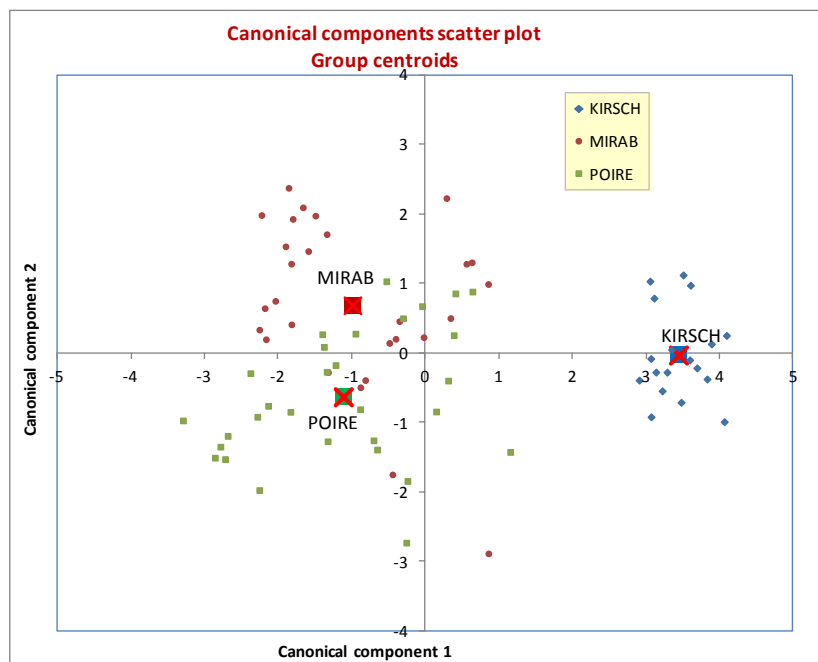
### 3.4.4 Class means on canonical variables

Tanagra provides the means on the axis of the three groups.

#### Group centroids on the canonical variables

TYPE	Root n1	Root n2
KIRSCH	3.439733	-0.031885
MIRAB	-0.981483	0.674773
POIRE	-1.115073	-0.63315
Sq Canonical corr.	0.789898	0.2544

It is especially interesting to visualize the class centroids within the graphical representation of the individuals. We observe that the closest group centroid to the instance to classify with the coordinates  $(-0.0243, 0.2245)$  is MIRAB  $(-0.981483, 0.674773)$ .



### 3.4.5 Classification rule – Method 1

How to assign the individual with the coordinates  $(-0.0243, 0.2245)$  to one of the classes? Visually, we observe that the MIRAB centroid  $(-0.981483, 0.674773)$  is the closest to the instance to classify (with the coordinates:  $-0.0243, 0.2245$ ). But we must confirm this visual impression with calculations.

To obtain a classification consistent with the one of the predictive discriminant analysis, we need – in addition to the group centroids - to know the proportion of the groups into the learning sample.

TYPE	Proportion
KIRSCH	0.233766
MIRAB	0.376623
POIRE	0.389610

For the instance  $\omega$  that we want to classify, we calculate the squared generalized distance to the centroids:

$$D^2(\omega, c) = \sum_{k=1}^K [f_k(\omega) - \mu_{kc}]^2 - 2 \times \ln \pi_c$$

Where  $c$  is one the classes,  $K$  is the number of factors,  $f_k(\omega)$  is the coordinate of the instance on the factor  $k$ ,  $\mu_{kc}$  is the conditional mean of the class  $c$  on the factor  $k$ ,  $\pi_c$  is the proportion of the class  $c$  in the learning sample.

Thus, for the various classes, we obtain:

$$D^2(\omega, \text{KIRSCH}) = (-0.0243 - 3.439733)^2 + (0.2245 + 0.031885)^2 - 2 \ln(0.233766) = \mathbf{14.9723}$$

$$D^2(\omega, \text{MIRAB}) = (-0.0243 + 0.981483)^2 + (0.2245 - 0.376623)^2 - 2 \ln(0.376623) = \mathbf{3.0719}$$

$$D^2(\omega, \text{POIRE}) = (-0.0243 + 1.115073)^2 + (0.2245 + 0.63315)^2 - 2 \ln(0.389610) = \mathbf{3.8106}$$

We assign the individual to the class for which the centroid is the closest. In this case, this is MIRAB since  $D^2(\omega, \text{MIRAB})$  takes the smallest value.

Furthermore, we can calculate the posterior probability of the class membership:

$$P(Y = y/X) = \frac{\exp[-0.5 \times D^2(y)]}{\sum_u \exp[-0.5 \times D^2(u)]}$$

For the instance above:

$$P(Y(\omega) = \text{KIRSCH} / X) = 0.00056 / 0.36459 = \mathbf{0.00154}$$

$$P(Y(\omega) = \text{MIRAB} / X) = 0.21525 / 0.36459 = \mathbf{0.59039}$$

$$P(Y(\omega) = \text{POIRE} / X) = 0.14878 / 0.36459 = \mathbf{0.40807}$$

We assign the instance  $\omega$  to the most likely group.

### 3.4.6 Classification rule - Method 2

By developing the formulas above and by multiplying them by  $-0.5$ , we obtain a linear classification functions based on the factorial coordinates. They are equivalent to the classification function provided by the predictive discriminant analysis, to within a constant which does not depend to the class membership. So, the classification characteristic is exactly the same.

We have:

$$S'(\omega, c) = \sum_{k=1}^K f_k(\omega) \times \mu_{kc} - \frac{1}{2} \sum_{k=1}^K \mu_{kc}^2 + \ln \pi_c$$

For the instance  $\omega$  to classify:

$$S'(\omega, \text{KIRSCH}) = -0.0243 \times 3.439733 + 0.2245 \times (-0.031885) - (3.439733^2 + (-0.031885)^2)/2 + \ln(0.233766) = \mathbf{-7.4606}$$

$$S'(\omega, \text{MIRAB}) = -0.0243 \times (-0.981483) + 0.2245 \times 0.674773 - ((-0.981483)^2 + 0.674773^2)/2 + \ln(0.376623) = \mathbf{-1.5105}$$

$$S'(\omega, \text{POIRE}) = -0.0243 \times (-1.115073) + 0.2245 \times (-0.63315) - ((-1.115073)^2 + (-0.63315)^2)/2 + \ln(0.389610) = \mathbf{-1.8798}$$

We assign the instance to the MIRAB class. The result is necessarily consistent with that of the predictive discriminant analysis (sections 3.2.3 and 3.4.5).

### 3.4.7 Classification rule - Return on the original variables - Method 3

In the previous section, the classification functions are a linear combination of the factors. These last ones are a linear combination of the original variables. So, we can produce a classification function defined on the original predictive variables.

The classification functions on factors are the following:

Score function 1	KIRSCH	MIRAB	POIRE
f1	3.439733	-0.981483	-1.115073
f2	-0.031885	0.674773	-0.633150
Const	-7.369825	-1.685824	-1.764742

For instance, we have for KIRSH:  $S'(KIRSH) = 3.439733 \times f1 - 0.031885 \times f2 - 7.369825$

The factors **F1** and **F2** are linear combinations of original predictive attributes (Figure 3). We can thus produce a new version of the classification functions defined on the predictive attributes  $S'(\cdot)$ :

Score function 2	KIRSCH	MIRAB	POIRE
MEOH	-0.011615	0.002934	0.004133
ACET	-0.000052	0.004447	-0.004267
BU1	-0.455268	0.140729	0.137123
MEPR	0.089836	-0.061155	0.005214
ACAL	0.140216	-0.059838	-0.026286
LNPRO1	0.972760	-0.536805	-0.064744
constant	-0.946902	-1.402493	-5.892384

We can compare them to the classification functions obtained from the predictive linear discriminant analysis  $S(\cdot)$  (Figure 1). We observe that the coefficients of  $S(\cdot)$  and  $S'(\cdot)$  are different, but the gap between the classes is the same for each variable.

S() - Analyse prédictive				S'() - Dédute de l'analyse descriptive			Ecart entre coefficients		
Attribute	KIRSCH	MIRAB	POIRE	KIRSCH	MIRAB	POIRE	KIRSCH	MIRAB	POIRE
MEOH	0.000659	0.015208	0.016407	-0.011615	0.002934	0.004133	0.0123	0.0123	0.0123
ACET	0.000445	0.004944	-0.003770	-0.000052	0.004447	-0.004267	0.0005	0.0005	0.0005
BU1	-0.039342	0.556654	0.553048	-0.455268	0.140729	0.137123	0.4159	0.4159	0.4159
MEPR	0.186892	0.035901	0.102271	0.089836	-0.061155	0.005214	0.0971	0.0971	0.0971
ACAL	0.039174	-0.160881	-0.127328	0.140216	-0.059838	-0.026286	-0.1010	-0.1010	-0.1010
LNPRO1	6.378935	4.869370	5.341430	0.972760	-0.536805	-0.064744	5.4062	5.4062	5.4062
constant	-24.686164	-25.141755	-29.631644	-0.946902	-1.402493	-5.892384	-23.7393	-23.7393	-23.7393

This is the reason for which we have not the same score value for the individual  $\omega$

$$S(\omega, \text{KIRSCH}) \neq S'(\omega, \text{KIRSCH}) ; S(\omega, \text{MIRAB}) \neq S'(\omega, \text{MIRAB}) ; S(\omega, \text{POIRE}) \neq S'(\omega, \text{POIRE})$$

But the difference depends on the individual and not to the class membership.

$$[S(\omega, \text{KIRSCH}) - S'(\omega, \text{KIRSCH})] = [S(\omega, \text{MIRAB}) - S'(\omega, \text{MIRAB})] = [S(\omega, \text{POIRE}) - S'(\omega, \text{POIRE})]$$

In the end, the instances are classified in identical way. This is the most important.

## 4 Discriminant analysis under SAS

### 4.1 Proc DISCRIM

We perform the same analysis under **SAS 9.3** using PROC DISCRIM. We use the following commands.

```

proc discrim data = alcohol;
  class type;
  var MEOH ACET BU1 MEPR ACAL LNPRO1;
  priors proportional;
run;

```

The **PRIORS** option enables to use the proportion measured on the learning set as classes' prior distribution in the learning process.

METHOD = NORMAL (multivariate normal distribution) and POOL = YES (used the pooled covariance matrix) are two important options for which the default values were used.

**Overall description.** SAS provides a description of the problem that we handle. We observe, among others, the proportion of classes into the learning sample.

The SAS System					
The DISCRIM Procedure					
Total Sample Size	77	DF Total		76	
Variables	6	DF Within Classes		74	
Classes	3	DF Between Classes		2	
Number of Observations Read		77			
Number of Observations Used		77			
Class Level Information					
TYPE	Variable Name	Frequency	Weight	Proportion	Prior Probability
KIRSCH	KIRSCH	18	18.0000	0.233766	0.233766
MIRAB	MIRAB	29	29.0000	0.376623	0.376623
POIRE	POIRE	30	30.0000	0.389610	0.389610
Pooled Covariance Matrix Information					
Covariance Matrix Rank		Natural Log of the Determinant of the Covariance Matrix			
6		30.87652			

**Distances between the group centroids.** We have the squared generalized distance between the group centroids.

Generalized Squared Distance to TYPE			
From TYPE	KIRSCH	MIRAB	POIRE
KIRSCH	2.90687	21.99954	22.99300
MIRAB	22.95339	1.95302	3.61372
POIRE	24.01465	3.68153	1.88522

The distance between the centroids of the same group is not null. In addition it is not symmetric!

This is really surprising. This is because the proportion of the groups is used in the formula<sup>4</sup>. For instance, the distance of the KIRSCH with itself is obtained with (see section 3.4.4 for the values of the centroids):

$$D^2(\text{KIRSCH}, \text{KIRSCH}) = (3.439733 - 3.439733)^2 + (-0.031885 + 0.031885)^2 - 2 \times \ln(0.233766) = 2.90687$$

The generalized distance is not symmetric. For instance, between KIRSCH and MIRAB:

$$D^2(\text{KIRSCH}, \text{MIRAB}) = (3.439733 + 0.981483)^2 + (-0.031885 - 0.674773)^2 - 2 \times \ln(0.233766) = 22.95339$$

$$D^2(\text{MIRAB}, \text{KIRSCH}) = (-0.981483 - 3.439733)^2 + (0.674773 + 0.031885)^2 - 2 \times \ln(0.376623) = 21.99954$$

**Classification functions.** SAS provides the same classification functions as Tanagra.

Linear Discriminant Function for TYPE				
Variable	Label	KIRSCH	MIRAB	POIRE
Constant		-24.68616	-25.14175	-29.63164
MEOH	MEOH	0.0006591	0.01521	0.01641
ACET	ACET	0.0004450	0.00494	-0.00377
BU1	BU1	-0.03934	0.55665	0.55305
MEPR	MEPR	0.18689	0.03590	0.10227
ACAL	ACAL	0.03917	-0.16088	-0.12733
LNPRO1	LNPRO1	6.37894	4.86937	5.34143

But SAS does not provide any information about the relevance of the variables.

**Confusion matrix and resubstitution error rate.** Last, SAS provides the confusion matrix.

The DISCRIM Procedure				
Classification Summary for Calibration Data: WORK.ALCOHOL				
Resubstitution Summary using Linear Discriminant Function				
Number of Observations and Percent Classified into TYPE				
From TYPE	KIRSCH	MIRAB	POIRE	Total
KIRSCH	18	0	0	18
	100.00	0.00	0.00	100.00
MIRAB	0	23	6	29
	0.00	79.31	20.69	100.00
POIRE	0	9	21	30
	0.00	30.00	70.00	100.00
Total	18	32	27	77
	23.38	41.56	35.06	100.00
Priors	0.23377	0.37662	0.38961	
Error Count Estimates for TYPE				
	KIRSCH	MIRAB	POIRE	Total
Rate	0.0000	0.2069	0.3000	0.1948
Priors	0.2338	0.3766	0.3896	

<sup>4</sup> <http://www.math.wpi.edu/saspdf/stat/chap25.pdf>

### 4.2 Variable selection with STEPDISC

SAS proposes STEPDISC, a tool for the variable selection which is consistent with the linear discriminant analysis principle. We perform a forward selection at 1% level (METHOD = FORWARD, SLENTRY = 1%):

```
proc stepdisc data = alcohol method = forward slentry = 0.01;
  class type;
  var MEOH ACET BU1 MEPR ACAL LNPRO1;
run;
```

We compare the detailed results with those of Tanagra (Figure 4).

N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(2, 74)	<b>BU1</b> L : 0.299 F : 86.75 p : 0.0000	BU1 L : 0.299 F : 86.75 p : 0.0000	MEOH L : 0.363 F : 64.82 p : 0.0000	LNPRO1 L : 0.771 F : 10.98 p : 0.0001	MEPR L : 0.838 F : 7.16 p : 0.0014	ACET L : 0.918 F : 3.29 p : 0.0429
2	(2, 73)	<b>MEPR</b> L : 0.249 F : 7.34 p : 0.0012	MEPR L : 0.249 F : 7.34 p : 0.0012	MEOH L : 0.251 F : 6.95 p : 0.0017	ACET L : 0.275 F : 3.17 p : 0.0478	LNPRO1 L : 0.276 F : 3.05 p : 0.0535	ACAL L : 0.298 F : 0.18 p : 0.8353
3	(2, 72)	<b>MEOH</b> L : 0.201 F : 8.53 p : 0.0005	MEOH L : 0.201 F : 8.53 p : 0.0005	ACET L : 0.228 F : 3.26 p : 0.0443	LNPRO1 L : 0.232 F : 2.68 p : 0.0751	ACAL L : 0.249 F : 0.05 p : 0.9543	-
4	(2, 71)	-	ACET L : 0.181 F : 4.02 p : 0.0222	LNPRO1 L : 0.182 F : 3.71 p : 0.0294	ACAL L : 0.191 F : 1.95 p : 0.1504	-	-

Figure 4 – Detailed results - Stepdisc at 1% level - Tanagra

Step 1: SAS picks BU1 after having listed the contributions of the candidate variables.

SAS

**The STEPDISC Procedure**  
**Forward Selection: Step 1**

Statistics for Entry, DF = 2, 74					
Variable	Label	R-Square	F Value	Pr > F	Tolerance
MEOH	MEOH	0.6366	64.82	<.0001	1.0000
ACET	ACET	0.0816	3.29	0.0429	1.0000
BU1	BU1	0.7010	86.75	<.0001	1.0000
MEPR	MEPR	0.1622	7.16	0.0014	1.0000
ACAL	ACAL	0.0232	0.88	0.4199	1.0000
LNPRO1	LNPRO1	0.2288	10.98	<.0001	1.0000

Variable BU1 will be entered.

TANAGRA

1	(2, 74)	<b>BU1</b> L : 0.299 F : 86.75 p : 0.0000	BU1 L : 0.299 F : 86.75 p : 0.0000	MEOH L : 0.363 F : 64.82 p : 0.0000	LNPRO1 L : 0.771 F : 10.98 p : 0.0001	MEPR L : 0.838 F : 7.16 p : 0.0014	ACET L : 0.918 F : 3.29 p : 0.0429
---	---------	--	---	--	--	---	---

SAS provides  $R\text{-Square} = 1 - \Lambda$ . For instance,  $R\text{-Square}(\text{MEOH}) = 1 - 0.363 = 0.6366$ . The F statistic is the same as Tanagra, it enables to check the significance of the variable e.g.  $F(\text{MEOH}) = 64.82$ , with the  $p\text{-value}(\text{MEOH}) = 0.0001$ . The tolerance statistic characterizes the redundancy with the variables already included in the model. At the first step, the initial set of selected variables is empty. There is no possible redundancy. It is therefore equal to 1 for all variables.

BU1 has the highest "F Value", and it is significant at the 1% level. It is thus included in the model. In the forward process, this decision is unchangeable. We cannot remove this variable later.

SAS pursues with the assessment of the current model

Multivariate Statistics					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.298992	86.75	2	74	<.0001
Pillai's Trace	0.701008	86.75	2	74	<.0001
Average Squared Canonical Correlation	0.350504				

The [stepdisc documentation](#) explains the reading of the various statistics (Pillai, Average Squared Canonical Correlation) provided by SAS. There is one variable at this step. Thus, the "F Value" is the same as the one of the BU1 variable in the table "Stepdisc Procedure – Step 1" above.

**Step 2:** SAS evaluates the contribution of the remaining variables. The "Partial R-Square" compares the lambda of the models containing or not an additional variable e.g. for MEOH,  $\text{Partial R-Square}(\text{MEOH}) = 1 - 0.251 / 0.299 \approx 0.160$ .

MEPR is the best variable according the "Partial R-Square" (or according to the F Value). It is significant at the 1% level. It is approved.

The STEPDISC Procedure  
Forward Selection: Step 2

Statistics for Entry, DF = 2, 73					
Variable	Label	Partial R-Square	F Value	Pr > F	Tolerance
MEOH	MEOH	0.1600	6.95	0.0017	0.3658
ACET	ACET	0.0799	3.17	0.0478	0.9981
MEPR	MEPR	0.1674	7.34	0.0012	0.9046
ACAL	ACAL	0.0049	0.18	0.8353	0.9573
LNPRO1	LNPRO1	0.0771	3.05	0.0535	0.8325

Variable MEPR will be entered.

TANAGRA

	MEPR	MEPR	MEOH	ACET	LNPRO1	ACAL
2	(2, 73)	L : 0.249 F : 7.34 p : 0.0012	L : 0.249 F : 7.34 p : 0.0012	L : 0.251 F : 6.95 p : 0.0017	L : 0.275 F : 3.17 p : 0.0478	L : 0.276 F : 3.05 p : 0.0535

We note also, even if this is not taken into account for the selection, that MEPR is weakly correlated with the already introduced variable because its tolerance is 0.9046. It means that the squared correlation between BU1 and MEPR is  $r^2 = 1 - 0.9046 = 0.0954$ .

For the overall evaluation of the model with 2 (BU1, MEPR) predictive variables, we have:



Variable(s) That Have Been Entered					
BU1	MEPR				

Multivariate Statistics					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.248931	36.66	4	146	<.0001
Pillai's Trace	0.851978	27.46	4	148	<.0001
Average Squared Canonical Correlation	0.425989				

SAS continues until it is no more possible to add variables. It then produces a table summarizing the process (Figure 5). The values (Wilks lambda, F Value) are consistent with those of Tanagra.

SAS

TANAGRA

Forward Selection Summary										
Step	Number In	Entered	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	BU1	BU1	0.7010	86.75	<.0001	0.29899190	<.0001	0.35050405	<.0001
2	2	MEPR	MEPR	0.1674	7.34	0.0012	0.24893122	<.0001	0.42598921	<.0001
3	3	MEOH	MEOH	0.1915	8.53	0.0005	0.20125392	<.0001	0.45385985	<.0001

N°	d.f	Best
1	(2, 74)	BU1 L : 0.299 F : 86.75 p : 0.0000
2	(2, 73)	MEPR L : 0.249 F : 7.34 p : 0.0012
3	(2, 72)	MEOH L : 0.201 F : 8.53 p : 0.0005

Figure 5 – Summary of the selection process - Stepdisc forward at 1% level

The selected variables are: BU1, MEPR and MEOH.

## 5 Discriminant Analysis under R – lda() [MASS package]

We use the **lda()** procedure of the « MASS » package to perform the linear discriminant analysis. This package is automatically installed with R.

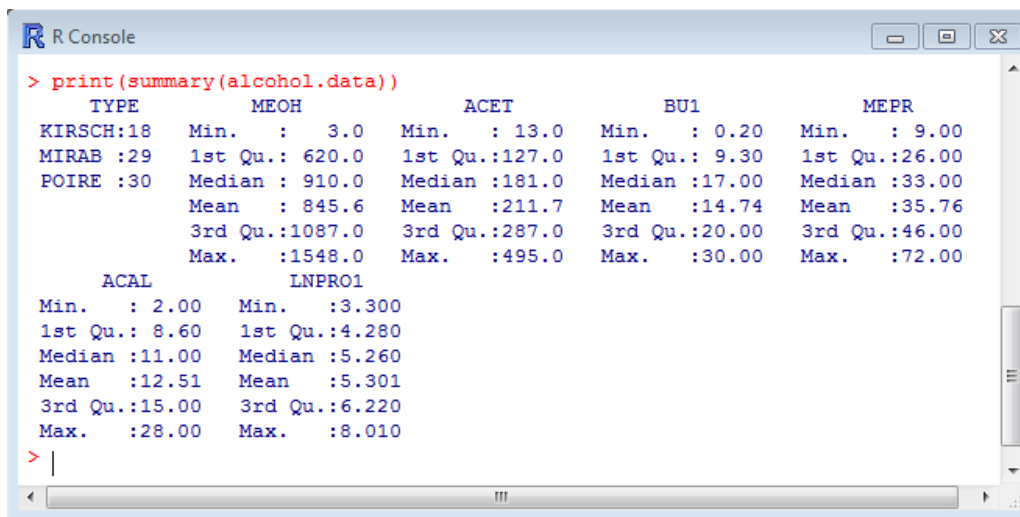
### 5.1 Importing the dataset

The **read.xlsx()** command of the package “xlsx”<sup>5</sup> enables to read a data file in the Excel format (XLS or XLSX). The **summary()** command describes shortly the variables of the dataset.

```
library(xlsx)
#sheetIndex: number of the sheet to read
#header: the first row corresponds to the name of the variables
alcohol.data <- read.xlsx(file="alcohol.xls", sheetIndex=1, header=T)
print(summary(alcohol.data))
```

We obtain the main features of the variables.

<sup>5</sup> <http://cran.r-project.org/web/packages/xlsx/index.html>



```

> print(summary(alcohol.data))
  TYPE      MEOH      ACET      BU1      MEPR
KIRSCH:18  Min.   : 3.0   Min.   : 13.0  Min.   : 0.20  Min.   : 9.00
MIRAB :29  1st Qu.: 620.0  1st Qu.:127.0  1st Qu.: 9.30  1st Qu.:26.00
POIRE :30  Median : 910.0  Median :181.0  Median :17.00  Median :33.00
          Mean  : 845.6   Mean  :211.7   Mean  :14.74   Mean  :35.76
          3rd Qu.:1087.0  3rd Qu.:287.0  3rd Qu.:20.00  3rd Qu.:46.00
          Max.  :1548.0   Max.  :495.0   Max.  :30.00   Max.  :72.00

  ACAL      LNPRO1
Min.   : 2.00   Min.   :3.300
1st Qu.: 8.60   1st Qu.:4.280
Median :11.00   Median :5.260
Mean  :12.51   Mean  :5.301
3rd Qu.:15.00   3rd Qu.:6.220
Max.  :28.00   Max.  :8.010
  
```

TYPE is the categorical attribute (3 values) that we want to explain.

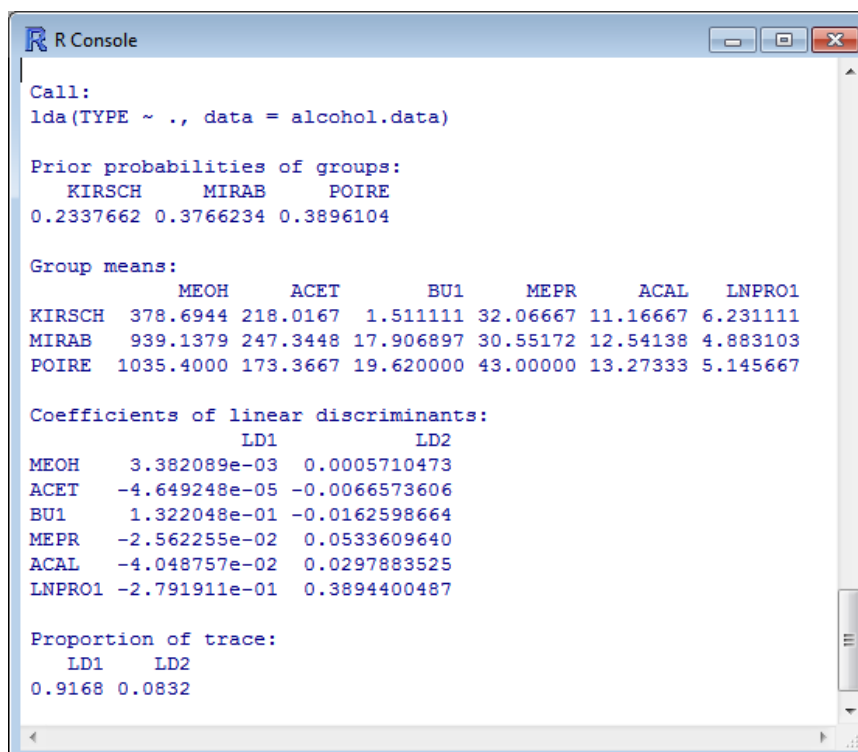
## 5.2 Using the lda() procedure

We load the MASS<sup>6</sup> package. We start the lda() procedure on the learning sample.

```

#linear discriminant analysis
library(MASS)
alcohol.lda <- lda(TYPE ~ ., data = alcohol.data)
print(alcohol.lda)
  
```

lda() provides: the proportions of the classes; the conditional mean of the variables according to the classes; the unstandardized coefficients of the canonical functions (Section 3.4.2), without the constant term; the proportion of variance explained by each factor.



```

Call:
lda(TYPE ~ ., data = alcohol.data)

Prior probabilities of groups:
  KIRSCH  MIRAB  POIRE
0.2337662 0.3766234 0.3896104

Group means:
      MEOH      ACET      BU1      MEPR      ACAL      LNPRO1
KIRSCH 378.6944 218.0167  1.511111 32.06667 11.16667  6.231111
MIRAB  939.1379 247.3448 17.906897 30.55172 12.54138  4.883103
POIRE 1035.4000 173.3667 19.620000 43.00000 13.27333  5.145667

Coefficients of linear discriminants:
          LD1          LD2
MEOH    3.382089e-03  0.0005710473
ACET   -4.649248e-05 -0.0066573606
BU1     1.322048e-01 -0.0162598664
MEPR   -2.562255e-02  0.0533609640
ACAL   -4.048757e-02  0.0297883525
LNPRO1 -2.791911e-01  0.3894400487

Proportion of trace:
  LD1  LD2
0.9168 0.0832
  
```

<sup>6</sup> <http://cran.r-project.org/web/packages/MASS/index.html>

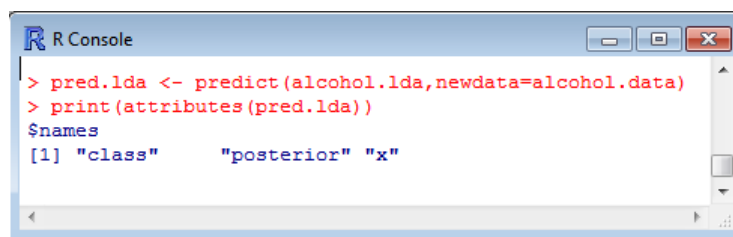
In principle, **lda()** seems only intended to the descriptive analysis. But as we have seen above, the connection between the descriptive analysis and the predictive analysis is strong (section 3.4.5). Thus, the **predict()** command enables to assign individual to the classes. The classification properties are identical to those of Tanagra or SAS i.e. each individual is assigned to the same group whatever the tool used.

### 5.3 Prediction

The **predict()** enables to apply the classifier on the individuals.

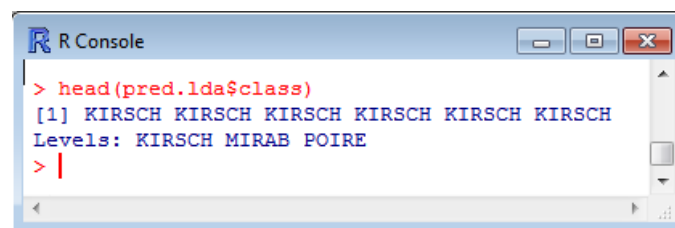
```
#prediction on the training set
pred.lda <- predict(alcohol.lda,newdata=alcohol.data)
print(attributes(pred.lda))
```

The object has 3 main properties: 'class', 'posterior' and 'x'.



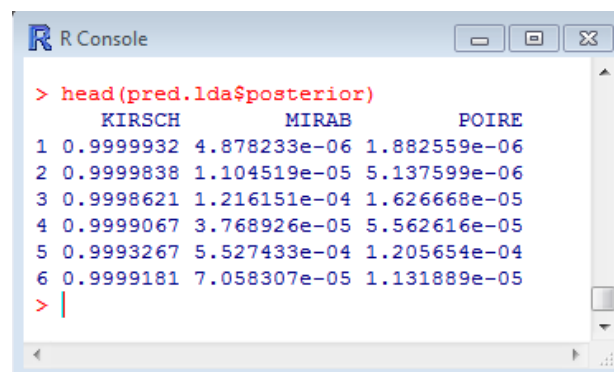
```
R Console
> pred.lda <- predict(alcohol.lda,newdata=alcohol.data)
> print(attributes(pred.lda))
$names
[1] "class"      "posterior"  "x"
```

'class' is a vector, its length is  $n = 77$ . It provides the predicted value for each individual of the sample. For instance, we show here the predicted values for the first 6 individuals.



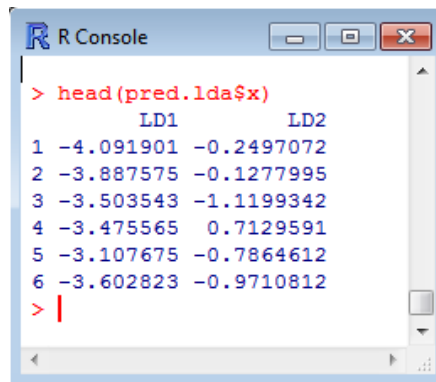
```
R Console
> head(pred.lda$class)
[1] KIRSCH KIRSCH KIRSCH KIRSCH KIRSCH
Levels: KIRSCH MIRAB POIRE
> |
```

'posterior' is a matrix with  $n = 77$  rows and 3 columns (because the class attribute has 3 values). Each column corresponds to the posterior probability of the class membership. For the first 6 instances, we have the following values:



```
R Console
> head(pred.lda$posterior)
      KIRSCH      MIRAB      POIRE
1 0.9999932 4.878233e-06 1.882559e-06
2 0.9999838 1.104519e-05 5.137599e-06
3 0.9998621 1.216151e-04 1.626668e-05
4 0.9999067 3.768926e-05 5.562616e-05
5 0.9993267 5.527433e-04 1.205654e-04
6 0.9999181 7.058307e-05 1.131889e-05
> |
```

'x' provides the canonical coordinates of the instances. This is a matrix with 77 rows and 2 columns (because we have two factors). Here are the values for the first 6 instances:



```

> head(pred.lda$x)
      LD1      LD2
1 -4.091901 -0.2497072
2 -3.887575 -0.1277995
3 -3.503543 -1.1199342
4 -3.475565  0.7129591
5 -3.107675 -0.7864612
6 -3.602823 -0.9710812
> |

```

## 5.4 Confusion matrix

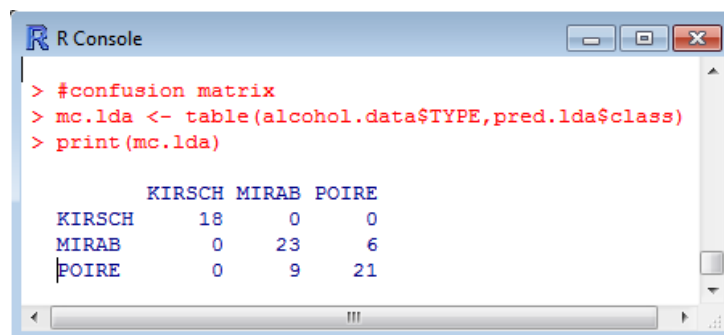
To obtain the confusion matrix, we create a contingency table from the observed values of the class attribute and the predictions of the model.

```

#confusion matrix
mc.lda <- table(alcohol.data$TYPE,pred.lda$class)
print(mc.lda)

```

We have the same matrix as Tanagra (section 3.2.1) and SAS (section 4.1).



```

> #confusion matrix
> mc.lda <- table(alcohol.data$TYPE,pred.lda$class)
> print(mc.lda)
      KIRSCH MIRAB POIRE
KIRSCH    18     0     0
MIRAB     0    23     6
POIRE     0     9    21

```

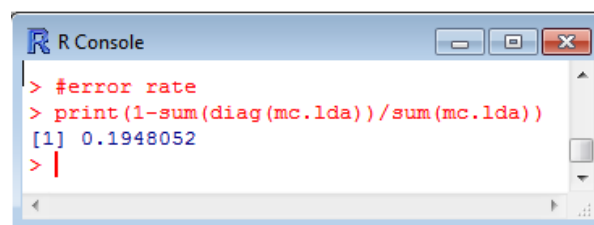
We can compute the error rate,

```

#error rate
print(1-sum(diag(mc.lda))/sum(mc.lda))

```

We have 19.48%



```

> #error rate
> print(1-sum(diag(mc.lda))/sum(mc.lda))
[1] 0.1948052
> |

```

## 5.5 Variable selection

We use the procedure **greedy.wilks()** of the package « klaR »<sup>7</sup> for the variable selection.

```

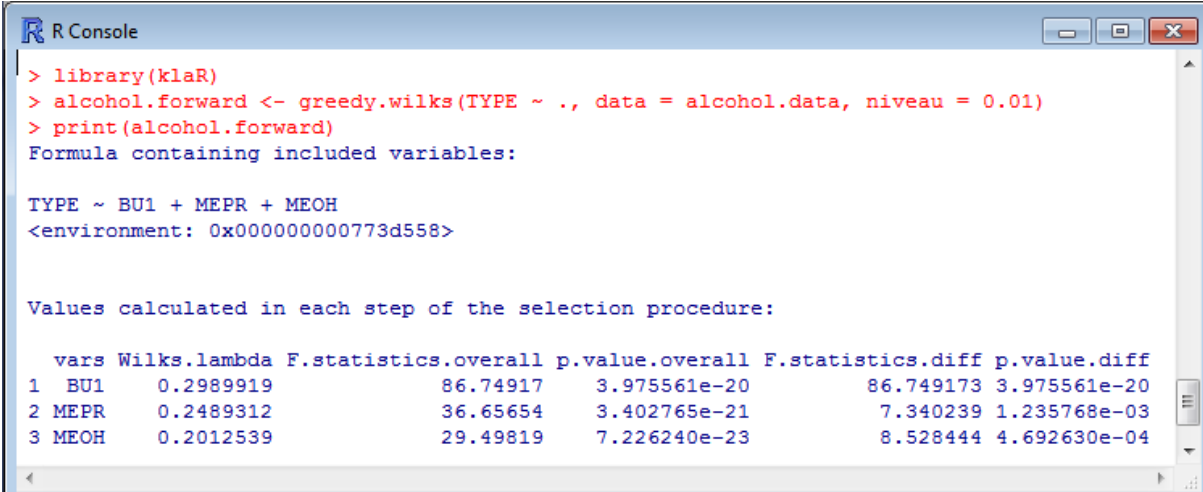
#variable selection
library(klaR)
alcohol.forward <- greedy.wilks(TYPE ~ .,data=alcohol.data, niveau = 0.01)

```

<sup>7</sup> <http://cran.r-project.org/web/packages/klaR/index.html>

```
print(alcohol.forward)
```

At each step, R provides the F statistic for the additional variable (**F.statistics.diff**), and the F statistic for the model with the current set of selected variables (**F.statistics.overall**).



```
> library(klaR)
> alcohol.forward <- greedy.wilks(TYPE ~ ., data = alcohol.data, niveau = 0.01)
> print(alcohol.forward)
Formula containing included variables:

TYPE ~ BU1 + MEPR + MEOH
<environment: 0x000000000773d558>

Values calculated in each step of the selection procedure:

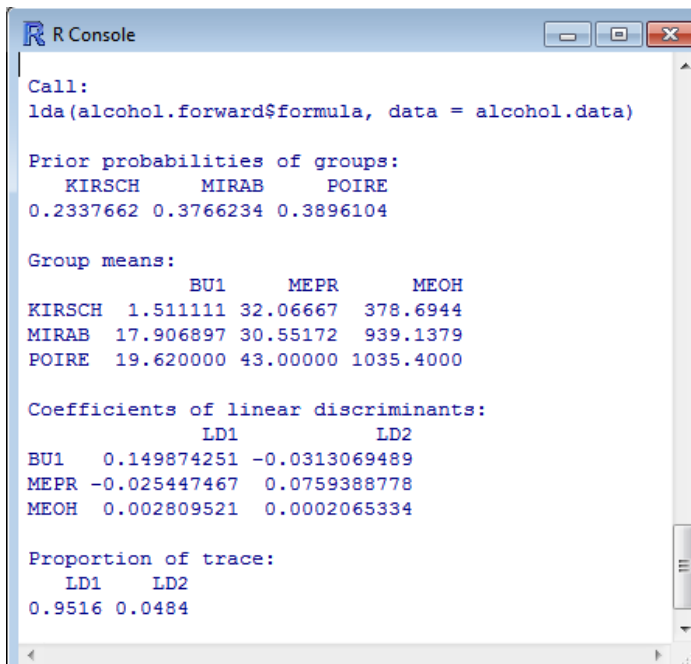
  vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff p.value.diff
1 BU1    0.2989919      86.74917      3.975561e-20      86.749173 3.975561e-20
2 MEPR   0.2489312      36.65654      3.402765e-21      7.340239 1.235768e-03
3 MEOH   0.2012539      29.49819      7.226240e-23      8.528444 4.692630e-04
```

The process is consistent with those of Tanagra and SAS (Figure 5).

We perform a new analysis on the variables selected by the **greedy.wilks()** procedure. This last one provides directly the formula with only the relevant variables. This feature is really useful if the initial number of candidate variables is very large.

```
#2nd model after variable selection
alcohol.lda.fwd <- lda(alcohol.forward$formula, data = alcohol.data)
print(alcohol.lda.fwd)
```

We obtain a new version of the model.



```
Call:
lda(alcohol.forward$formula, data = alcohol.data)

Prior probabilities of groups:
  KIRSCH  MIRAB  POIRE
0.2337662 0.3766234 0.3896104

Group means:
      BU1  MEPR  MEOH
KIRSCH 1.511111 32.06667 378.6944
MIRAB 17.906897 30.55172 939.1379
POIRE 19.620000 43.00000 1035.4000

Coefficients of linear discriminants:
      LD1  LD2
BU1  0.149874251 -0.0313069489
MEPR -0.025447467  0.0759388778
MEOH  0.002809521  0.0002065334

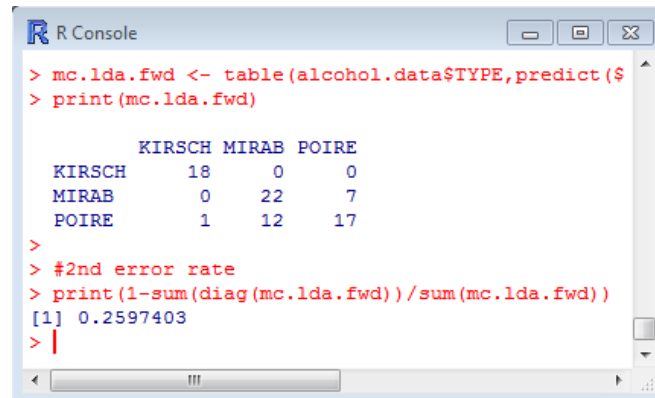
Proportion of trace:
  LD1  LD2
0.9516 0.0484
```

We apply this model on the sample to obtain the corresponding confusion matrix and error rate.

```
#2nd confusion matrix
mc.lda.fwd <- table(alcohol.data$TYPE,predict(alcohol.lda.fwd,newdata=alcohol.data)$class)
print(mc.lda.fwd)

#2nd error rate
print(1-sum(diag(mc.lda.fwd))/sum(mc.lda.fwd))
```

We obtain:

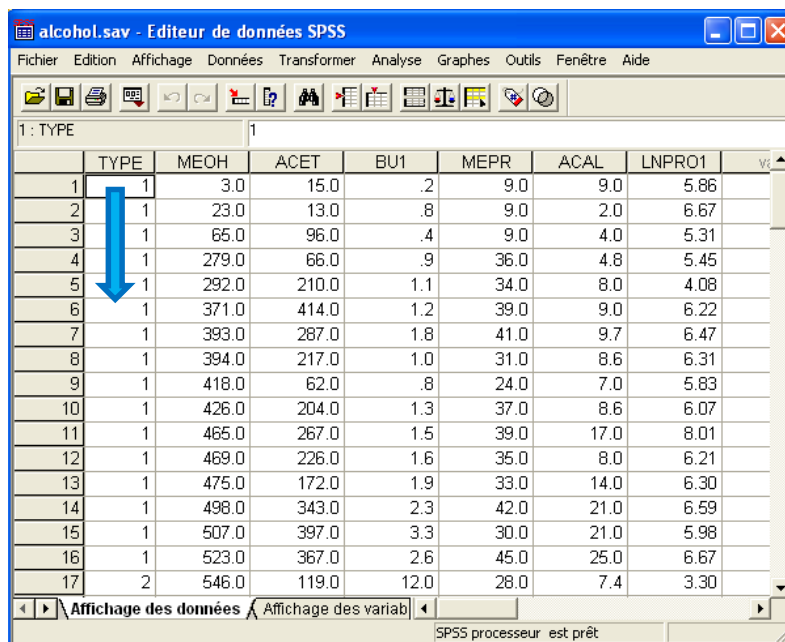


```
R Console
> mc.lda.fwd <- table(alcohol.data$TYPE,predict($
> print(mc.lda.fwd)

      KIRSCH MIRAB POIRE
KIRSCH    18     0     0
MIRAB     0    22     7
POIRE     1    12    17
>
> #2nd error rate
> print(1-sum(diag(mc.lda.fwd))/sum(mc.lda.fwd))
[1] 0.2597403
> |
```

## 6 Discriminant analysis under SPSS

We use the French version of **SPSS 12.0.1**. We import the “alcohol.xls” data file. We transform TYPE in a numerical attribute before (KIRSCH = 1, MIRAB = 2, POIRE = 3)<sup>8</sup>.

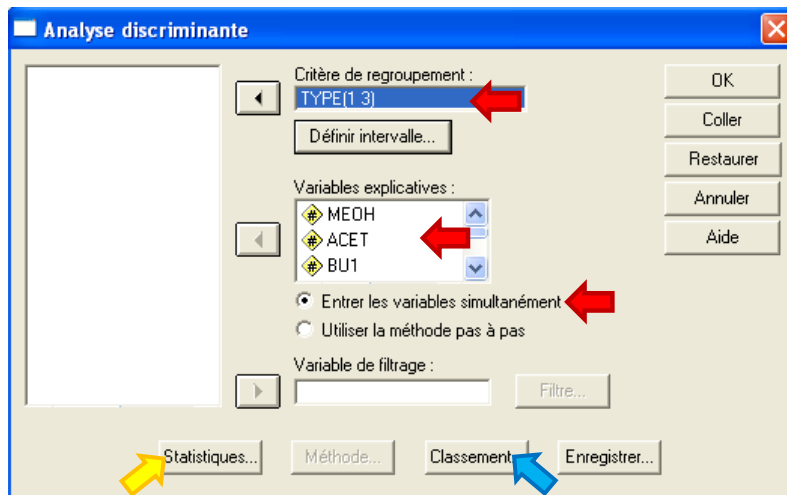


1	TYPE	MEOH	ACET	BU1	MEPR	ACAL	LNPRO1	vr
1	1	3.0	15.0	.2	9.0	9.0	5.86	
2	1	23.0	13.0	.8	9.0	2.0	6.67	
3	1	65.0	96.0	.4	9.0	4.0	5.31	
4	1	279.0	66.0	.9	36.0	4.8	5.45	
5	1	292.0	210.0	1.1	34.0	8.0	4.08	
6	1	371.0	414.0	1.2	39.0	9.0	6.22	
7	1	393.0	287.0	1.8	41.0	9.7	6.47	
8	1	394.0	217.0	1.0	31.0	8.6	6.31	
9	1	418.0	62.0	.8	24.0	7.0	5.83	
10	1	426.0	204.0	1.3	37.0	8.6	6.07	
11	1	465.0	267.0	1.5	39.0	17.0	8.01	
12	1	469.0	226.0	1.6	35.0	8.0	6.21	
13	1	475.0	172.0	1.9	33.0	14.0	6.30	
14	1	498.0	343.0	2.3	42.0	21.0	6.59	
15	1	507.0	397.0	3.3	30.0	21.0	5.98	
16	1	523.0	367.0	2.6	45.0	25.0	6.67	
17	2	546.0	119.0	12.0	28.0	7.4	3.30	

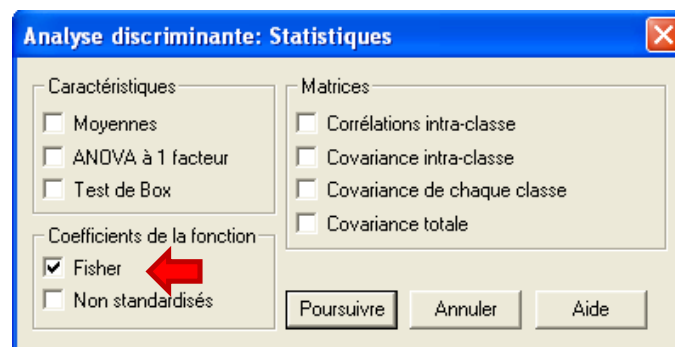
### 6.1 Construction of the model including all the predictive variables

We click on the **ANALYSE / CLASSIFICATION / ANALYSE DISCRIMINANTE** menu. We set into the settings dialog:

<sup>8</sup> See also “[Annotated SPSS Output – Discriminant Analysis](#)”.

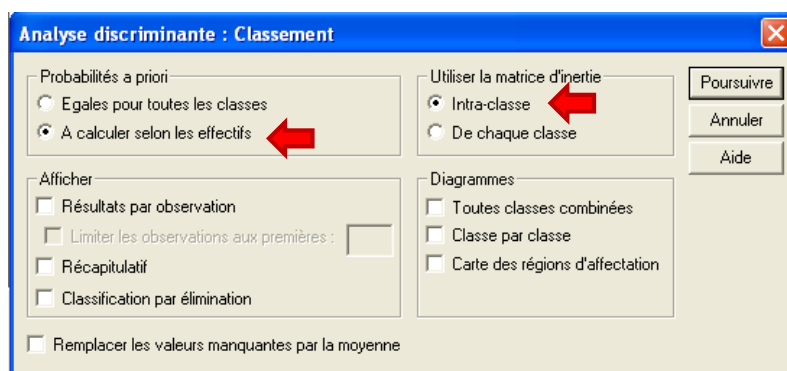


We set TYPE in « **Critère de regroupement** », the range (Définir intervalle) of the value is MIN = 1 to MAX = 3. All the others variables are predictive variables « **Variables explicatives** ». At the moment, we do not perform a variable selection « **Entrer les variables simultanément** ».



Then, we click on the « **Statistiques** » button. We ask the “Fisher” classification function.

From the initial dialog box, we click on the « **Classements** » button. We want to estimate the prior probabilities of the groups from the proportion measured on the learning sample. We use the pooled covariance matrix.

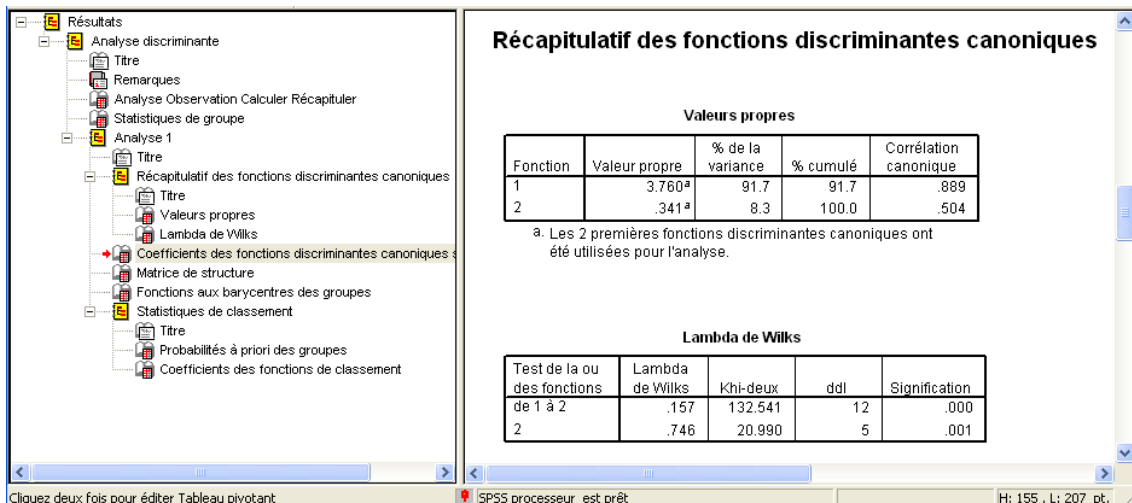


We validate these settings. A new window describing the results of the calculations appears.

SPSS mixes the outputs of the canonical analysis and predictive analysis. This is not a problem. Indeed, the two approaches can converge as we shown previously. We must know simply discern the appropriate information in the outputs.

Firstly, we have the results in the **descriptive point of view**.

We have the eigenvalues related to the factors and the tests of significance.



Then, we have: (1) the canonical functions; (2) the canonical structure; (3) the conditional centroids.

**Coefficients des fonctions discriminantes canoniques standardisées**

	Fonction	
	1	2
MEOH	.6811478	.1150081
ACET	-.0051822	-.7420478
BU1	.6469978	-.0795743
MEPR	-.3454686	.7194653
ACAL	-.2160076	.1589256
LNPRO1	-.2753705	.3841107

**(1)**

**Matrice de structure**

	Fonction	
	1	2
BU1	.787683*	.187184
MEOH	.676771*	.296316
LNPRO1	-.272151*	.231331
MEPR	.086008	.697111*
ACET	-.021247	-.505278*
ACAL	.073841	.097353*

Les corrélations intra-groupes combinés entre variables discriminantes et les variables des fonctions discriminantes canoniques standardisées sont ordonnées par tailles absolues des corrélations à l'intérieur de la fonction.

\*. Plus grande corrélation absolue entre chaque variable et une fonction discriminante quelconque.

**Fonctions aux barycentres des groupes**

TYPE	Fonction	
	1	2
1	-3.439733	.031885
2	1.115073	.633150
3	.981483	-.674773

Fonctions discriminantes canoniques non standardisées évaluées aux moyennes des groupes

**(3)**



Secondly, we have the results in the **predictive point of view**.

We have, among others, the classification functions: the “Fisher’s Linear Discriminant Functions” according to SPSS.

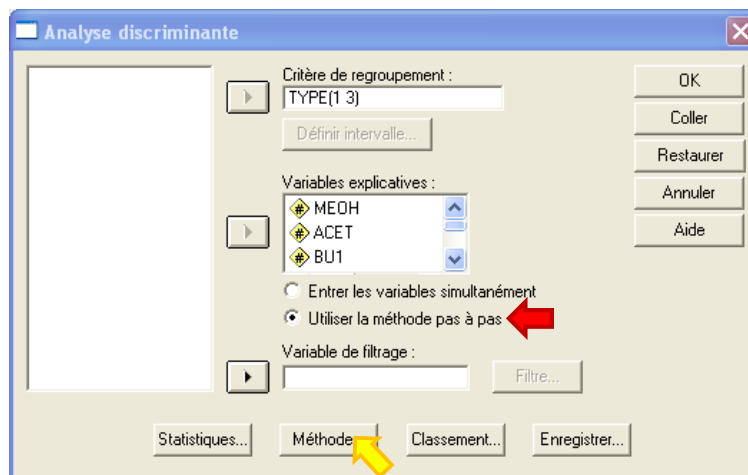
**Coefficients des fonctions de classement**

	TYPE		
	1	2	3
MEOH	.001	.016	.015
ACET	.000	-.004	.005
BU1	-.039	.553	.557
MEPR	.187	.102	.036
ACAL	.039	-.127	-.161
LNPRO1	6.379	5.341	4.869
(Constante)	-24.686	-29.632	-25.142

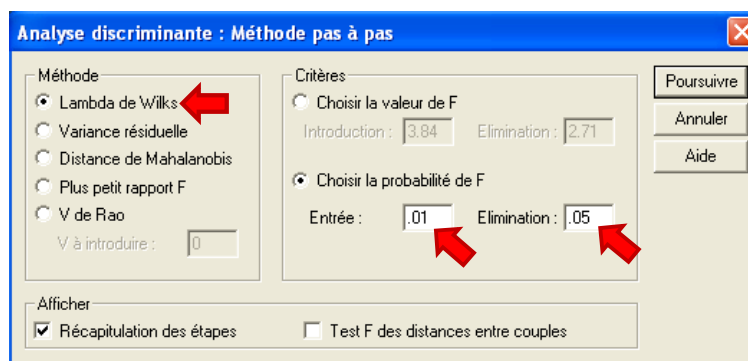
Fonctions discriminantes linéaires de Fisher

## 6.2 Variable selection

To perform a variable selection, we must restart the analysis by modifying the treatment of the independent variables. We select the stepwise approach. We click on the “Méthode” button.



Only the bidirectional approach is available i.e. at each step, we check if the adding of a variable does not imply the removing of another already selected variable. Two significance levels enable to guide the process: 0.01 for the adding, 0.05 for the removing. We select the Wilks’ lambda (Méthode) to obtain consistent results with the other tools.



A table summarizes the process. It is very similar to the table provided by the procedure `greedy.wilks()` (KlaR package) under R.

Variables introduites/éliminées<sup>a,b,c,d</sup>

Pas	Introduite	Lambda de Wilks							
		Statistique	ddl1	ddl2	ddl3	F exact			
						Statistique	ddl1	ddl2	Signification
1	BU1	.299	1	2	74.000	86.749	2	74.000	.000
2	MEPR	.249	2	2	74.000	36.657	4	146.000	.000
3	MEOH	.201	3	2	74.000	29.498	6	144.000	.000

Then we have more details on the various versions of the models. As we seen above, the tolerance criterion enables to evaluate the redundancy between the selected variables. It varies between 0 and 1. The higher is the value, the weaker is the redundancy between the variables.

Variables de l'analyse

Pas		Tolérance	Signification du F pour éliminer	Lambda de Wilks
1	BU1	1.000	.000	
2	BU1	.899	.000	.838
	MEPR	.899	.001	.299
3	BU1	.811	.000	.303
	MEPR	.830	.000	.251
	MEOH	.781	.000	.249

In the last column, we have the Wilks' lambda of the model if we remove a variable.

Last, a table provides a full detailed description of the process. We observe for instance that BU1 is the best variable in the first step. It proposes the weaker value of the Wilks' lambda (0.299). In the second step, MEPR is the most interesting (0.249). Etc.

Variables absentes de l'analyse

Pas		Tolérance	Tolérance minimale	Signification du F pour introduire	Lambda de Wilks
0	MEOH	1.000	1.000	.000	.363
	ACET	1.000	1.000	.043	.918
	BU1	1.000	1.000	.000	.299
	MEPR	1.000	1.000	.001	.838
	ACAL	1.000	1.000	.420	.977
	LNPRO1	1.000	1.000	.000	.771
1	MEOH	.846	.846	.002	.251
	ACET	1.000	1.000	.048	.275
	MEPR	.899	.899	.001	.249
	ACAL	.975	.975	.835	.298
	LNPRO1	.996	.996	.053	.276
2	MEOH	.781	.781	.000	.201
	ACET	.995	.895	.044	.228
	ACAL	.966	.886	.954	.249
	LNPRO1	.996	.896	.075	.232
3	ACET	.940	.738	.022	.181
	ACAL	.801	.647	.150	.191
	LNPRO1	.966	.758	.029	.182

## 7 Conclusion

Discriminant analysis is an attractive method. It is available in almost all statistical software. In this tutorial, we tried to highlight the similarities and the differences between the outputs of Tanagra, R, SAS, and SPSS software. The main conclusion is that, if the presentation is not always the same, ultimately we have exactly the same results. This is the most important.