1 Objectif

Analyse discriminante linéaire – Comparaison des logiciels Tanagra, SAS, SPSS et R.

L'analyse discriminante linéaire est une technique prédictive très populaire auprès des statisticiens et des data scientists. En effet, elle cumule des qualités intéressantes : elle est relativement rapide sur des grandes bases ; elle sait traiter directement les problèmes multi-classes (les variables cibles à plus de 2 modalités) ; elle génère un classifieur linéaire, facile à interpréter ; elle est robuste et plutôt stable, même appliquée sur des petites bases ; elle dispose d'un mécanisme intégré de sélection de variables. Pour ma part, je l'apprécie énormément car elle se prête à de multiples interprétations (probabilistes, géométriques), éclairant ainsi différentes facettes de l'apprentissage supervisé.

L'analyse discriminante est à la fois une méthode prédictive et descriptive. Dans le premiers cas, on se réfère souvent à l'analyse discriminante linéaire. On cherche à produire un système de classement qui permet d'affecter un groupe à un individu selon ses caractéristiques (les valeurs prises par les variables indépendantes). Dans le second cas, on parle d'analyse factorielle discriminante. L'objectif est de produire un système de représentation synthétique où l'on distinguerait au mieux les groupes, en fournissant les éléments d'interprétation permettant de comprendre ce qui les réunit ou les différencie. Les finalités ne sont donc pas intrinsèquement identiques même si, en creusant en peu, on se rend compte que les deux approches se rejoignent. Certaines références bibliographiques entretiennent d'ailleurs la confusion en les présentant dans un cadre unique.

Tanagra opère clairement la distinction en proposant les deux méthodes dans des composants différents : LINEAR DISCRIMINANT ANALYSIS (onglet SPV LEARNING) pour la prédiction, CANONICAL DISCRIMINANT ANALYSIS (onglet FACTORIAL ANALYSIS) pour la description. Il en est de même pour SAS avec respectivement les procédures DISCRIM et CANDISC. D'autres en revanche les associent. C'est le cas des logiciels SPSS et R, mélangeant des résultats de teneur différente. Pour les spécialistes qui savent distinguer les éléments importants selon le contexte, cet amalgame n'est pas un problème. Pour les néophytes, c'est un peu plus problématique. On peut être perturbé par des informations qui ne semblent pas en rapport direct avec les finalités de l'étude.

Dans ce tutoriel, nous détaillons[1] dans un premier temps les sorties de Tanagra concernant l'analyse discriminante linéaire. Dans un second temps, nous les mettrons en parallèle avec les

[1] Voir Références (section 8) pour les supports de cours associés.

résultats fournis par les logiciels R, SAS et SPSS. L'objectif est de discerner les informations importantes pour l'analyse prédictive c.-à-d. obtenir un système simple d'affectation des individus aux classes, avoir des indications sur le rôle (interprétation) et la pertinence (significativité) des variables, et disposer d'un mécanisme de sélection de variables.

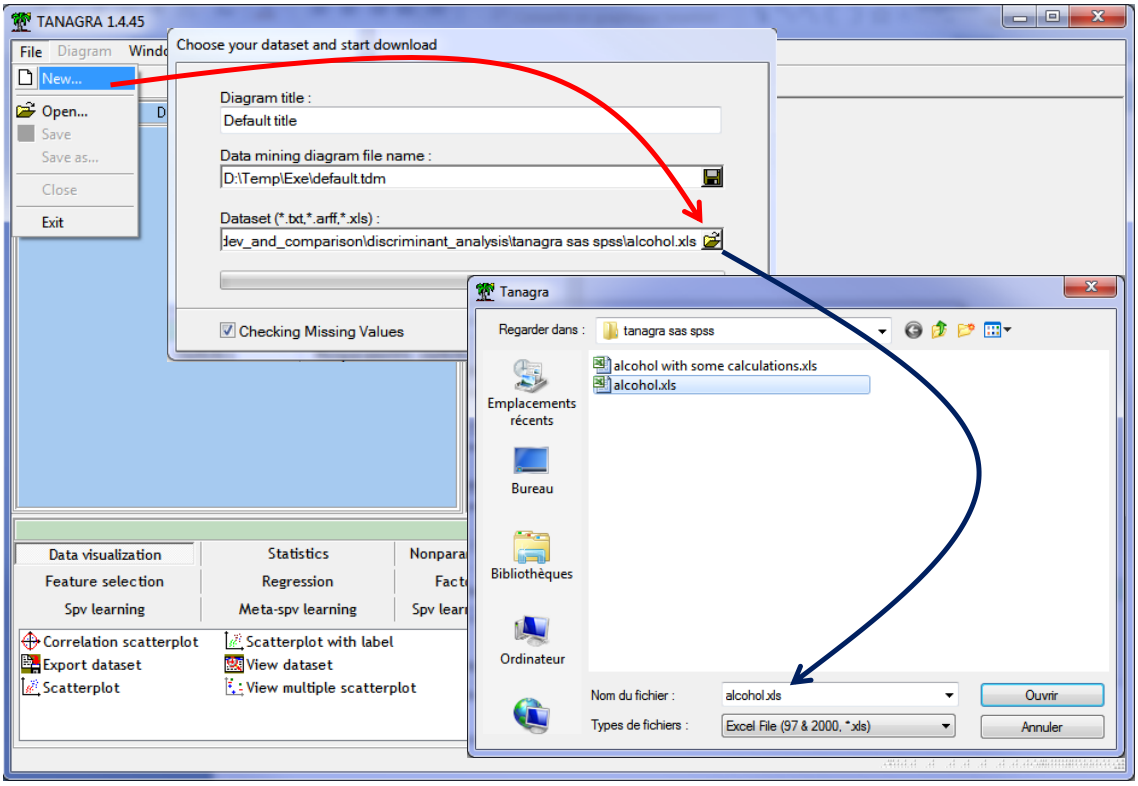
2 Données

Nous traitons le fichier « alcohol.xls ». Nous souhaitons prédire le type (kirsch, mirabelle, poire ; 3 modalités) d'un liquide alcoolisé contenu dans un verre selon sa composition (butanol, méthanol, etc. ; 6 variables). Nous disposons de 77 observations.

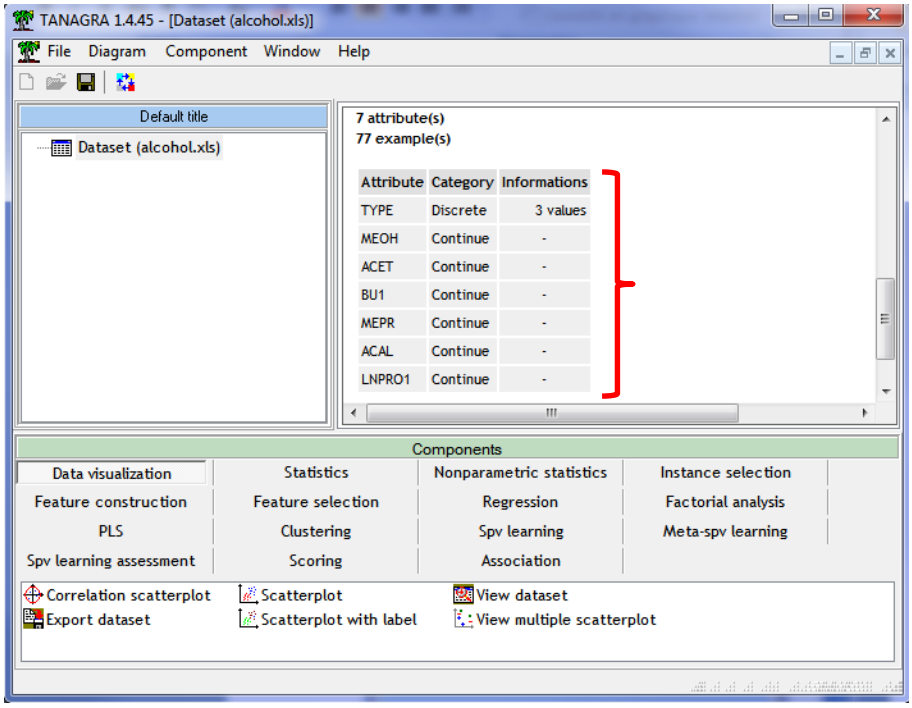
3 Analyse discriminante avec Tanagra

3.1 Importation des données

Après avoir démarré Tanagra, nous créons un nouveau diagramme en actionnant le menu FILE / NEW. Nous sélectionnons le fichier '**alcohol.xls**'.

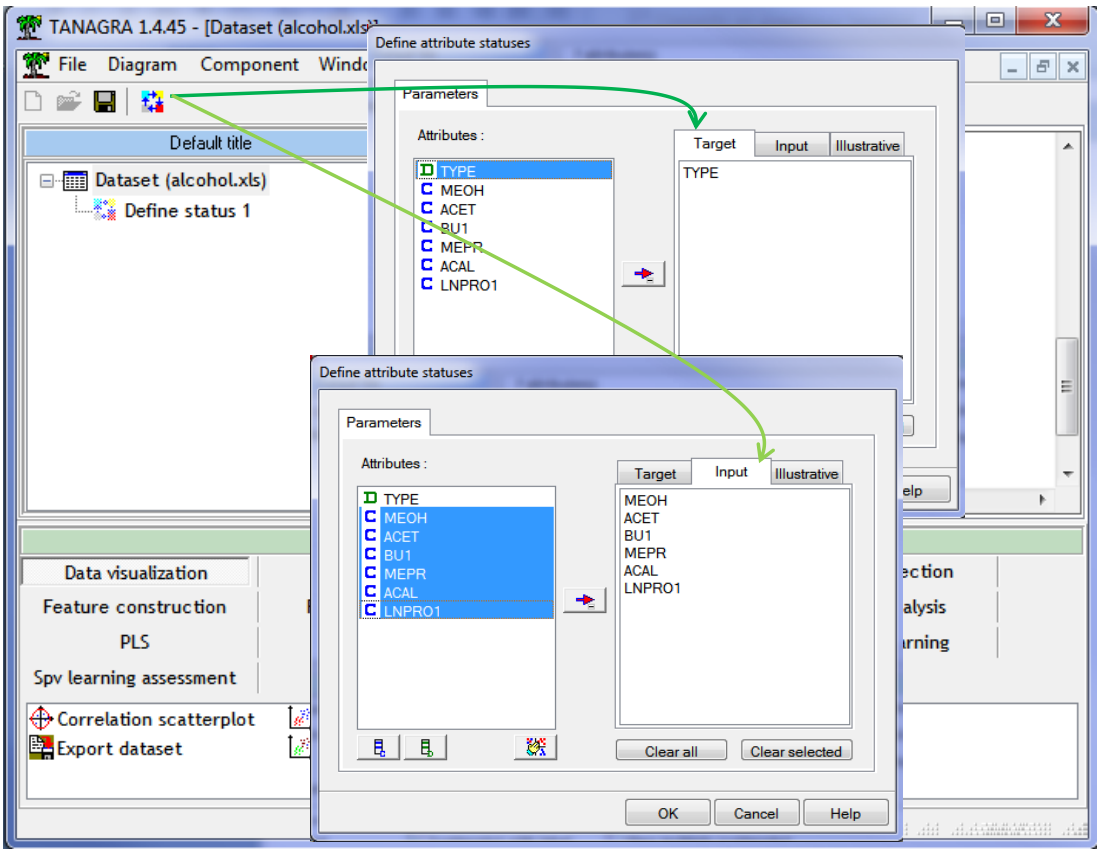


Nous validons. Tanagra indique le nombre d'observations (77 individus) et de variables (7, en comptant la variable cible TYPE) chargées.

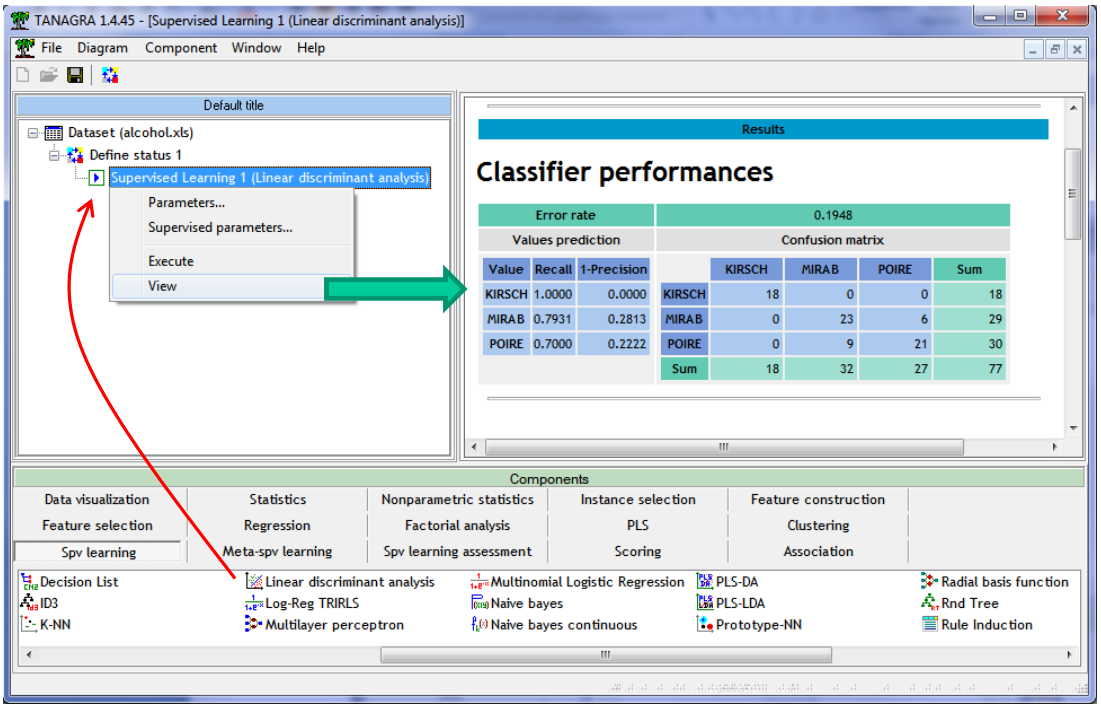


3.2 Analyse discriminante linéaire

Nous utilisons le composant DEFINE STATUS pour indiquer le rôle des variables : TYPE est la cible (TARGET), les autres (MEOH … LNPRO1) sont les prédictives (INPUT).



Nous plaçons le composant LINEAR DISCRIMINANT ANALYSIS (onglet SPV LEARNING). Il n'y a pas de paramétrage à effectuer. Nous cliquons sur VIEW pour accéder aux résultats.



3.2.1 Matrice de confusion et erreur en resubstitution

En appliquant le modèle sur les données d'apprentissage, nous obtenons la matrice de confusion en resubstitution et le taux d'erreur associé : 19.48% c.-à-d. (9 + 6) = 15 individus mal classés sur 77.

Classifier performances

Error rate	0.1948
Values prediction	Confusion matrix

Value	Recall	1-Precision
KIRSCH	1.0000	0.0000
MIRAB	0.7931	0.2813
POIRE	0.7000	0.2222

	KIRSCH	MIRAB	POIRE	Sum
KIRSCH	18	0	0	18
MIRAB	0	23	6	29
POIRE	0	9	21	30
Sum	18	32	27	77

L'échantillon d'apprentissage étant juge et partie dans ce processus, on sait que ce taux est souvent trop optimiste.

3.2.2 Evaluation globale du modèle

Le lambda de Wilks est l'indicateur privilégié pour l'évaluation statistique du modèle. Il indique dans quelle mesure les centres de classes sont distincts les uns des autres dans l'espace de représentation. Il varie entre 0 et 1 : vers 0, le modèle sera bon parce que les nuages sont bien

distincts ; vers 1, les nuages sont confondus, il est difficile de discerner les individus appartenant à des classes différentes. Il s'agit en réalité d'un test d'analyse de variance multivariée, on parle de MANOVA[2].

MANOVA

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

Dans notre cas, Λ = 0.1567, c'est plutôt bon signe. L'indicateur n'étant pas tabulé, nous nous tournons vers les transformations de Bartlett (C = 132.5414, avec ddl = 12 ; loi du χ^2) ou de Rao (F = 17.5556, avec ddl 1 = 12 et ddl 2 = 138 ; loi de Fisher) pour statuer sur la significativité des écarts. Tous deux aboutissent à la même conclusion : à 5%, on rejette l'hypothèse selon laquelle les centres de classes sont confondus.

En couplant la lecture du test avec celle de la matrice de confusion, nous comprenons que la bonne tenue du modèle tient surtout à la situation de KIRSCH qui se distingue parfaitement des autres modalités de la variable cible. L'analyse descriptive nous le confirmera.

3.2.3 Fonctions de classement

Le tableau suivant décrit les fonctions de classement (partie bleue). Elles servent à affecter les groupes (KIRSCH, POIRE ou MIRAB) aux individus à classer.

	Classification functions		
Attribute	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 - Fonctions de classement - Modèle complet

Mettons que nous avons un nouvel individu ω avec la description suivante :

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

Nous calculons le score de chaque classe. Par exemple, pour KIRSH, nous avons :

[2] « Comparaison de populations – Tests paramétriques », septembre 2009, chapitre 7.

S(ω, KIRSCH) = 0.000659 x 707 + 0.000445 x 131 – 0.039342 x 15 + 0.186892 x 28 + 0.039174 x 9 + 6.378935 x 4.89 – 24.68616 = **12.0264**

Nous faisons de même pour les autres modalités, nous obtenons finalement :

	KIRSCH	MIRAB	POIRE
S(.)	12.0264	17.9763	17.6072

La modalité MIRAB présente le score le plus élevé [**S(MIRAB) = 17.9763**]. La meilleure prédiction pour cet individu sera donc « TYPE = MIRAB ».

Ce processus de classement est une des principales finalités du data mining prédictif.

3.2.4 Evaluation individuelle des variables prédictives

Les variables ne contribuent pas de manière identique. Il se peut même que certaines aient une faible influence la prédiction. L'analyse discriminante nous propose des outils d'évaluation de la contribution des variables dans la partie « Statistical Evaluation » du tableau (couleur verte).

	Classification functions			Statistical Evaluation			
Attribute	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	-			

Rappelons que le Λ du modèle dans sa globalité est de 0.1567 (Section 3.2.2).

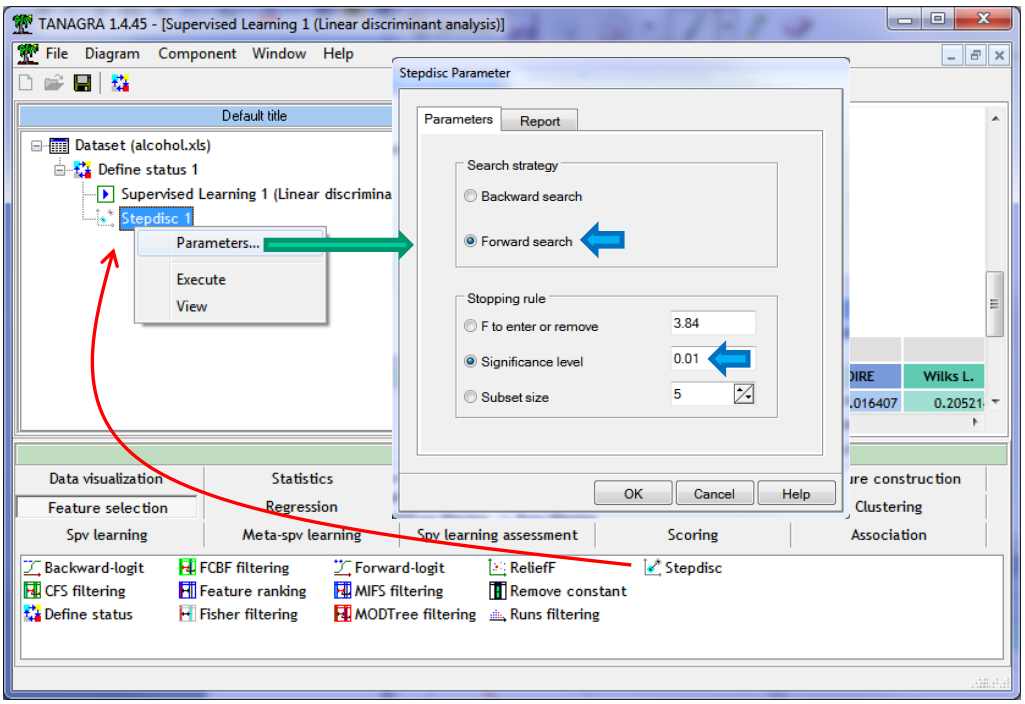
- La première colonne « Wilks L. » indique le Λ du modèle que l'on obtiendrait si l'on retirait la variable. Par exemple, si on décide de reconstruire le modèle après avoir retiré MEOH (et donc en conservant toutes les autres), nous obtiendrons $\Lambda_{\{-MEOH\}}$ = 0.205214. Plus élevé est le différentiel avec le lambda initial, plus la variable est importante puisque son retrait entraîne une forte dégradation (rappelons que plus le lambda est élevé, plus mauvais est le modèle prédictif).
- « Partial L. » indique le rapport entre les lambdas. Ainsi, Partial $L._{\{-MEOH\}}$ = 0.1567 / 0.205214 = 0.763.
- Les deux dernières colonnes servent à tester la contribution des variables. Le test est toujours basé sur la confrontation des lambdas. La statistique suit une loi de Fisher. Les probabilités critiques (p-value) permettent de statuer. Dans notre exemple, seule ACAL ne semble pas contribuer significativement à 5% ($p\text{-value}_{\{-ACAL\}}$ = 0.346369 > 5%).

3.3 Sélection de variables

Nous ne pouvons pas laisser le modèle en l'état. Il faut mettre en œuvre un processus de sélection de manière à intégrer uniquement les variables pertinentes.

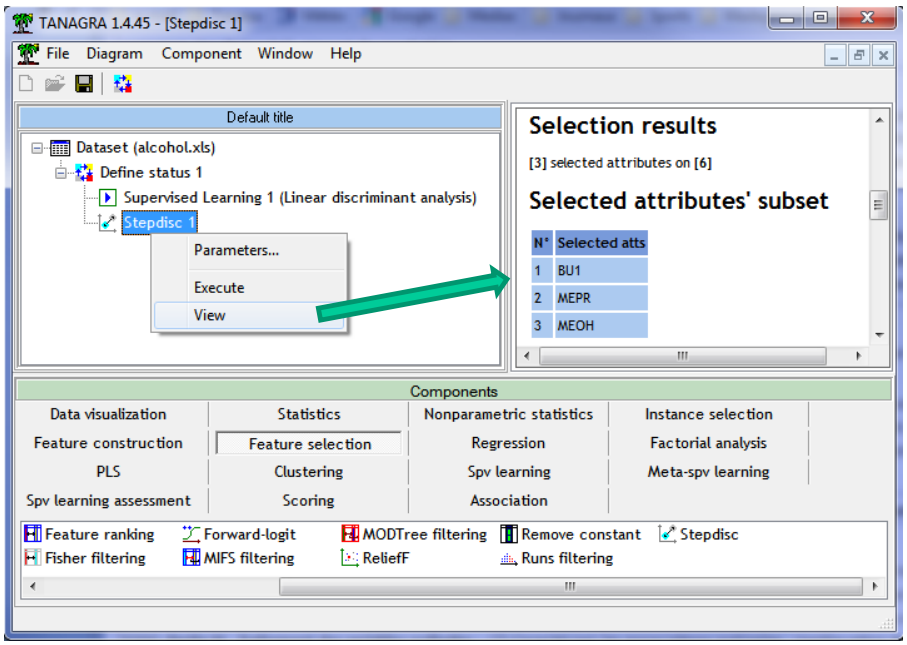
La tentation est grande de retirer en bloc l'ensemble des variables non significatives. Cette attitude est à bannir. En effet, elles ne sont pas indépendantes. Il se peut qu'une variable, du fait de sa forte liaison avec une autre, devienne significative lorsque l'on retire cette dernière.

Sélection des variables. Pour contourner cet écueil, on s'appuie sur les procédures pas à pas pour sélectionner les bonnes variables. Soit en adoptant une approche descendante (backward) : on commence avec le modèle complet, on retranche unes à unes les variables – en retirant la moins contributive à chaque étape – jusqu'à ce que les restantes soient toutes significatives. Soit à l'aide d'une approche ascendante (forward) : on part du modèle trivial réduit à la constante, on rajoute unes à unes les variables – en choisissant la plus contributive à chaque étape – tant qu'elles sont significatives.



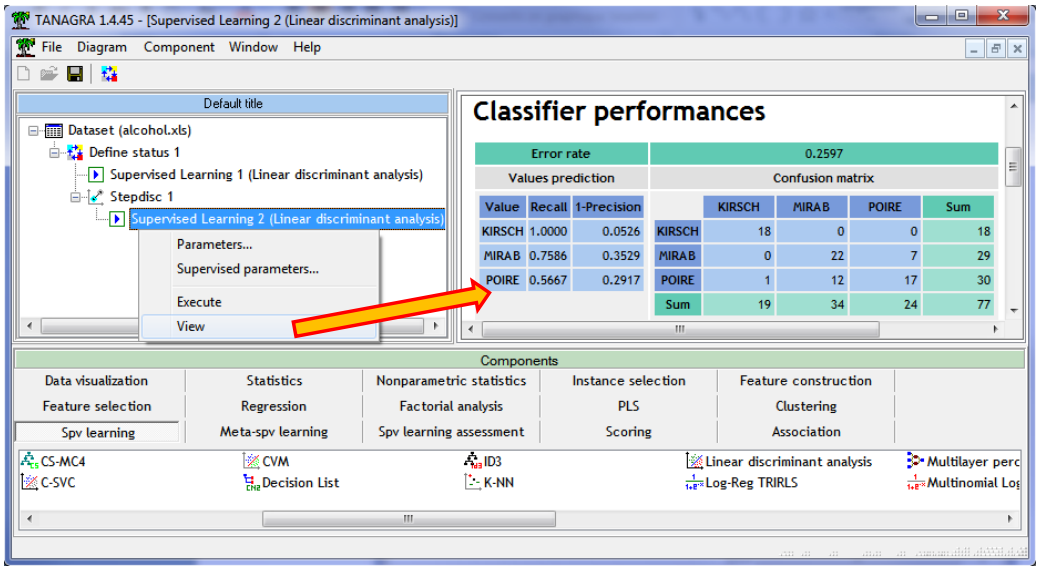
Le composant STEPDISC (onglet FEATURE SELECTION) se charge de cette opération. Nous le plaçons à la suite du ''DEFINE STATUS 1'' où les variables candidates ont été placées en INPUT. Nous actionnons le menu PARAMETERS. Nous choisissons une stratégie ascendante. Le processus d'ajout est stoppé lorsqu'à une étape donnée, la meilleure variable présente une p-value supérieure à 1% (Note : Il n'y a pas de règle établie concernant le choix du seuil. On sait seulement que plus on la diminue, c.-à-d. nous sommes plus exigeants quant aux contributions

des variables, moins nous en obtiendront à la sortie. Et inversement). Nous validons et nous cliquons sur le menu VIEW, 3 variables sont sélectionnées : BU1, MEPR, MEOH.



Tanagra décrit en détail les étapes. Nous y reviendrons lors de la présentation des sorties de SAS.

Modélisation à partir des variables sélectionnées. Il ne nous reste plus qu'à placer de nouveau le composant LINEAR DISCRIMINANT ANALYSIS à la suite de STEPDISC, Tanagra réalise l'analyse discriminante à partir des 3 variables sélectionnées.



Le taux d'erreur en resubstitution est de 25.97%. Le modèle simplifié semble moins bon que le précédent. Mais n'oublions pas que **ce taux calculé sur l'échantillon d'apprentissage** n'est pas crédible. Il **avantage toujours les modèles plus complexes, comportant un nombre plus élevé de variables prédictives**. Il faudrait approfondir l'évaluation en utilisant un échantillon à

part (schéma « holdout ») ou les techniques de ré-échantillonnage (ex. validation croisée ou bootstrap) pour pouvoir réellement statuer.

Voici les coefficients des nouvelles fonctions de classement.

	Classification functions			Statistical Evaluation			
Attribute	**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 - Fonctions de classement après sélection

Déployées sur l'individu ω (Section 3.2.3, les variables inutilisées sont grisées) :

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

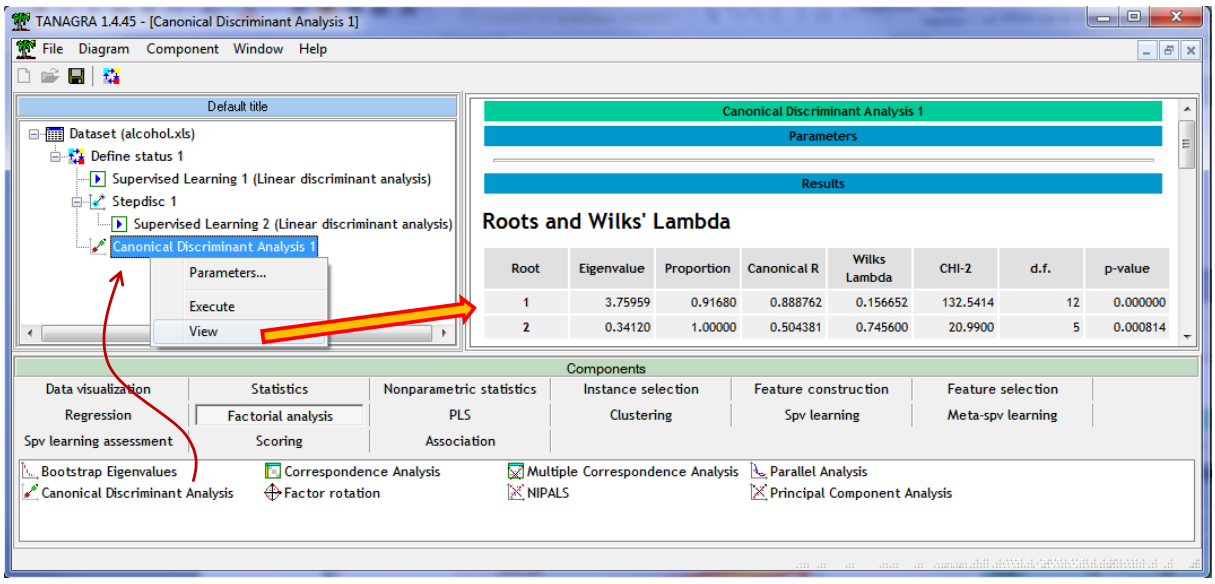
Nous obtenons les scores :

	KIRSCH	MIRAB	POIRE
S(.)	1.4610	6.2670	5.5103

Ici aussi, la classe MIRAB est attribuée à l'individu ω.

3.4 Analyse factorielle discriminante

L'analyse discriminante descriptive n'est pas le sujet principal de ce tutoriel. Nous nous y attarderons quand même pour mieux situer les résultats fournis automatiquement par les autres logiciels. Nous plaçons le composant CANONICAL DISCRIMINANT ANALYSIS (onglet FACTORIAL ANALYSIS) dans le diagramme. Nous actionnons le menu VIEW.



3.4.1 Tableau des valeurs propres

Le tableau des valeurs propres et des proportions de variance expliquées associées à chaque axe factoriel est accompagné des tests de significativité.

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 Fonctions canoniques – Coordonnées factorielles

Les coefficients des fonctions de projection, applicables sur les variables originelles ou standardisées, permettent de calculer les coordonnées factorielles des individus.

Canonical Discriminant Function

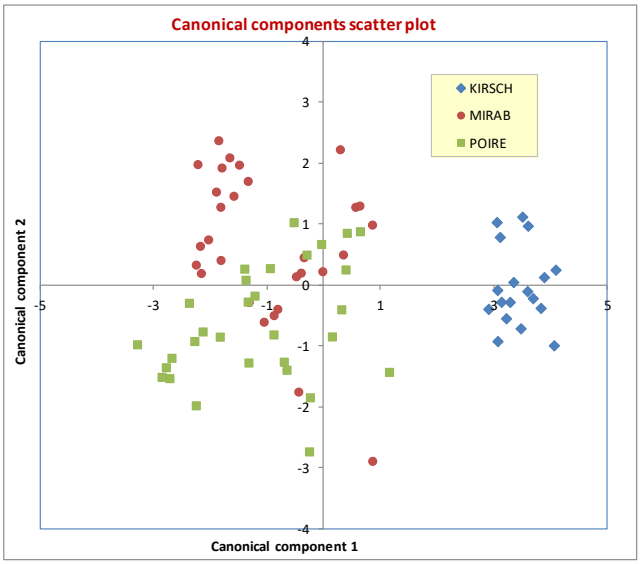
Coefficients	Unstandardized		Standardized	
Attribute	Root n1	Root n2	Root n1	Root n2
MEOH	-0.0033821	-0.000571	-0.6811478	-0.1150081
ACET	0.0000465	0.0066574	0.0051822	0.7420478
BU1	-0.1322048	0.0162599	-0.6469978	0.0795743
MEPR	0.0256226	-0.053361	0.3454686	-0.7194653
ACAL	0.0404876	-0.0297884	0.2160076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

Figure 3 - Fonctions canoniques

Pour l'individu ω avec les caractéristiques suivantes :

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

Nous aurions pour coordonnées sur le premier axe :



Axe 1 = -0.0033821 x 707 + 0.0000465 x 131 – 0.1322048 x 15 + 0.0256226 x 28 + 0.0404876 x 9 + 0.2791911 x 4.89 + 1.89673943 = **-0.0243**

Et sur le second, **Axe 2** = -0.000571 x 707 + … - 0.3894401 x 4.89 + 3.17877051 = **0.2245**

En appliquant ces fonctions sur les observations de l'échantillon d'apprentissage, nous obtenons une représentation des points dans le premier

plan factoriel. Elle est d'autant plus intéressante que nous pouvons distinguer les individus selon leur classe d'appartenance.

Le groupe KIRSCH est bien discriminé, essentiellement sur le 1er axe. MIRAB et POIRE sont plus mélangés, ils se différencient sur le 2nd facteur. Ces conclusions rejoignent celles de la matrice de confusion où nous constations déjà que le groupe KIRSH était prédit sans erreur (en rappel et en précision), contrairement à MIRAB et POIRE (Section 3.2.1).

3.4.3 Structure canonique

La structure canonique indique les corrélations des variables avec les facteurs. Différents types de corrélations peuvent être calculées : globalement, sans tenir compte de l'appartenance aux groupes (TOTAL) ; intra-classe, annihilant l'appartenance aux groupes (WITHIN) ; interclasse, exacerbant les structures de groupe (BETWEEN).

Factor Structure Matrix - Correlations

Root	Root n1			Root n2		
Descriptors	Total	Within	Between	Total	Within	Between
MEOH	**-0.890038**	-0.676771	**-0.991413**	-0.206864	-0.296316	-0.130768
ACET	0.044423	0.021247	0.138243	**0.560792**	0.505278	**0.990398**
BU1	**-0.939650**	-0.787683	**-0.997447**	-0.118534	-0.187184	-0.071407
MEPR	-0.171747	-0.086008	-0.378992	**-0.738951**	-0.697111	**-0.925400**
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

Un rapide coup d'œil sur ce tableau montre que :

1. La distinction entre KIRSCH et les autres, sur le premier axe factoriel, est essentiellement imputable aux variables MEOH et BU1. KIRSCH présente, en moyenne, des valeurs plus faibles sur ces deux variables.
2. La distinction entre MIRAB et POIRE sur le second facteur est essentiellement basée sur l'opposition entre ACET et MEPR. MIRAB a tendance à prendre des valeurs plus élevées que POIRE pour ACET, inversement pour MEPR.

Ces résultats rejoignent celles de l'analyse prédictive. En effet, les variables sélectionnées par STEPDISC à 1% étaient BU1, MEPR et MEOH. Elles caractérisent bien l'appartenance aux groupes.

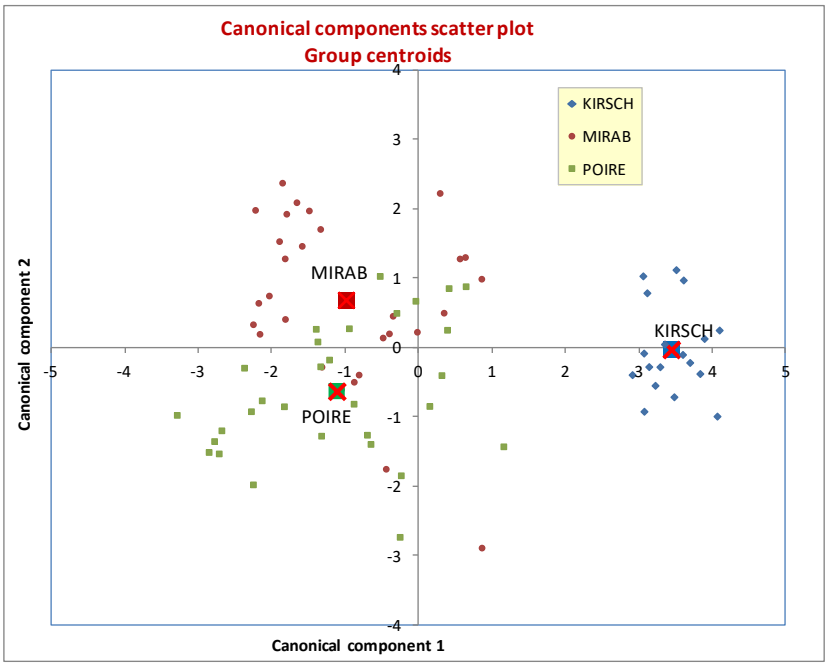
3.4.4 Centres de classes sur les facteurs

Enfin, Tanagra fournit les coordonnées des centres de classes sur les variables canoniques.

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

Elles sont surtout intéressantes représentées dans le nuage de points.



On se rend compte – visuellement, mais nous pouvons le constater par le calcul – que le barycentre de MIRAB est le plus proche de notre point additionnel de coordonnées (**-0.0243, 0.2245**). Ce qui expliquerait son affectation à ce groupe dans l'analyse prédictive ? Nous détaillons les calculs dans la section suivante.

3.4.5 Règle d'affectation aux groupes – Méthode 1

A quel groupe peut-on affecter l'individu de coordonnées (-0.0243, 0.2245) ? Visuellement, il est situé dans une région où MIRAB et POIRE sont majoritaires ; de surcroît, il semble plus proche du barycentre de la classe MIRAB. Il faudrait donc lui attribuer cette dernière étiquette ?

Ce ne sont que des « impressions ». Pour pouvoir y répondre de manière calculatoire, et obtenir un résultat équivalent à l'analyse discriminante prédictive, deux informations supplémentaires sont nécessaires : (1) la proportion des groupes dans l'échantillon de données, (2) la position de leurs barycentres dans le repère factoriel.

Voyons dans le détail la procédure. Nous disposons des barycentres conditionnels depuis la section précédente (Section 3.4.4, « Group centroids ») ; nous donnons ici la proportion des classes.

TYPE	Proportion
KIRSCH	0.233766
MIRAB	0.376623
POIRE	0.389610

Pour le point ω à classer, nous devons calculer le carré de la distance généralisée aux centres de groupes dont voici l'expression générique :

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

Où *c* est une des modalités de la variable cible, K est le nombre de facteurs, $f_k(\omega)$ est la coordonnée de l'individu sur l'axe k, μ_{kc} est la moyenne de la modalité *c* sur le facteur k, π_c est la prévalence de la classe *c*.

Ainsi, respectivement pour les classes KIRSCH, MIRAB et POIRE :

D²(ω, KIRSCH) = (-0.0243 – 3.439733)² + (0.2245 + 0.031885)² - 2 ln(0.233766) = **14.9723**

D²(ω, MIRAB) = (-0.0243 + 0.981483)² + (0.2245 – 0.376623)² - 2 ln(0.376623) = **3.0719**

D²(ω, POIRE) = (-0.0243 + 1.115073)² + (0.2245 + 0.63315)² - 2 ln(0.389610) = **3.8106**

On affecte l'individu ω à la classe qui lui est le plus proche au sens de cette distance, en l'occurrence MIRAB puisque D²(ω, MIRAB) présente la valeur minimale.

A partir de ces éléments, on peut obtenir les probabilités conditionnelles d'appartenance aux classes.

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

Nous reprenons le calcul pour chacune des modalités de la variable cible :

P(Y(ω) = KIRSCH / X) = 0.00056 / 0.36459 = **0.00154**

P(Y(ω) = MIRAB / X) = 0.21525 / 0.36459 = **0.59039**

P(Y(ω) = POIRE / X) = 0.14878 / 0.36459 = **0.40807**

La probabilité conditionnelle est en réalité un score, mais avec les propriétés – très intéressantes – d'une probabilité (ex. possibilité de couplage avec les coûts de mauvaise affectation). On attribue la modalité qui présente la valeur la plus élevée, en l'occurrence la classe MIRAB pour ce qui est de l'individu ω.

3.4.6 Règle d'affectation aux groupes – Méthode 2

En développant la distance généralisée, en la multipliant par -0.5, et en supprimant tous les éléments qui ne dépendent pas des classes, on peut produire des fonctions de classement linéaires – à maximiser – basées sur les coordonnées factorielles[3]. Elles sont équivalentes aux fonctions fournies par l'analyse discriminante prédictive c.-à-d. elles classent exactement de manière identique, les valeurs sont différentes à une constante près (qui ne dépend pas des classes).

Nous avons :

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

Pour notre individu ω à classer :

S'(ω, KIRSCH) = -0.0243 x 3.439733 + 0.2245 x (-0.031885) – (3.439733² + (-0.031885)²)/2 + ln(0.233766) = **-7.4606**

S'(ω, MIRAB) = -0.0243 x (-0.981483) + 0.2245 x 0.674773 – ((-0.981483)² + 0.674773²)/2 + ln(0.376623) = **-1.5105**

S'(ω, POIRE) = -0.0243 x (-1.115073) + 0.2245 x (-0.63315) – ((-1.115073)² + (-0.63315) ²)/2 + ln(0.389610) = **-1.8798**

Nous lui affectons la modalité MIRAB. La démarche est forcément cohérente avec les précédentes (Sections 3.2.3 et 3.4.5).

3.4.7 Fonctions scores – Retour sur les variables originelles (Méthode 3)

En développant l'expression précédente, nous trouvons une fonction linéaire définie sur les facteurs. Nous avons respectivement :

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

Pour KIRSH par exemple : **S'(KIRSH)** = 3.439733 x f1 -0.031885 x f2 – 7.369825

Les facteurs **F1** et **F2** étant eux-mêmes des combinaisons linéaires des variables originelles (Figure 3), il est facile de déduire une fonction définie sur ces variables. Nous obtenons ainsi une autre version de la fonction score S'.

[3] W. Venables, B. Ripley, « Modern Applied Statistics with S », Springer, 2002 ; page 334.

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

Que l'on peut rapprocher avec les fonctions fournies par l'analyse discriminante prédictive (Figure 1).

Ce que je n'ai pas manqué de faire bien évidemment. J'ai constaté que les coefficients de S() et S'() étaient différents, mais que les écarts entre les modalités étaient les mêmes pour chaque variable.

S() - Analyse prédictive			
Attribute	KIRSCH	MIRAB	POIRE
MEOH	0.000659	0.015208	0.016407
ACET	0.000445	0.004944	-0.003770
BU1	-0.039342	0.556654	0.553048
MEPR	0.186892	0.035901	0.102271
ACAL	0.039174	-0.160881	-0.127328
LNPRO1	6.378935	4.869370	5.341430
constant	-24.686164	-25.141755	-29.631644

S'() - Déduite de l'analyse descriptive		
KIRSCH	MIRAB	POIRE
-0.011615	0.002934	0.004133
-0.000052	0.004447	-0.004267
-0.455268	0.140729	0.137123
0.089836	-0.061155	0.005214
0.140216	-0.059838	-0.026286
0.972760	-0.536805	-0.064744
-0.946902	-1.402493	-5.892384

Ecarts entre coefficients		
KIRSCH	MIRAB	POIRE
0.0123	0.0123	0.0123
0.0005	0.0005	0.0005
0.4159	0.4159	0.4159
0.0971	0.0971	0.0971
-0.1010	-0.1010	-0.1010
5.4062	5.4062	5.4062
-23.7393	-23.7393	-23.7393

C'est la raison pour laquelle nous n'avions pas les mêmes valeurs de score pour l'individu à classer ω

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

En revanche, l'écart entre S() et S'() dépend de l'individu et non pas des modalités de la variable cible

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

Au final, les individus sont classés de manière identique par les deux procédés. C'est ce qui importe.

4 Analyse discriminante avec SAS

4.1 Proc DISCRIM

Voyons maintenant ce que nous dit **SAS 9.3** pour la même analyse. L'importation des données dans une banque a déjà été décrite dans un précédent document (voir « La Proc Logistic de SAS 9.3 », avril 2012). Nous soumettons les commandes suivantes pour lancer l'analyse discriminante prédictive.

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

Nous faisons appel à la procédure DISCRIM en spécifiant la base à traiter (DATA), la variable cible (CLASS) et les variables prédictives (VAR). Avec l'option PRIORS, SAS utilise les fréquences des classes calculées sur l'échantillon d'apprentissage.

Deux options par défaut sont importantes : METHOD = NORMAL, on fait l'hypothèse d'une distribution gaussienne des variables conditionnellement aux classes ; POOL = YES, on utilise la matrice de variance covariance intra-classes. A la sortie, nous obtenons bien un classifieur linéaire.

Informations générales. SAS fournit une indication globale sur l'analyse que nous menons. Nous disposons entres autres des observations par classe. KIRSCH est la modalité la plus rare.

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 entres centres de classes. Dans la deuxième partie, nous avons tout d'abord le carré de la distance généralisée – de Mahalanobis – entre les centres de classes.

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

Deux éléments interpellent dans ce tableau : la distance d'un centre de classe avec lui-même n'est pas nulle, et la matrice n'est pas symétrique. Clairement, il y a autre chose dans la formule. Nous avons la réponse en farfouillant dans la documentation de SAS. La prévalence des classes est introduite dans les calculs (http://www.math.wpi.edu/saspdf/stat/chap25.pdf).

En se basant sur les centres de classes sur les facteurs (Section 3.4.4), la distance de KIRSCH avec elle-même devient[4] :

D^2(KIRSCH, KIRSCH) = (3.439733 - 3.439733)² + (-0.031885 + 0.031885)² – 2 x ln(0.233766) = **2.90687**

La « distance » de KIRSCH avec MIRAB (dans un sens puis dans l'autre) :

D^2(KIRSCH, MIRAB) = (3.439733 + 0.981483)² + (-0.031885 – 0.674773)² - 2 x ln(0.233766) = **22.95339**

D^2(MIRAB, KIRSCH) = (-0.981483 – 3.439733)² + (0.674773 + 0.031885)² -2 x ln(0.376623) = **21.99954**

Fonctions de classement. SAS fournit ensuite les fonctions de classement. Elles sont strictement identiques à celles de Tanagra. Les règles d'affectation sont exactement les mêmes.

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

Curieusement, SAS ne donne aucune indication sur la pertinence des variables.

Matrice de confusion et taux d'erreur. Enfin, SAS produit la matrice de confusion et le taux d'erreur par classe (il s'agit plutôt de « 1 – sensibilité »).

[4] La distance euclidienne dans le repère factoriel équivaut à la distance de Mahalanobis dans le repère originel.

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	

4.2 Sélection de variables avec STEPDISC

SAS propose la procédure STEPDISC pour une sélection de variables en adéquation avec l'analyse discriminante. Je m'en suis inspiré pour programmer le composant du même nom dans Tanagra.

Nous soumettons les commandes suivantes pour un processus de sélection METHOD = FORWARD contrôlée par un risque de SLENTRY = 1% :

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

Nous comparerons les sorties avec celles de Tanagra (Figure 4).

Detailed results

N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(2, 74)	BU1 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)	MEPR 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)	MEOH 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 - Détail des étapes - Stepdisc à 1% de Tanagra

Etape 1 : SAS sélectionne BU1 après avoir listé les contributions des variables candidates.

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)	BU1 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 fournit R-Square = 1 - Λ. Pour MEOH par exemple, R-Square(MEOH) = 1 – 0.363 = 0.6366. La statistique F est équivalente à celle de Tanagra, elle permet d'évaluer la significativité de la variable F(MEOH) = 64.82, avec la p-value(MEOH) = 0.0001. La tolérance indique la redondance avec les variables déjà intégrées dans le modèle. A la première étape, l'ensemble initial est vide. Il n'y a pas de redondance possible, elle est donc égale à 1 pour toutes les variables.

La variable BU1 présente le « F Value » le plus élevé et elle est significative (p-value(BU1) < 1%). Elle est incorporée dans le modèle. Sa présence ne sera plus remise en cause.

SAS enchaîne avec l'évaluation du modèle courant.

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				

Le modèle ne comportant qu'une variable, il n'y a pas d'éléments réellement nouveaux ici. La trace du Pillai est une variante du test d'écartement significatif entres les barycentres conditionnels[5], et « Average Squared Canonical Correlation » est égal à la statistique du Pillai divisé par le « nombre de groupes – 1 » (dixit la documentation de Stepdisc).

Etape 2 : SAS évalue les variables restantes. Le « Partial R-Square » confronte les Λ des modèles incluant ou non la variable candidate. Pour MEOH par exemple, Partial R-Square(MEOH) = 1 – 0.251 / 0.299 ≈ 0.160. MEPR présente la valeur la plus élevée et elle est significative. Elle est validée.

SAS

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	L : 0.249	L : 0.251	L : 0.275	L : 0.276	L : 0.298
		F : 7.34	F : 7.34	F : 6.95	F : 3.17	F : 3.05	F : 0.18
		p : 0.0012	p : 0.0012	p : 0.0017	p : 0.0478	p : 0.0535	p : 0.8353

Notons également, même si cela n'est pas pris en compte pour la sélection, que MEPR est faiblement corrélée avec la variable déjà introduite puisque sa tolérance est de 0.9046. Cela veut dire que le carré de la corrélation entre BU1 et MEPR est égal $r^2 = 1 - 0.9046 = 0.0954$.

Pour l'évaluation du modèle à deux variables, nous avons :

[5] Voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf, section 7.1.4.

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 continue ainsi jusqu'à ce qu'il ne soit plus possible d'introduire de variables. Il édite alors un tableau récapitulant le processus (Figure 5).

SAS

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

TANAGRA

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 - Récapitulatif de la sélection - Stepdisc forward à 1%

Nous obtenons un sous ensemble de 3 variables : BU1, MEPR et MEOH.

5 Analyse discriminante avec R – Procédure lda() [MASS]

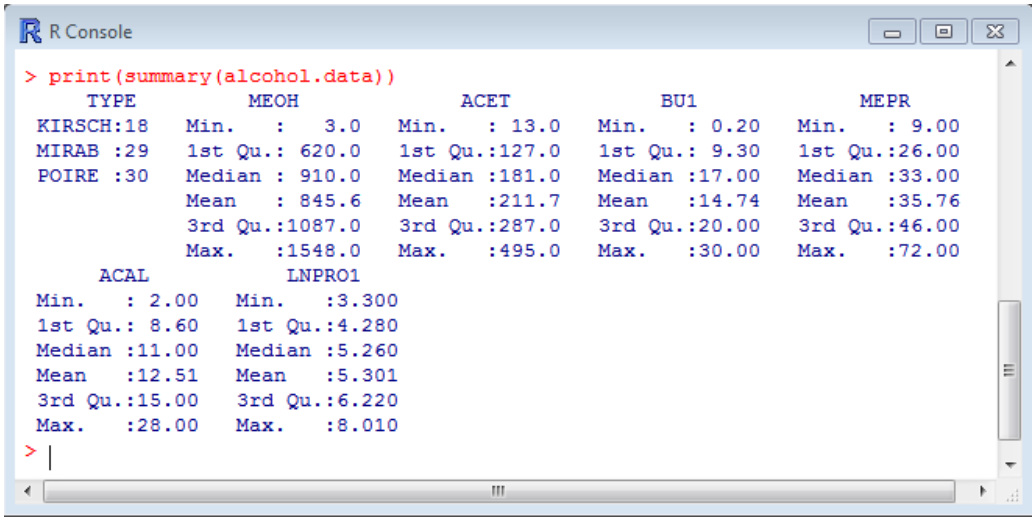
Nous utilisons la procédure **lda()** du package « MASS » pour réaliser l'analyse discriminante. Ce package est automatiquement installé avec R, il n'est pas nécessaire de le charger à partir du web.

5.1 Chargement des données

La commande **read.xlsx()** du package « xlsx » permet d'importer directement un fichier au format Excel (XLS et XLSX). Nous effectuons une vérification en affichant un résumé statistique.

```
library(xlsx)
#sheetIndex : numéro de la feuille à lire
#header : la première ligne correspond aux noms des variables
alcohol.data <- read.xlsx(file="alcohol.xls",sheetIndex=1,header=T)
print(summary(alcohol.data))
```

Nous obtenons une description succincte de chaque variable.



TYPE est la seule variable catégorielle, c'est celle que l'on cherche à prédire.

5.2 Modélisation – Propriétés de l'objet

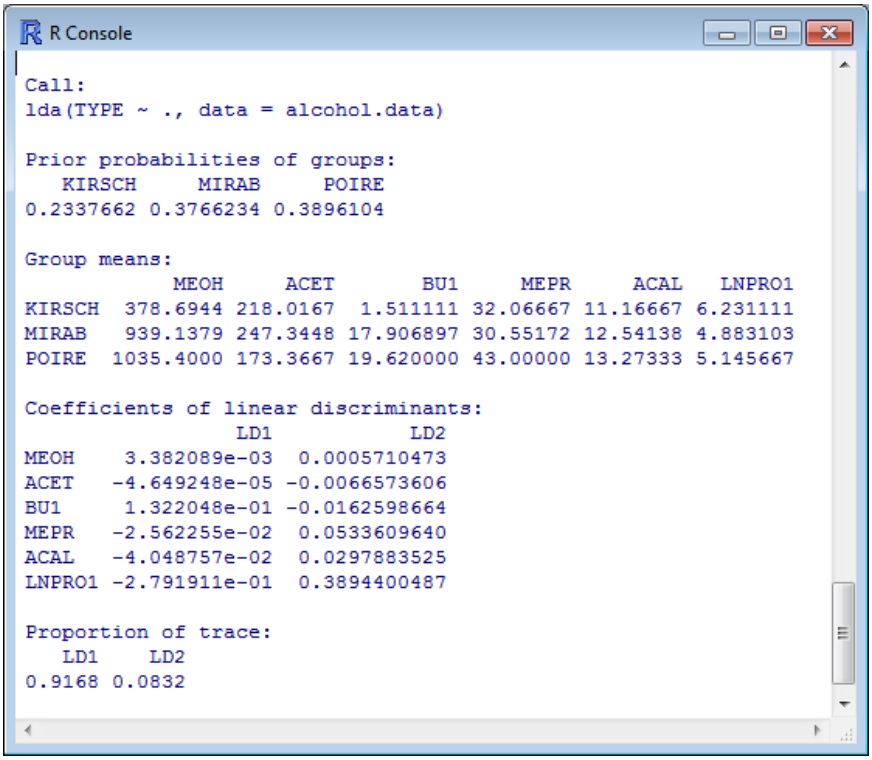
Après avoir chargé le package MASS[6], nous lançons l'apprentissage avec la méthode **lda()**. Nous affichons l'objet résultat.

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

Nous obtenons : les proportions des classes ; les moyennes des variables conditionnellement aux classes ; les coefficients non standardisés de la fonction canonique (Section 3.4.2), sans la constante néanmoins ; les proportions de variance expliquées par les facteurs.

Remarque : A priori, lda() semble taillée pour l'analyse discriminante descriptive. Mais nous avions vu plus haut que le chemin du descriptif au prédictif est finalement très court (Section 3.4.5). La fonction **predict()** appliquée à l'objet 'lda' permet ainsi de réaliser des prédictions identiques à celles de Tanagra et SAS c.-à-d. les individus en généralisation sont affectés aux mêmes classes.

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

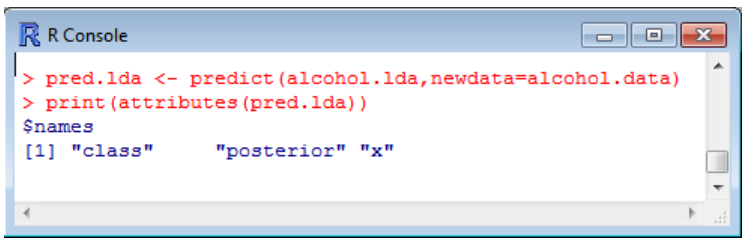


5.3 Prédiction

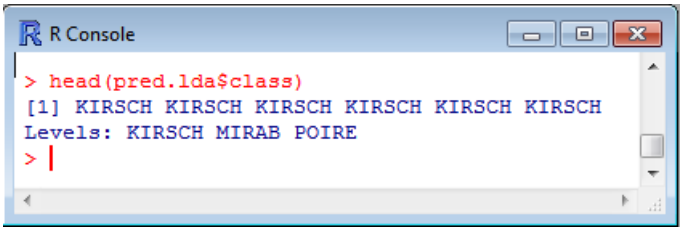
Nous utilisons la commande **predict()** pour réaliser la prédiction sur l'échantillon d'apprentissage.

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

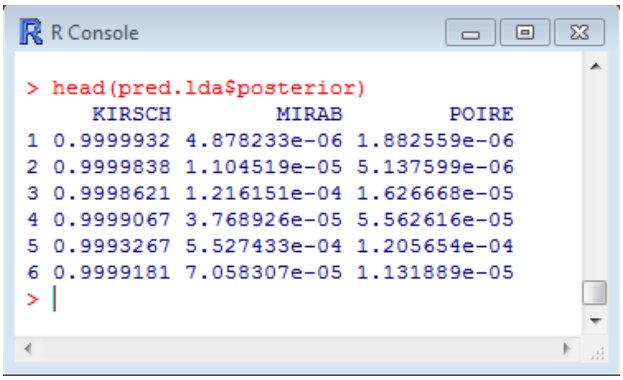
L'objet qui en est issu propose 3 propriétés : 'class', 'posterior' et 'x' :



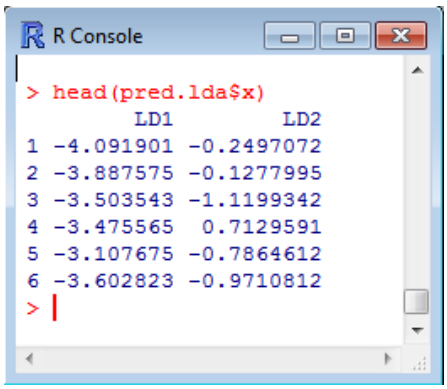
- 'class' est un vecteur de n = 77 lignes, il correspond aux classes prédites pour chaque observation. Voici les prédictions pour les 6 premières observations



– 'posterior' est une matrice de 77 lignes et 3 colonnes (parce la variable cible possède 3 modalités). Elle fournit les probabilités conditionnelles d'appartenance aux classes pour chaque observation. Pour les 6 premières observations, nous avons



– 'x' correspond enfin aux coordonnées factorielles, il s'agit d'une matrice de 77 lignes et 2 colonnes (parce que 2 facteurs). Pour les 6 premiers individus,

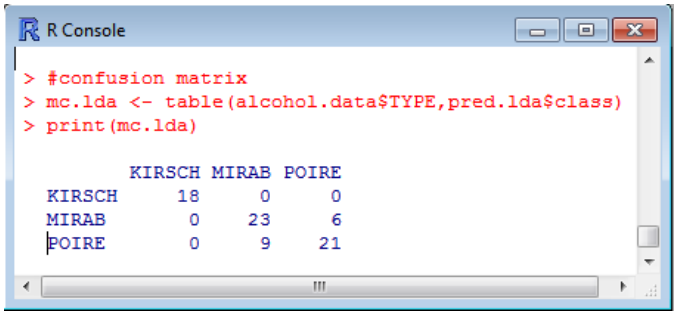


5.4 Matrice de confusion

Pour calculer la matrice de confusion, nous croisons les classes observées et prédites par le modèle.

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

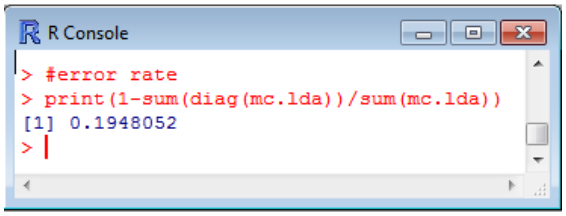
Nous retrouvons la matrice proposée par Tanagra (Section 3.2.1) et SAS.



Nous en déduisons le taux d'erreur

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

Soit,

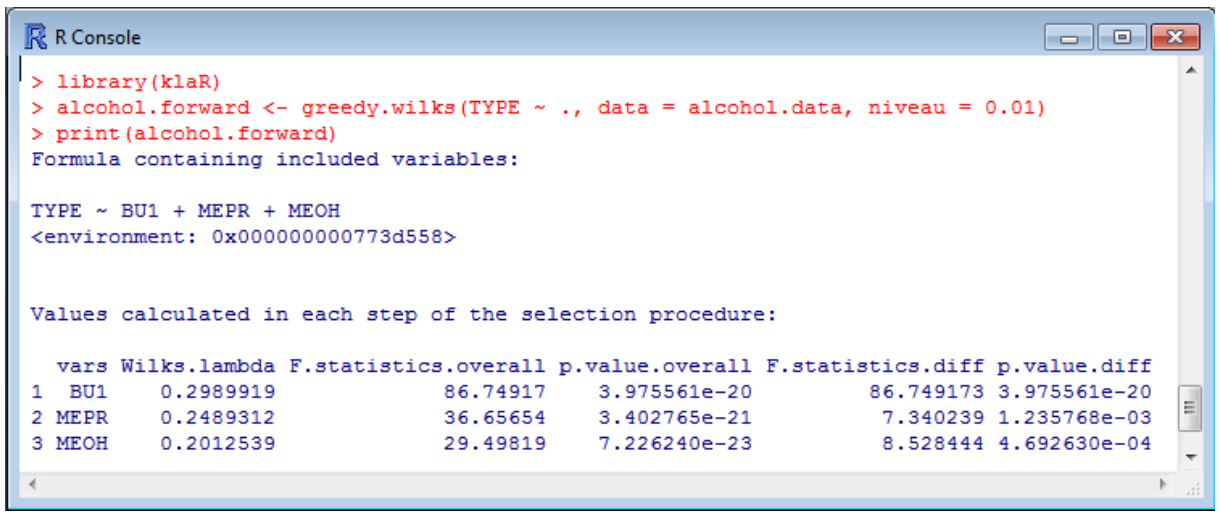


5.5 Sélection de variables

Nous utilisons la procédure **greedy.wilks()** du package « klaR » pour la sélection de variables.

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

Pour chaque étape, R fournit les statistiques de test pour le choix des variables (**F.statistics.diff**), et pour le modèle courant incorporant les variables sélectionnées (**F.statistics.overall**).

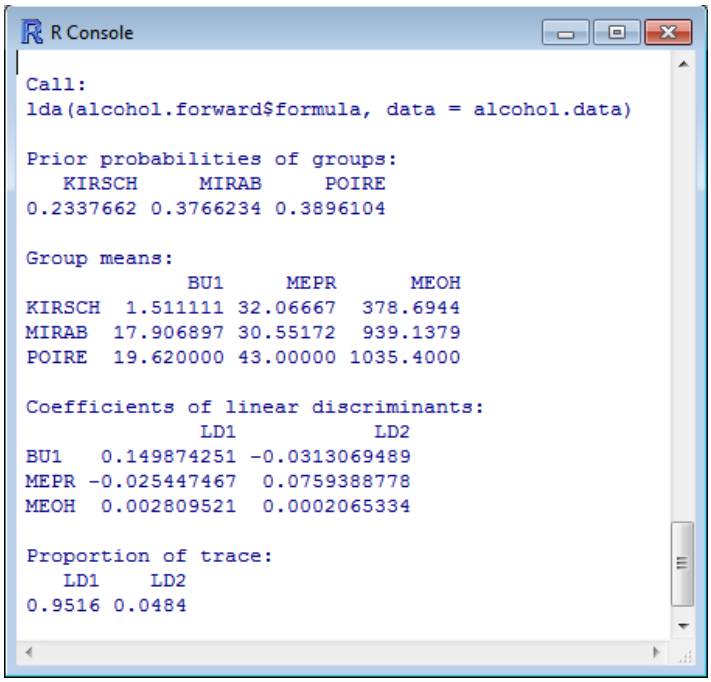


Les valeurs présentées rejoignent en tous points celles proposées par Tanagra et SAS (Figure 5). La démarche et les formules utilisées sont exactement les mêmes.

Nous lançons la modélisation sur les variables sélectionnées par **greedy.wilks()**. Cette dernière fournit directement la formule réduite. C'est très appréciable si nous avons à traiter une grande base. Il ne sera pas nécessaire de recopier manuellement la liste des variables.

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

Voici la nouvelle version du modèle.

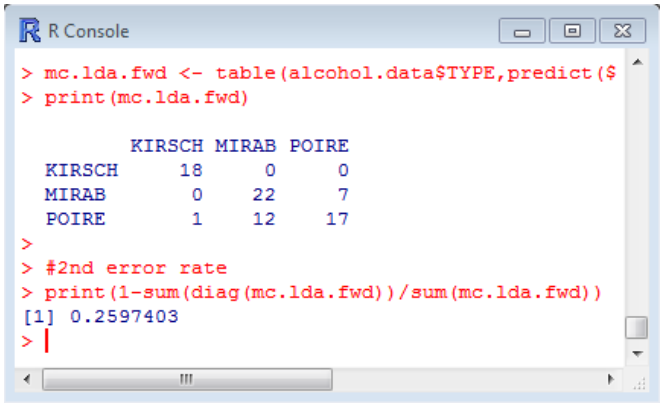


Nous le réappliquons sur l'échantillon d'apprentissage pour obtenir la matrice de confusion et le taux d'erreur en resubstitution.

```
#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))
```

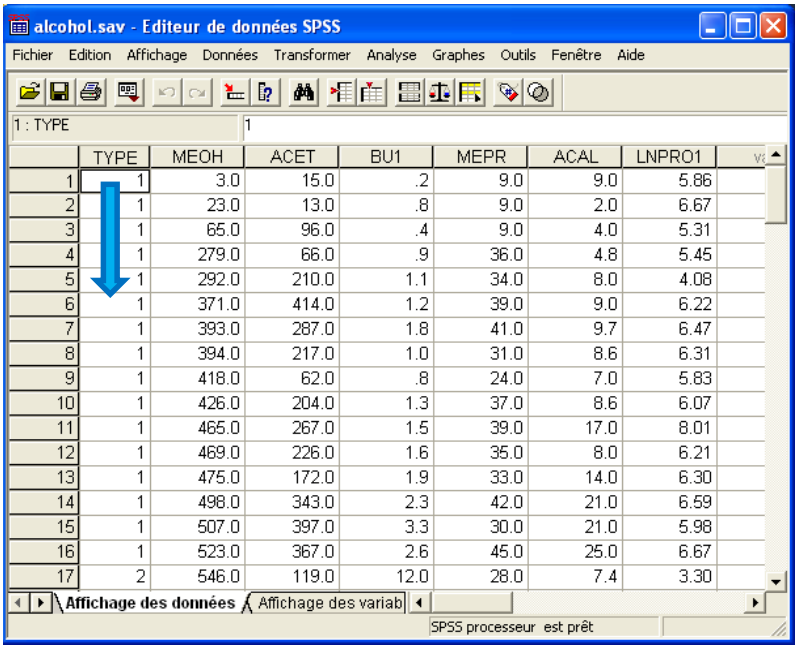
Nous avons :



Encore une fois, le taux d'erreur étant calculé sur l'échantillon d'apprentissage, nous ne savons pas vraiment si ce second modèle est réellement moins bon que celui intégrant toutes les prédictives.

6 Analyse discriminante avec SPSS

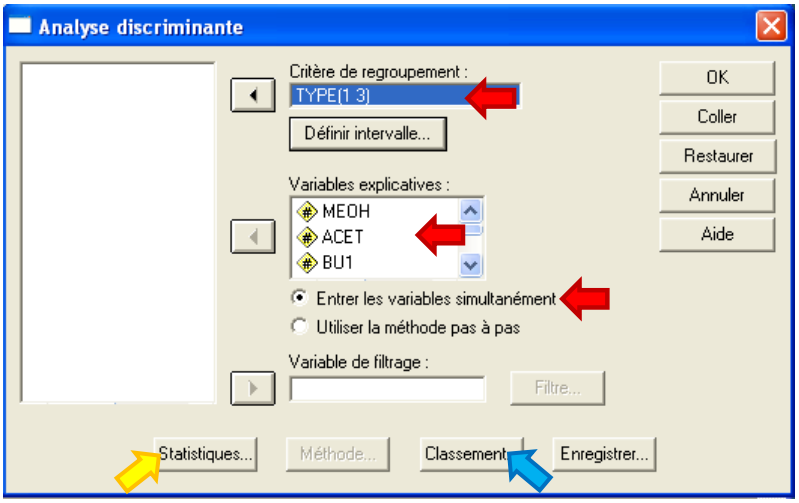
Nous utilisons la version française de **SPSS 12.0.1**. Nous importons le fichier « alcohol.xls » après avoir recodé la variable TYPE en numérique (KIRSCH = 1, MIRAB = 2, POIRE = 3).



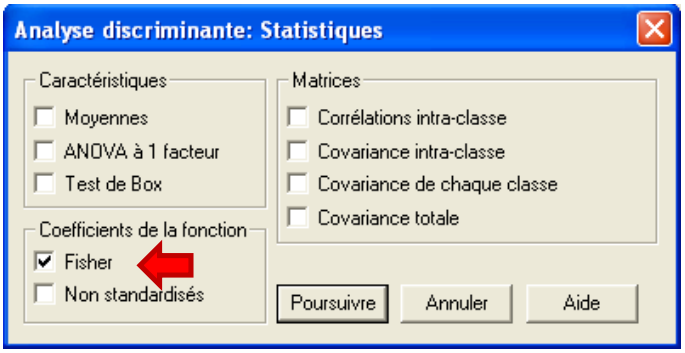
	TYPE	MEOH	ACET	BU1	MEPR	ACAL	LNPRO1
1	1	3.0	15.0	.2	9.0	9.0	5.86
2	1	23.0	13.0	.8	9.0	2.0	6.67
3	1	65.0	96.0	.4	9.0	4.0	5.31
4	1	279.0	66.0	.9	36.0	4.8	5.45
5	1	292.0	210.0	1.1	34.0	8.0	4.08
6	1	371.0	414.0	1.2	39.0	9.0	6.22
7	1	393.0	287.0	1.8	41.0	9.7	6.47
8	1	394.0	217.0	1.0	31.0	8.6	6.31
9	1	418.0	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 du modèle complet

Nous actionnons le menu **ANALYSE / CLASSIFICATION / ANALYSE DISCRIMINANTE**. La boîte de paramétrage apparaît.

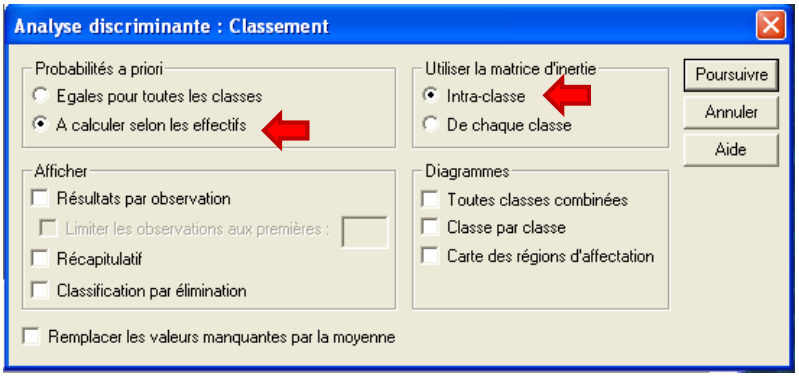


Nous plaçons TYPE en « **Critère de regroupement** », l'intervalle est de MIN = 1 à MAX = 3. Toutes les autres sont en « **Variables explicatives** ». Pour l'instant, nous faisons « **Entrer les variables simultanément** » c.-à-d. nous n'opérons pas de sélection de variables.



Nous cliquons ensuite sur le bouton « **Statistiques** » pour demander l'affichage des fonctions de classement « Coefficients de la fonction = Fisher ».

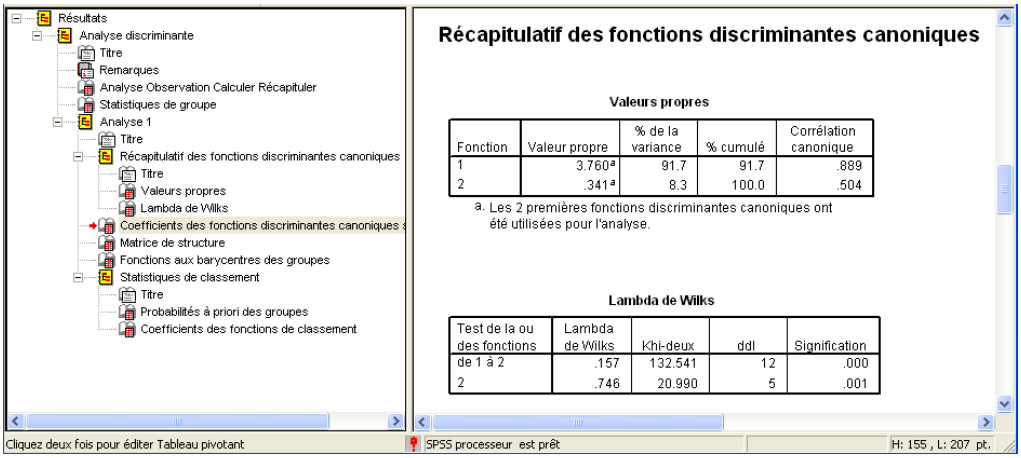
De retour dans la boîte de paramétrage, nous cliquons sur « **Classements** ». Nous souhaitons que les distributions des classes soient estimées à partir des effectifs du fichier d'apprentissage, et que la matrice de variance covariance intra-classe soit utilisée.



Nous validons nos choix. Un rapport est édité dans une nouvelle fenêtre à l'issue des calculs.

SPSS mélange les sorties des analyses canoniques et prédictives. Ce n'est pas choquant. En effet, les deux approches se rejoignent comme nous l'avions précisé précédemment. Il faut le savoir simplement et discerner les informations adéquates dans les sorties.

A. En premier lieu, nous avons les résultats relatifs à l'analyse canonique. Ainsi, nous disposons des valeurs propres liées aux facteurs et les tests de significativité associés.



Récapitulatif des fonctions discriminantes canoniques

Valeurs propres

Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	3.760[a]	91.7	91.7	.889
2	.341[a]	8.3	100.0	.504

a. Les 2 premières fonctions discriminantes canoniques ont été utilisées pour l'analyse.

Lambda de Wilks

Test de la ou des fonctions	Lambda de Wilks	Khi-deux	ddl	Signification
de 1 à 2	.157	132.541	12	.000
2	.746	20.990	5	.001

Puis, nous avons successivement : (1) les fonctions canoniques de projection, (2) les corrélations intra-classes (within) des variables avec les facteurs, (3) les barycentres conditionnels.

Coefficients des fonctions discriminantes canoniques standardisées

	Fonction 1	Fonction 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	Fonction 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.

(2)

Fonctions aux barycentres des groupes

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

Fonctions discriminantes canoniques non standardisées évaluées aux moyennes des groupes

(3)

B. En second lieu, nous avons les résultats relatifs à la prédiction avec, entres autres, les fonctions de classement appelées « Fonctions discriminantes linéaires de Fisher » dans SPSS.

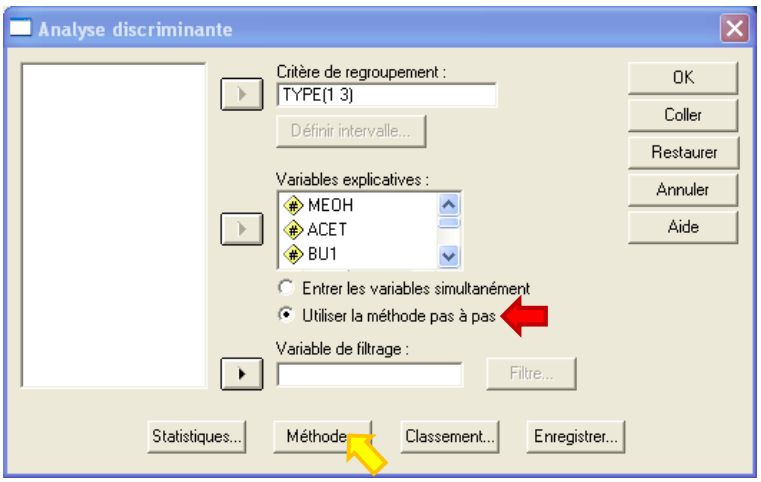
Coefficients des fonctions de classement

	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

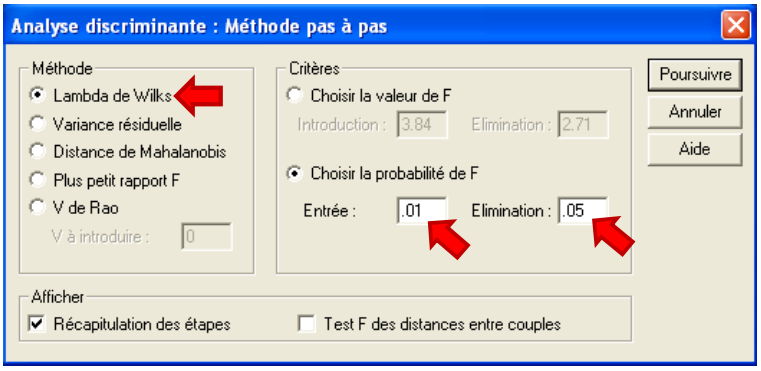
Fonctions discriminantes linéaires de Fisher

6.2 Sélection de variables

Pour effectuer une sélection de variable, il faut lancer une nouvelle analyse en précisant le mode de traitement des prédictives. Nous choisissons « Utiliser la méthode pas à pas ». Puis nous cliquons sur le bouton « Méthode » qui est maintenant activé.



Seule la méthode STEPWISE est disponible. A chaque ajout, SPSS se donne la possibilité de retirer une variable introduite aux étapes précédentes. Deux niveaux de signification permettent de piloter le processus, pour l'entrée (0.01) et pour le retrait (0.05). Comme pour les autres logiciels, le calcul est basé sur la comparaison des lambdas de Wilks (SPSS peut s'appuyer sur d'autres critères).



Plusieurs tableaux viennent décrire le processus de sélection.

Le premier décrit la pertinence des modèles successifs : à 1 explicative, à 2 explicatives, etc. Le tableau n'est pas sans rappeler le récapitulatif de **greedy.wilks()** de R.

Variables introduites/éliminées[a,b,c,d]

Pas	Introduite	Lambda de Wilks							
						F exact			
		Statistique	ddl1	ddl2	ddl3	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

Ensuite, nous avons une description plus détaillée des différents modèles. La tolérance permet de situer la redondance entre les variables explicatives. Elles le sont faiblement en ce qui nous concerne (tolérance = 1 veut dire que la variable est orthogonale aux autres ; tolérance = 0 veut dire que la variable peut être déduite d'une combinaison linéaire des autres).

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

En dernière colonne, nous observons le lambda de Wilks lorsqu'une variable déjà sélectionnée venait à être retirée. Par exemple, dans le modèle à 3 variables (BU1, MEPR et MEOH), BU1 semble avoir l'impact le plus fort, MEPR et MEOH en revanche sont d'égales importances. Ces valeurs sont les mêmes que le « Wilks L. » qui accompagnent les fonctions de classement fournies par Tanagra après sélection de variables (Figure 2).

Enfin, un dernier tableau détaille tous les calculs pour chaque étape. Nous pouvons le rapprocher avec celui de Tanagra (Figure 4).

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

L'analyse discriminante prédictive est séduisante à bien des égards. Elle est proposée dans quasiment tous les logiciels de statistique. Dans ce tutoriel, nous avons essayé de situer les points communs et les différences entre les sorties des logiciels Tanagra, R, SAS et SPSS. La très bonne nouvelle est qu'en définitive, nous avons exactement les mêmes résultats.

8 Références

Cours Data Science, http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html

- « Analyse discriminante prédictive ».
- « Analyse factorielle discriminante ».