Analyse discriminante descriptive

ou

Analyse factorielle discriminante

Caractériser de manière multidimensionnelle (à l'aide de plusieurs variables, simultanément) l'appartenance des individus à des groupes prédéfinis

Ricco RAKOTOMALALA

PLAN

1. Position du problème
2. Détermination des variables discriminantes (axes factoriels)
3. Analyse des résultats
4. Etude de cas
5. Classement (prédiction) avec l'analyse discriminante
6. Les logiciels (Tanagra, R avec lda, SAS avec PROC CANDISC)
7. Conclusion
8. Bibliographie

Position du problème

Construire un nouveau système de représentation (variables latentes)

qui permet de mettre en évidence les groupes

Une population est subdivisée en K groupes (classes), elle est décrite par une série de **J** caractères (variables) quantitatives.

Ex. Les vins de Bordeaux
(Tenenhaus, 2006; page 353)
Les lignes correspondent aux années (1924 à 1957)

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

Description

Groupe d'appartenance

Objectif(s) :

(1) Descriptif (Schéma d'explication) : Mettre en évidence les caractéristiques qui permettent de distinguer au mieux les groupes → Objectif principal

(2) Prédictif (Schéma de prédiction) : Classer automatiquement un nouvel individu (l'affecter à un groupe) à partir de ses caractéristiques → Objectif secondaire *dans notre contexte (!)* (*on se reproche de l'AD Prédictive dans ce cas, cf. support associé – Analyse discriminante linéaire*)

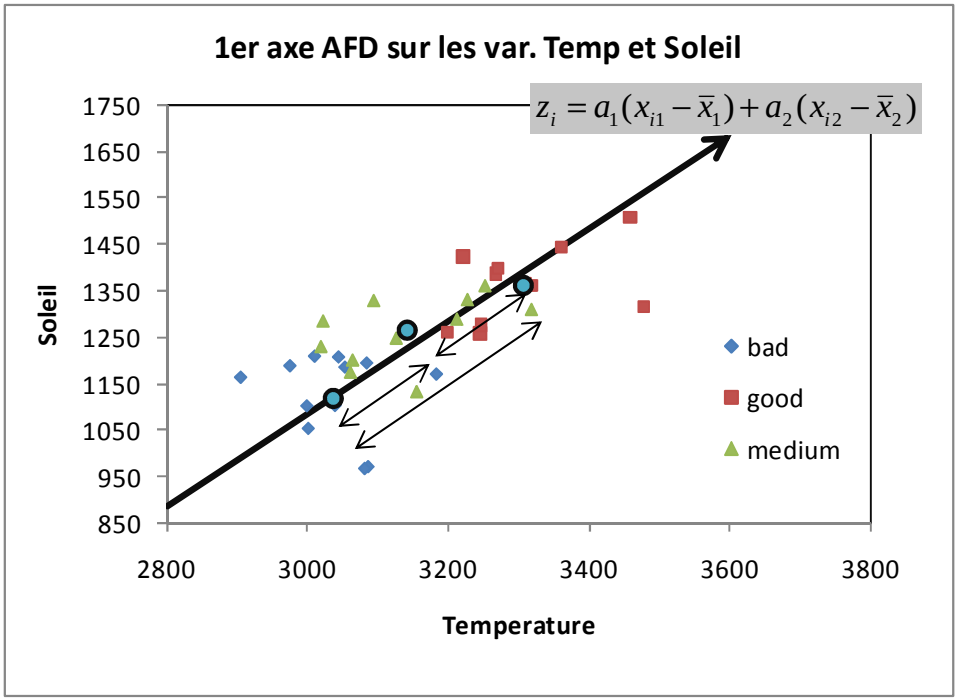
Principe : Trouver une succession de combinaisons linéaires des variables initiales (on parle de variables latentes ou variables discriminantes, elles sont deux à deux orthogonales) qui permet de distinguer au mieux (au sens des barycentres) les groupes → analyse factorielle discriminante

On souhaite que les barycentres conditionnels, projetés sur l'axe factoriel, soient le plus écartés possibles.

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

SCT = SCE + SCR

SC totaux = SC expliqués (groupes) + SC résiduels



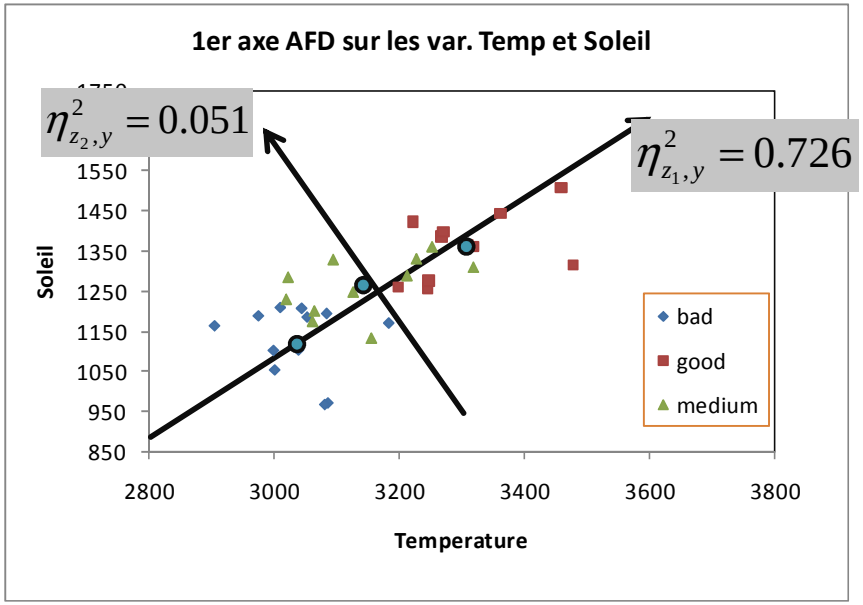
Analyse discriminante descriptive – Démarche (suite)

Un indicateur de qualité de la séparation des groupes à maximiser : le rapport de corrélation

$$\eta^2_{z,y} = \frac{SCE}{SCT} \quad \text{avec} \quad 0 \leq \eta^2_{z,y} \leq 1$$

1 → discrimination parfaite, les points associés aux groupes sont agglutinés sur leurs barycentres (SCR = 0)

0 → discrimination impossible, barycentres confondus (SCE = 0)



→ Trouver les coefficients (a1,a2) qui définissent la variable discriminante Z (ou axe factoriel) maximisant le rapport de corrélation

→ Le nombre d'axes factoriels est égal à M = MIN(J, K-1)

→ Les axes sont deux à deux orthogonaux

→ Les axes suivants maximisent l'écart entre les barycentres en contrôlant l'effet des axes précédents c.-à-d. ils essaient d'expliquer les écarts entre les barycentres non pris en compte encore par les axes précédents

→ Le pouvoir discriminatoire est quantifié par le rapport de corrélation

Solution

Recherche d'une représentation « optimale » au sens du rapport de corrélation

Analyse discriminante descriptive

Formulation mathématique

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

« a » est le vecteur des coefficients permettant de définir le premier axe factoriel Z c.-à-d. $z = a_1(x_1 - \bar{x}_1) + \cdots + a_J(x_J - \bar{x}_J)$

Matrice de variance covariance totale

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

$\Rightarrow$ $SCT = a'Va$ *[à un facteur (1/n) près]*

Matrice de variance covariance intraclasses

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

$\Rightarrow$ $SCR = a'Wa$

Matrice de variance covariance interclasses

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

$\Rightarrow$ $SCE = a'Ba$

Théorème d'Huyghens → V = B + W

L'ADD consiste à chercher le vecteur de coefficients « a » qui permet de définir une axe (variable latente Z) qui maximise le rapport de corrélation avec Y

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

Analyse discriminante descriptive

Solution mathématique

$\max_a \frac{a'Ba}{a'Va}$ est équivalent à $\max_a a'Ba$

Sous la contrainte $a'Va = 1$ *Le vecteur « a » est normé*

Solution : former le lagrangien, et annuler la dérivée c.-à-d.

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

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

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

λ est la première valeur propre de $V^{-1}B$

« a » est le vecteur propre associé

De manière générale, les axes factoriels de l'ADD sont définis par les valeurs et vecteurs propres de la matrice $V^{-1}B$.

Au plus, nous avons M = min(K-1, J) valeurs propres non nulles, et donc autant d'axes factoriels.

$\lambda = \eta^2$ La valeur propre est égal au rapport de corrélation associé à l'axe ($0 \leq \lambda \leq 1$)

$\eta = \sqrt{\lambda}$ Est la « corrélation canonique »

Analyse discriminante descriptive

Vins de Bordeaux (X1 : Température et X2 : Soleil)

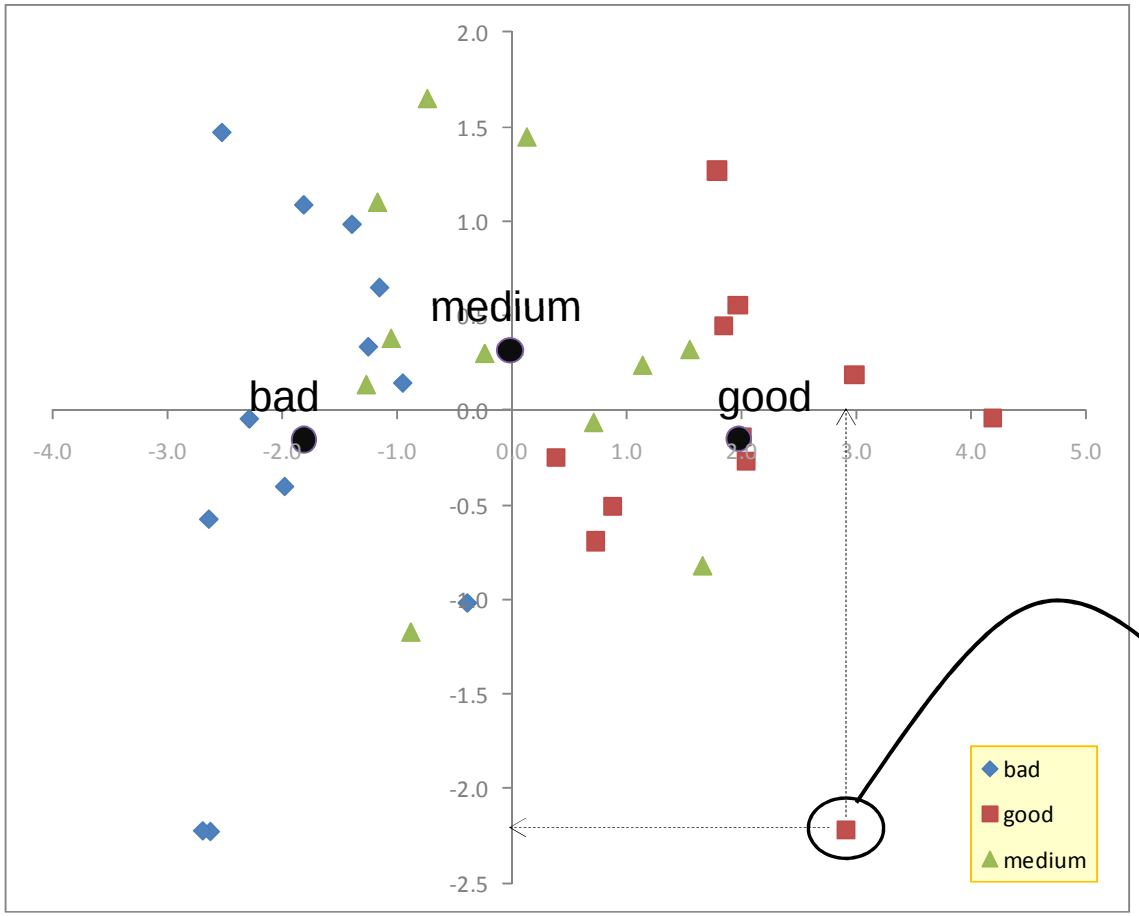
Nombre d'axes

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

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

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

L'écartement entre les barycentres est moindre sur cet axe.



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

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

L'écartement entre les barycentres est élevé sur cet axe.

(2.91; -2.22) : les coordonnées factorielles d'un individu sont appelées « score » dans les logiciels anglo-saxons (SAS, SPSS, R…)

Analyse discriminante descriptive

Solution mathématique (bis) – Logiciels anglo-saxons

Puisque V = B + W, on peut reformuler le problème à résoudre de la manière suivante :

$\max_a \frac{a'Ba}{a'Wa}$ Qui est équivalent à $\max_a a'Ba$ Sous la contrainte $a'Wa = 1$ *(Le vecteur « a » est normé)*

Les axes factoriels de l'AFD sont définis par les valeurs et vecteurs propres de la matrice $W^{-1}B$.

➡ Les vecteurs propres « a » de $W^{-1}B$ sont identiques à ceux de $V^{-1}B$ → les axes factoriels sont définis de la même manière.

Les valeurs propres sont reliées par la relation suivante : ρ = SCE / SCR pour l'axe associé

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

Ex. Fichier « Vins de Bordeaux »

Avec les variables « Température » et « Soleil » uniquement

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

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

Ex. le 1er axe explique 98% de l'écartement entre les barycentres dans l'espace initial : 98% = 2.6432 / (2.6432 + 0.0534).

Les 2 premiers axes expliquent 100% de cet écartement.

➔ Clairement, le 1er axe suffit largement ici !!!

On peut aussi exprimer les axes en termes de « pouvoir discriminatoire relatif »

Analyse des résultats
Choix du nombre d'axes
Interprétation des axes

Analyse discriminante descriptive – Choisir le nombre d'axes adéquat

On veut tester

H0 : les « q » derniers rapports de corrélation sont tous nuls

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

$\Leftrightarrow$ H0 : on peut négliger les « q » derniers axes

N.B. Tester individuellement un axe intermédiaire n'a pas de sens (ex. essayer de retirer le 1er axe tout en conservant le 2nd). Parce qu'ils ont un pouvoir discriminatoire décroissant ; et parce que l'explication fournie par un axe dépend du pouvoir discriminatoire des précédents.

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

Plus la statistique prend une valeur faible, plus intéressants sont les axes factoriels.

Dans le cadre de populations gaussiennes [c.-à-d. X = (X1,…,XJ) suit une loi multi normale dans chaque sous-groupe], on peut utiliser les transformations de Bartlett (loi du KHI-2) et de Rao (loi de Fisher)

Tanagra

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

SAS

Valeurs propres de Inv(E)*H = CanRsq/(1-CanRsq) — Test de H0 : les corrélations canoniques de la ligne en cours et suivantes sont égales à zéro

	Valeur propre	Différence	Proportion	Cumulé	Rapport de vraisemblance	Valeur de F approchée	DDL Num.	DDL Res.	Pr > F
1	2.6432	2.5898	0.9802	0.9802	0.26056848	14.39	4	60	<.0001
2	0.0534		0.0198	1.0000	0.94930704	1.66	1	31	0.2070

On ne peut pas retirer les deux premiers axes à 5% ; le second seul en revanche n'est pas significatif.

Analyse discriminante descriptive – Tester tous les axes

Un test particulier

H0 : les rapports de corrélation sont tous nuls

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

$\Leftrightarrow$ H0 : il est impossible de discerner les groupes dans l'espace de représentation

Test MANOVA c.-à-d. test d'égalité des moyennes multidimensionnelles

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

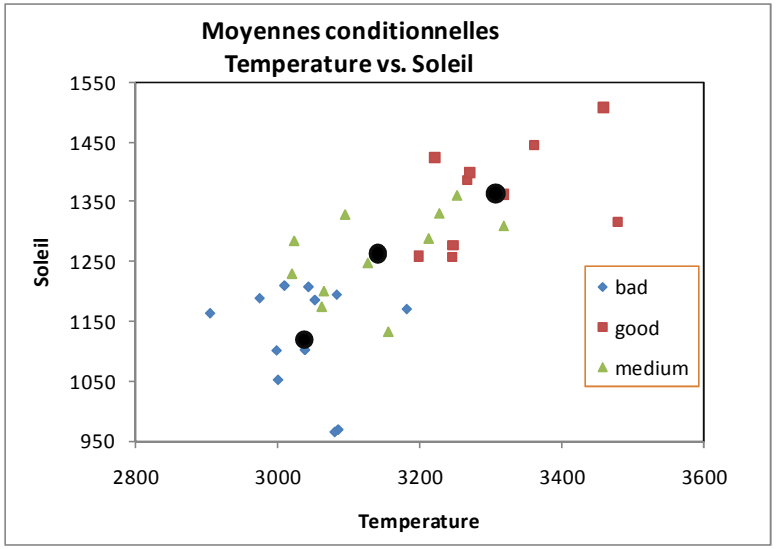
Simultanément

Statistique de test :

Le LAMBDA de Wilks

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

Plus le LAMBDA est petit, plus les (moyennes des) groupes sont écartés dans l'espace de représentation ($0 \leq \Lambda \leq 1$).



LAMBDA de Wilks = 0.26

Transformation de Bartlett

KHI-2 = 41.02 ; p-value < 0.0001

Transformation de Rao

F = 14.39 ; p-value < 0.0001

Conclusion : Une des moyennes (barycentres) conditionnelles au moins s'écarte des autres

Analyse discriminante descriptive – Interprétation des axes

Coefficients canoniques bruts et normalisés (intra-classes)

Coefficients bruts

Pour un axe, les vecteurs propres permettent de définir les coefficients de projection c.-à-d. ils permettent de calculer le score (coordonnée) des individus sur l'axe.

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

Coefficients non interprétables (comparables) parce variables non définies dans les mêmes unités.

Coefficients normalisés (standardisés)

On réalise l'AFD sur les variables centrées et réduites avec l'écart type intra-classes.

On peut avoir directement le résultat en multipliant le coefficient par l'écart-type intra-classes de la variable.

Les valeurs sont comparables d'une variable à l'autre

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

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

Est la variance intra-classes de la variable Xj

Indique la contribution des variables pour la discrimination sur l'axe

Attention, il s'agit de contributions partielles, tenant compte des autres variables

Deux variables peuvent se « gêner » si elles sont corrélées, partageant leurs contributions. Au point de prendre parfois des signes opposés (*cf. W.R. Klecka, « Discriminant Analysis », 1980 ; page 33*).

On préfèrera les structures canoniques (corrélation des variables avec les axes) pour interpréter les axes

Qualité = AFD (Température, Soleil) >>

Canonical Discriminant Function

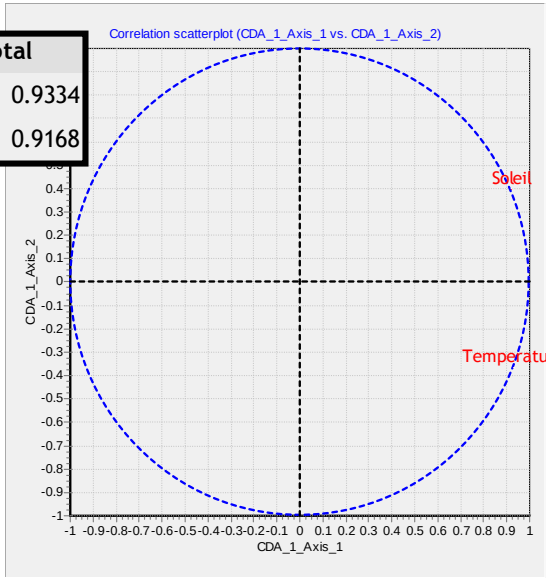
Coefficients	Unstandardized		Standardized	
Attribute	Root n°1	Root n°2	Root n°1	Root n°2
Temperature	0.007465	-0.009214	-0.653736	-0.806832
Sun	0.007479	0.010459	-0.604002	0.844707
constant	32.903185	16.049255	-	

Analyse discriminante descriptive – Interprétation des axes

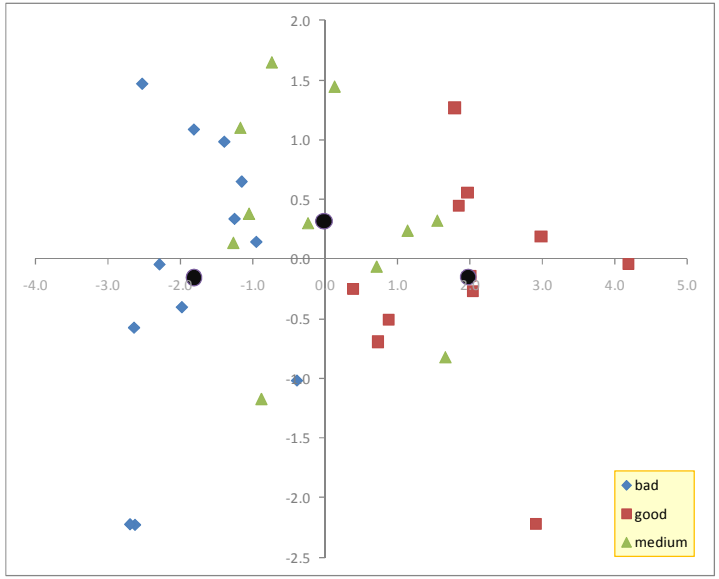
Structure canonique totale

Corrélation brute entre les variables et les axes factoriels, sans tenir compte de la structuration en classes. On peut même produire un « cercle des corrélations » comme en ACP

Descriptors	Total
Temperature	0.9334
Sun	0.9168



Le 1er axe correspond à la conjonction des journées chaudes (température) et des durées d'ensoleillement élevées



Les années où il a fait chaud avec beaucoup de soleil, le millésime a été bon.

Permet d'interpréter simplement les axes, caractériser l'importance des variables. A privilégier.

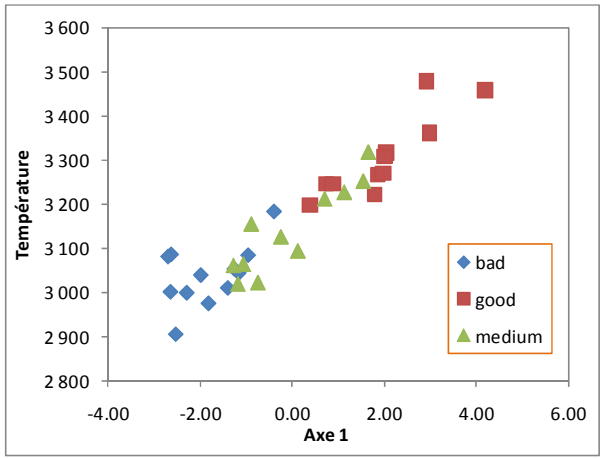
A comparer avec les coefficients standardisés, si signes différents ➔ colinéarité entre les variables

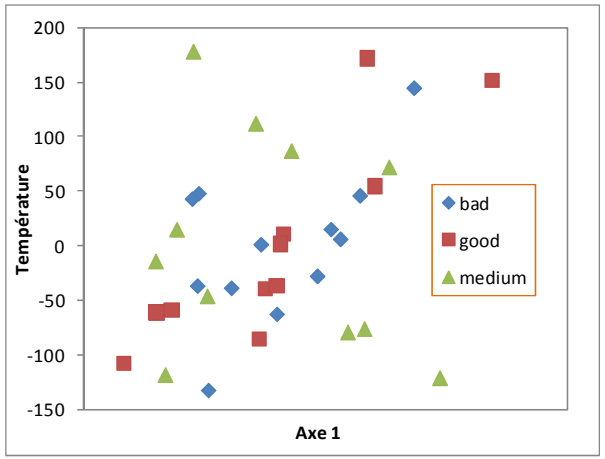
Analyse discriminante descriptive – Interprétation des axes

Structure canonique intra-classes

Corrélation après centrage des variables et des axes sur les sous-populations. Permet de caractériser le lien après avoir annihilé la structuration en classes.

Bof.





$r = 0.9334$

$r_w = 0.8134$

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

Souvent plus faible que la corrélation brute (pas toujours).

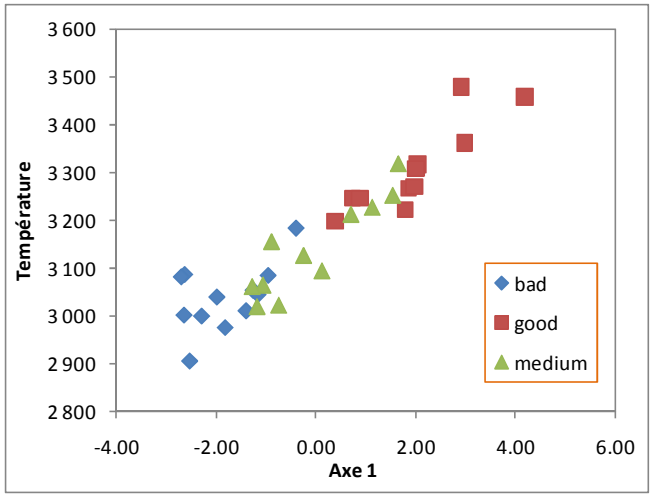
Permet de comprendre l'orientation des sous-nuages.

Analyse discriminante descriptive – Interprétation des axes

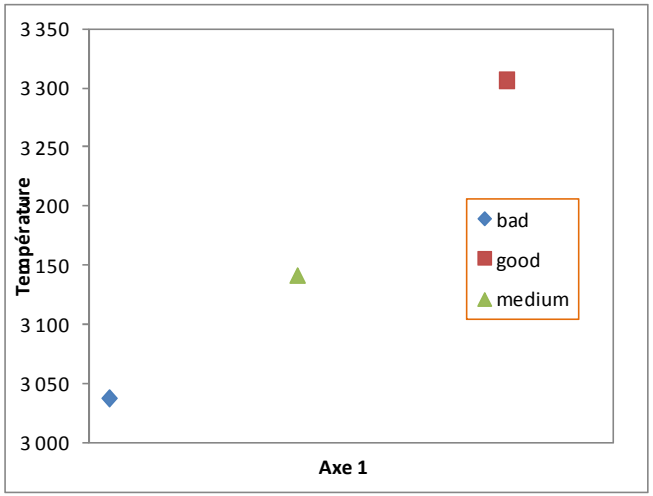
Structure canonique inter-classes

Corrélation après réduction des variables et des axes aux moyennes conditionnelles. Permet d'exalter la structuration en classes.

Séduisant en théorie. Difficile à lire en pratique (ex. « 1 » ou « -1 » systématiquement pour K = 2).



$r = 0.9334$



$r_B = 0.9949$

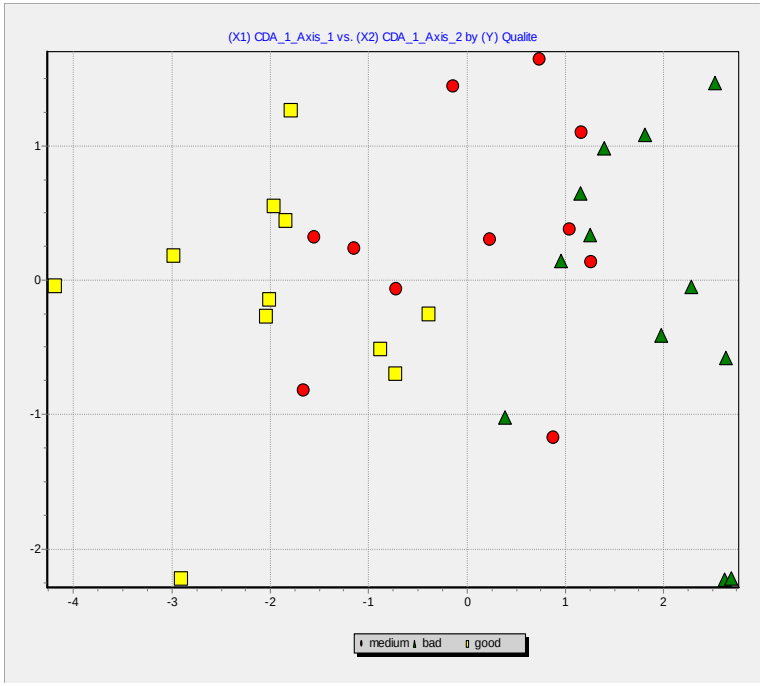
Root		Root n°1	
Descriptors	Total	Within	Between
Temperature	0.9334	0.8134	0.9949
Sun	0.9168	0.777	0.9934

Analyse discriminante descriptive – Interprétation des axes

Moyenne conditionnelles sur les axes

Calculer les moyennes conditionnelles sur les axes.

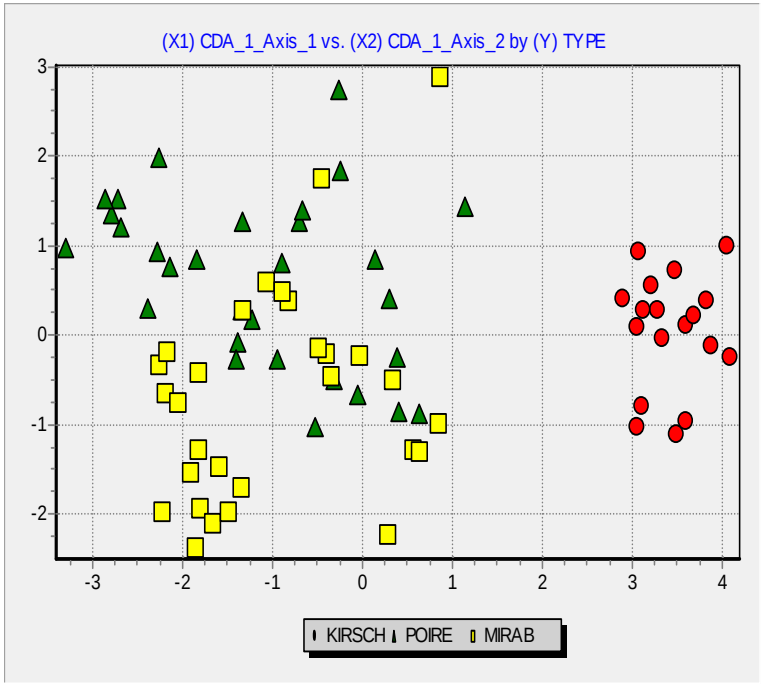
Permet de comprendre les groupes bien discriminés sur un axe.



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

Les 3 groupes bien discriminés sur le 1er axe

Rien d'intéressant sur le 2nd axe (rapport de corrélation très faible)



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

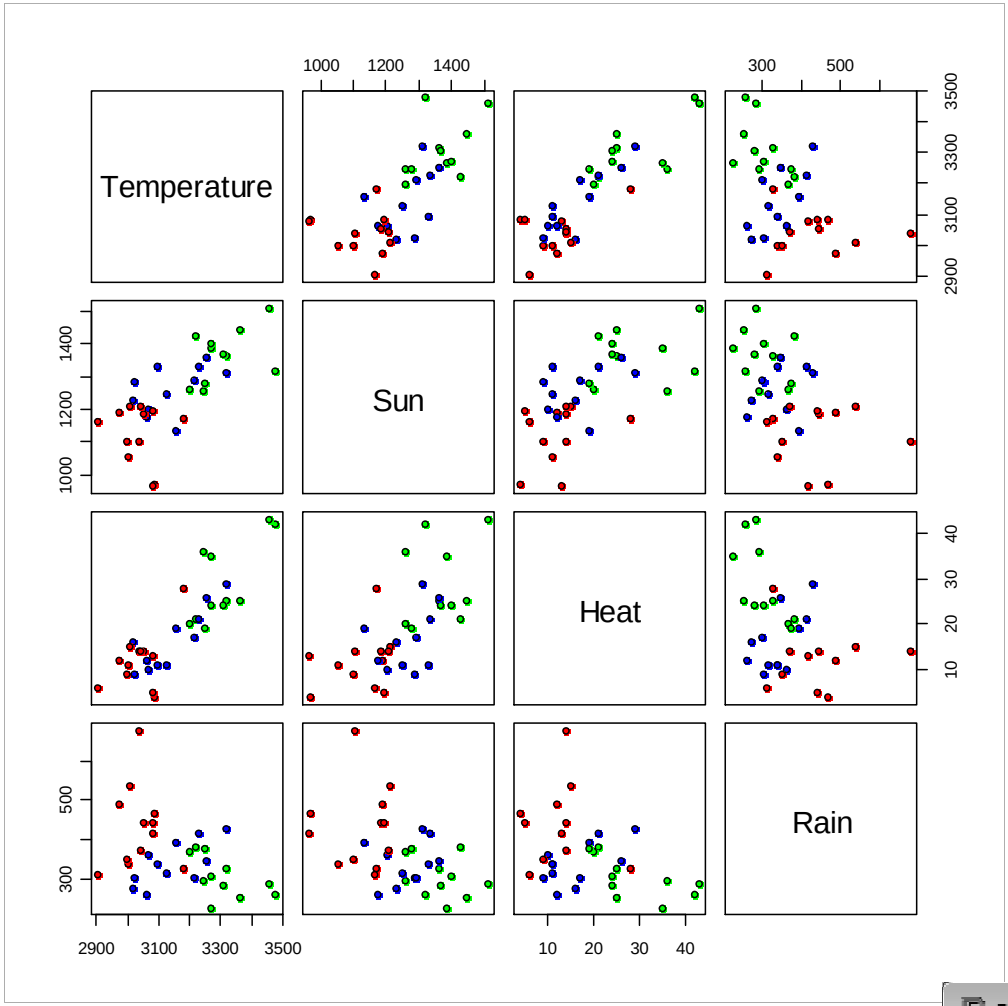
KIRSCH vs. les deux autres sur le 1er axe

POIRE vs. MIRAB sur le 2nd (rapport de corrélation reste élevé)

Etude de cas

Les « vins de Bordeaux » (Tenenhaus, 2007 ; page 353)

Vins de Bordeaux - Description rapide des données



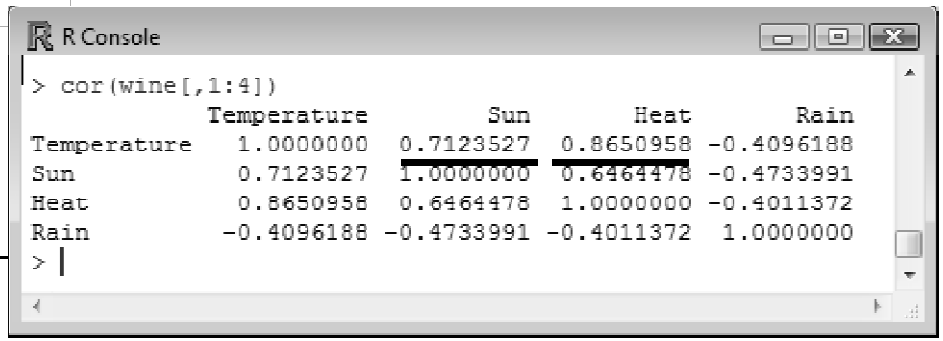
Certaines variables sont assez fortement corrélées (cf. matrice des corrélations)

(Rouge : Bad ; bleu : Medium ; vert : Good). On distingue les groupes, surtout sur certaines variables.

L'influence sur la qualité n'est pas la même selon (température, soleil, chaleur) et (pluie).

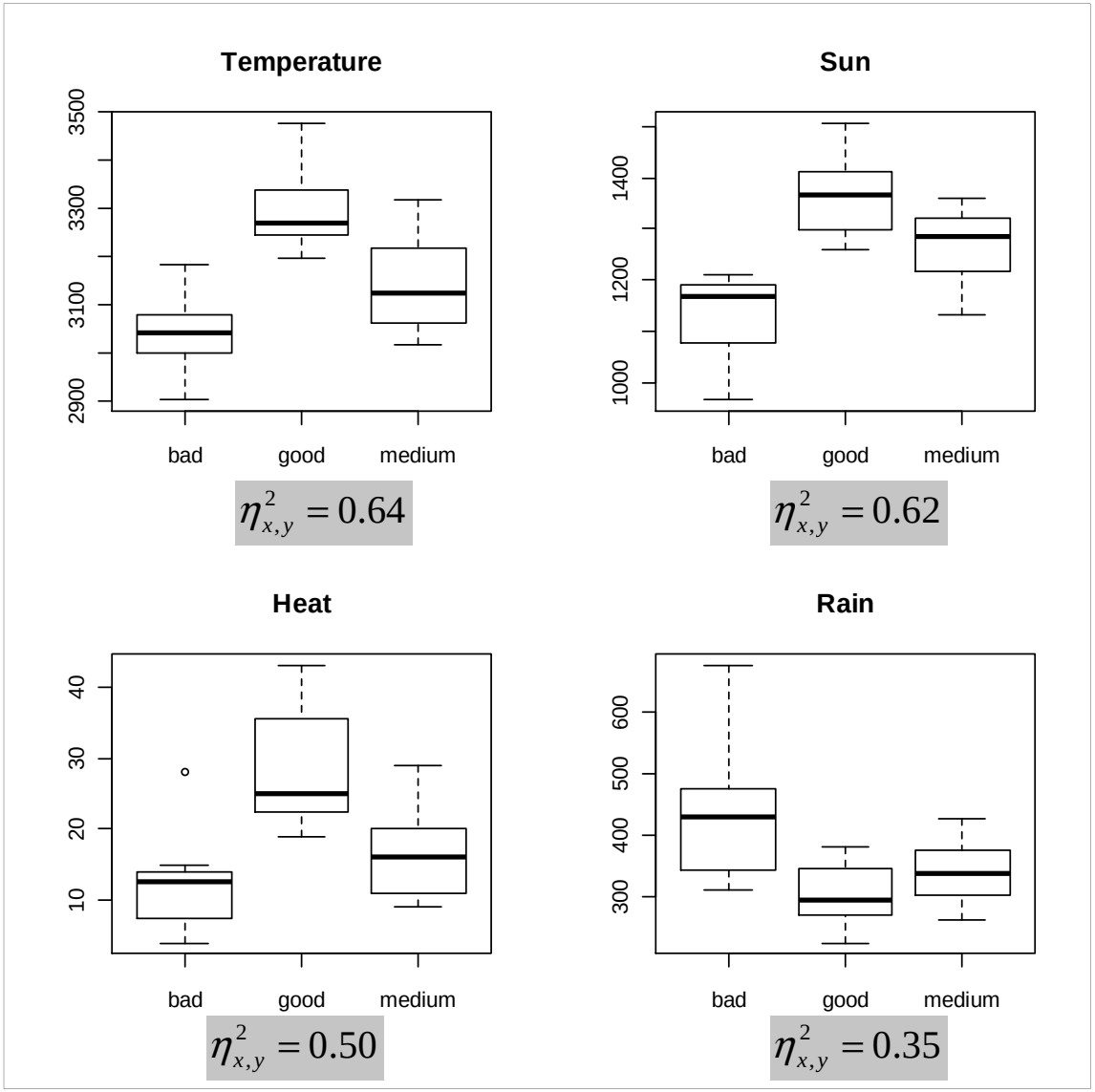
Il y a un point manifestement atypique, via la variable « Rain » (pluie)

Matrice des corrélations



Vins de Bordeaux – Discrimination selon les variables

Prises individuellement



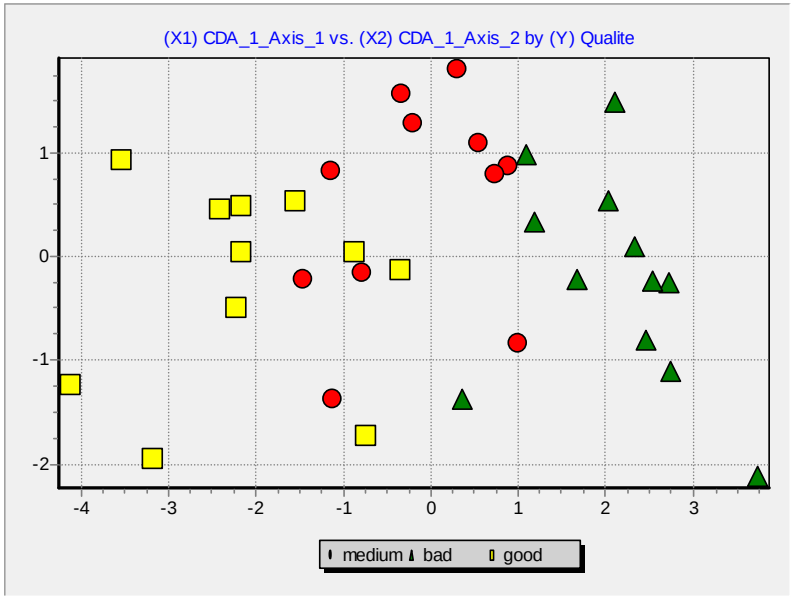
Température, Soleil et Chaleur permet déjà de distinguer les « bons » vins des « mauvais ».

Pour toutes les variables, l'ANOVA indique un écart significatif entre les moyennes à 5%.

Vins de Bordeaux – Résultats de l'AFD

Roots and Wilks' Lambda

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



Group centroids on the canonical variables

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

Le 2nd axe ne permet pas discerner les groupes, on peut la négliger.

96% du pouvoir discriminatoire de l'AFD est porté par le 1er axe. Nous allons y concentrer notre analyse.

Sur le 1er axe, nous distinguons les 3 groupes. Les moyennes conditionnelles nous indique (de gauche à droite) : les « bons » vins, les « moyens » et les « mauvais ».

Le rapport de corrélation sur l'axe est de 0.76; plus élevé que n'importe quelle variable prise individuellement (la meilleure était « température » avec 0.64).

Vins de Bordeaux – Caractérisation des groupes

Via la caractérisation des axes

Canonical Discriminant Function

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

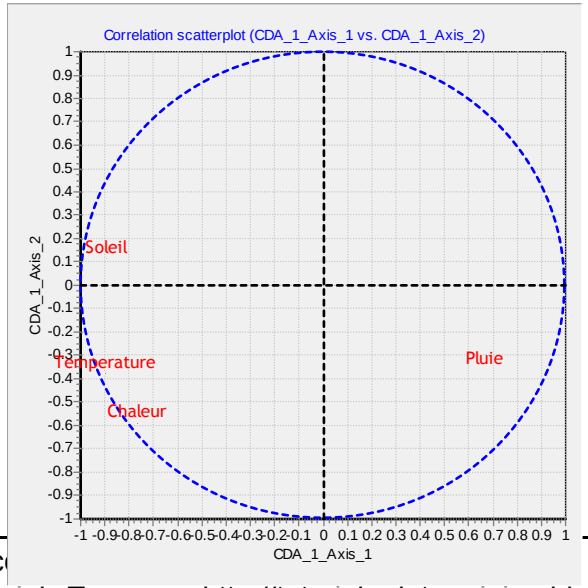
Factor Structure Matrix - Correlations

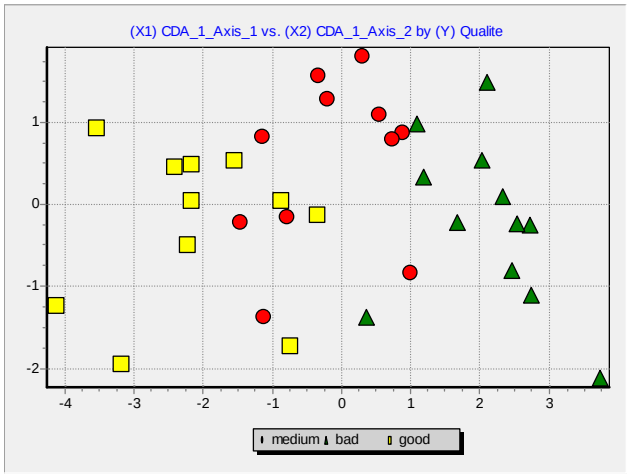
Root	Root n°1			Root n°2		
Descriptors	Total	Within	Between	Total	Within	Between
Temperature	-0.9006	-0.7242	-0.9865	-0.3748	-0.5843	-0.1636
Soleil	-0.8967	-0.7013	-0.9987	0.1162	0.1761	0.0516
Chaleur	-0.7705	-0.5254	-0.9565	-0.59	-0.7799	-0.2919
Pluie	0.6628	0.3982	0.9772	-0.3613	-0.4208	-0.2123

L'axe oppose la température et l'ensoleillement (plus ils sont élevés, meilleur sera le vin) et la pluie (plus il pleut, mauvais sera le vin ; l'impact est moindre que pour les autres variables).

Attention, le rôle de « chaleur » est ambigu. En réalité, il est fortement corrélé avec « température » (cf. la matrice des corrélations).

« Chaleur » influe positivement sur la qualité mais, par rapport à « température », son apport d'information additionnel dans l'explication de la qualité est négligeable. Le rapport de corrélation conditionnel (cf. Tenenhaus, page 376) est $\eta^2_{x_3, y/x_1} = 0.0348$





Coordonnées des individus + Groupe d'appartenance. Cercle des corrélations.

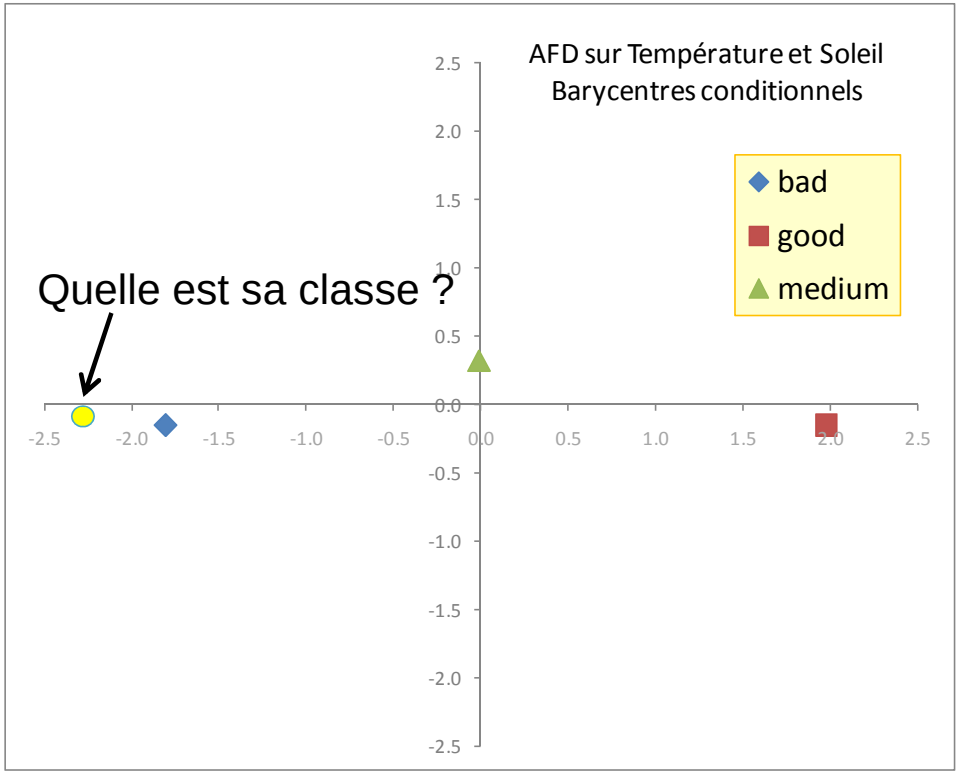
Classement des nouveaux individus

S'appuyer sur l'AFD pour prédire le groupe d'appartenance

Préambule

L'analyse discriminante linéaire (prédictive) propose un cadre théorique plus séduisant pour la prédiction. Notamment en explicitant les hypothèses probabilistes.

Néanmoins, on peut s'appuyer sur les résultats de l'AFD pour classer les individus, en s'appuyant sur des règles géométriques.



Etapes :

1. A partir des valeurs de X, et des coefficients canoniques bruts : calculer la position de l'individu dans le repère factoriel.
2. Calculer les distances aux barycentres dans ce repère (distance euclidienne simple)
3. On attribuera à l'individu la classe dont le barycentre est le plus proche

Classement de nouveaux individus – AFD sur Température (X1) et Soleil (X2)

X1 = 3000 – X2 = 1100 – Millésime 1958 (Basé sur les prévisions météo)

1. Calcul des coordonnées factorielles

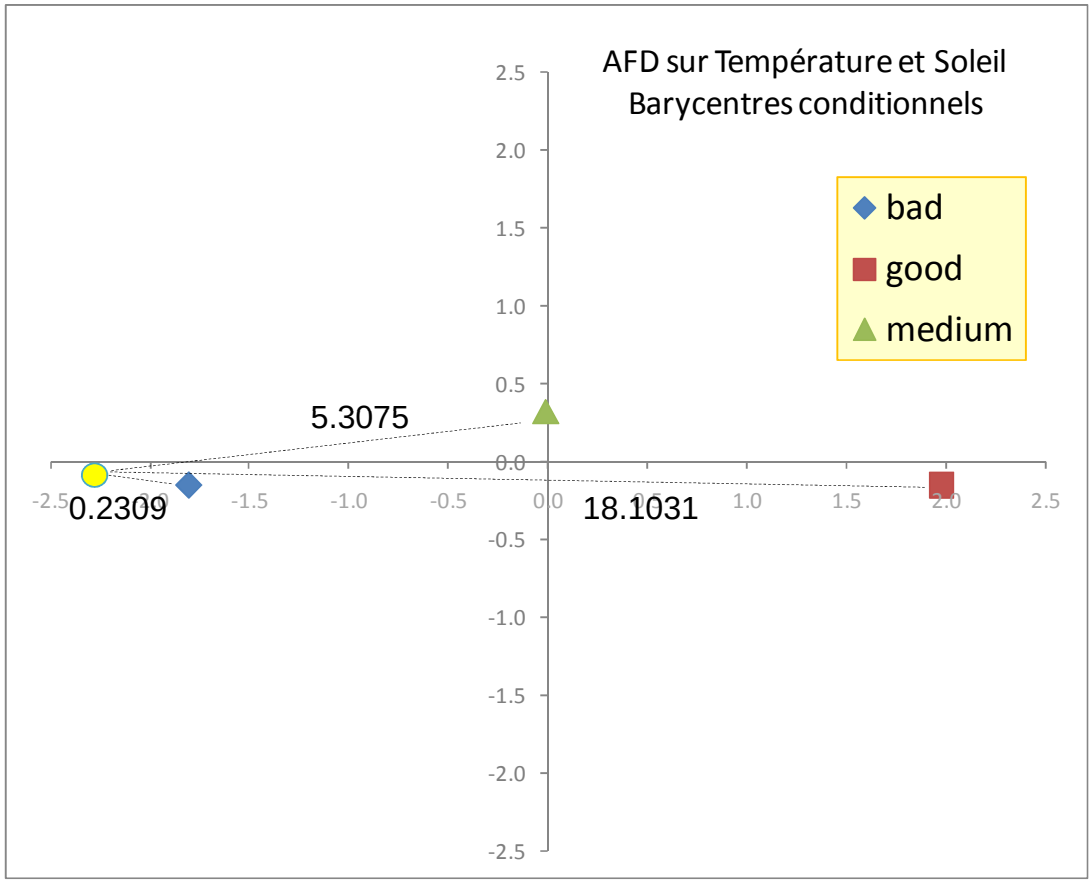
$$
\begin{aligned}
z_1 &= 0.007457 \times x_1 + 0.007471 \times x_2 - 32.868122 \\
&= 0.007457 \times 3000 + 0.007471 \times 1100 - 32.868122 \\
&= -2.2780 \\
z_2 &= -0.009204 \times x_1 + 0.010448 \times x_2 + 16.032152 \\
&= -0.009204 \times 3000 + 0.010448 \times 1100 + 16.032152 \\
&= -0.0862
\end{aligned}
$$

2. Calcul des distances par rapport aux barycentres

$$
\begin{aligned}
d^2(bad) &= (-2.2780 - (-1.8023))^2 + (-0.0832 - (-0.1538))^2 \\
&= 0.2309 \\
d^2(good) &= 18.1031 \\
d^2(medium) &= 5.3075
\end{aligned}
$$

3. Conclusion

Le millésime 1958 a de fortes chances d'être « bad ». Il a très peu de chances d'être « good ».



Classement de nouveaux individus

Distance euclidienne dans l'espace factoriel = distance de Mahalanobis dans l'espace initial

On peut calculer la même distance que précédemment dans l'espace initial en utilisant la métrique W^{-1} : on parle de distance de MAHALANOBIS.

Pour le même individu que précédemment, on calcule sa distance par rapport au barycentre de « bad » avec….

$$
\begin{aligned}
d^2(bad) &= (x-\mu_{bad})' W^{-1} (x-\mu_{bad}) \\
&= (3000-3037.3;1100-1126.4) \begin{pmatrix} 7668.46 & 1880.15 \\ 1880.15 & 6522.33 \end{pmatrix}^{-1} \begin{pmatrix} 3000-3037.3 \\ 1100-1126.4 \end{pmatrix} \\
&= (-37.33 \quad -26.42) \begin{pmatrix} 0.000140 & -0.000040 \\ -0.000040 & 0.000165 \end{pmatrix} \begin{pmatrix} -37.33 \\ -26.42 \end{pmatrix} \\
&= 0.2309
\end{aligned}
$$

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

Est la matrice de variance covariance intra-classes [multiplié par les degrés de libertés (n-K)]

Pourquoi s'enquiquiner à passer par une AFD alors ?

1. On dispose d'une explication du classement, le vin est mauvais parce que faible température et pas de soleil
2. On peut ne utiliser que les axes significatifs pour le classement (*uniquement le 1er axe pour notre exemple*) : on introduit une forme de régularisation (« *reduced rank LDA* ») (Hastie et al., 2001)

Classement de nouveaux individus

Produire une fonction de classement explicite

Pour un individu « i » à classer, on calcule sa distance euclidienne par rapport au barycentre de la classe « k » dans l'espace défini par les Q axes factoriels (Q = M si on prend tous les axes)

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

Minimisation par rapport à « k », donc on peut supprimer tout ce qui n'en dépend pas… et on multiplie par -0.5 → on passe à une maximisation

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

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

Fonction canonique pour le facteur « m »

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

→ La fonction de classement est linéaire !!!

Ex. Vins de Bordeaux avec Température (x1) et Soleil (x2) – Un seul axe factoriel (Q = 1)

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

Pour l'individu (x1 = 3000; x2 = 1100)

$$\begin{aligned} f(bad) &= 2.4815 \\ f(good) &= -6.5447 \\ f(medium) &= 0.0230 \end{aligned}$$

Conclusion : le millésime 1958 sera un « bad »

Classement de nouveaux individus

Quel rapport avec l'analyse discriminante linéaire [ADL] (prédictive) ?

L'analyse discriminante linéaire prédictive fait l'hypothèse de multi-normalité des distributions conditionnelles des X et d'homoscédasticité

http://fr.wikipedia.org/wiki/Analyse_discriminante_lin%C3%A9aire

Fonction de classement de l'ADL

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

Règle de classement issu de l'AFD où l'on prend tous les M axes factoriels

Bref, la règle d'affectation de l'AFD équivaut à celle de l'ADL avec l'hypothèse que les classes sont équiprobables c.-à-d.

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

Equivalence

Certains logiciels font par défaut cette hypothèse, même pour l'ADL (ex. PROC DISCRIM de SAS)

Introduire la correction avec les probabilités estimées sur les données permet de réduire le taux d'erreur (cf. Hastie et al., 2001 ; page 95)

Utilisation des logiciels

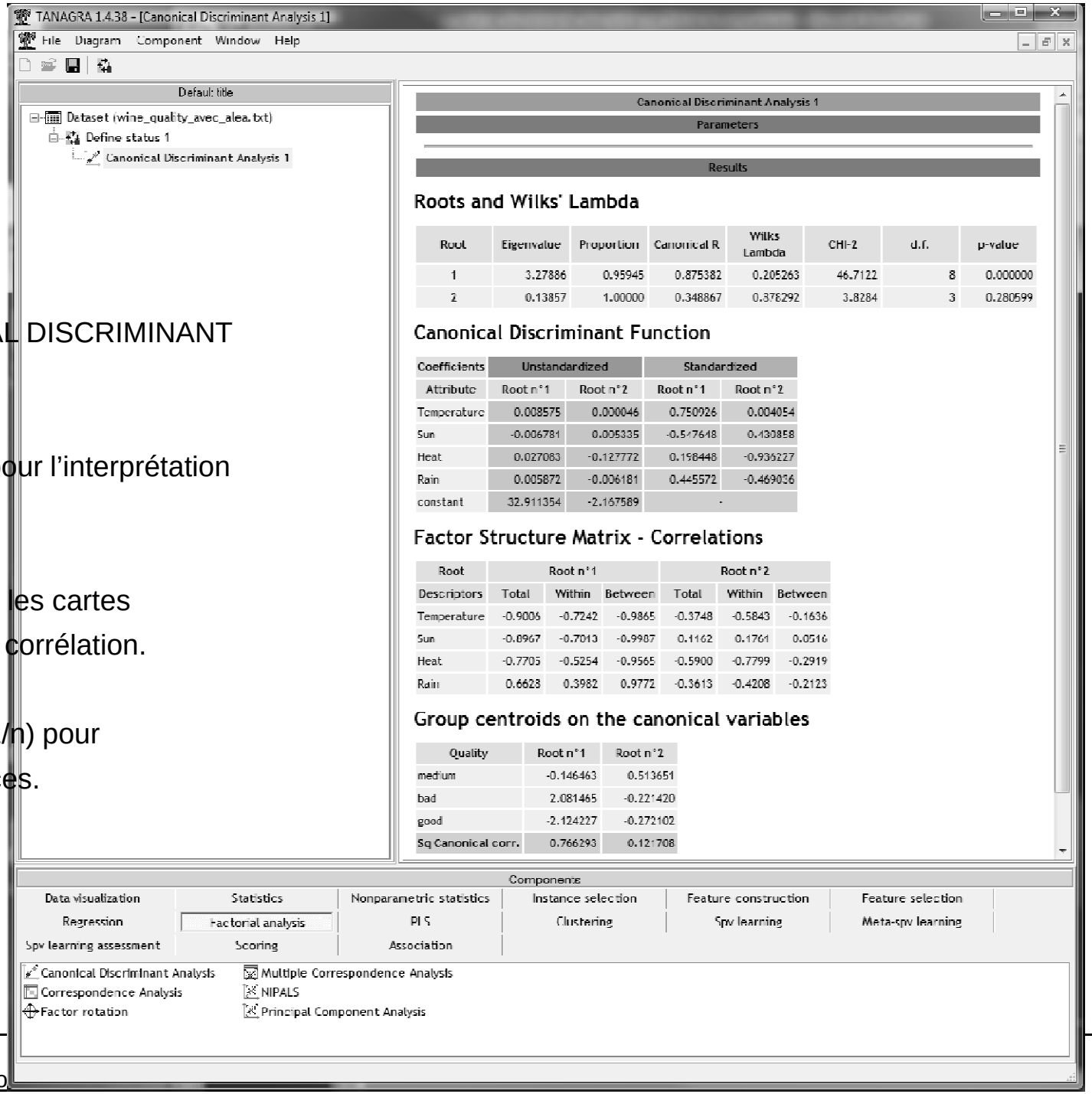
Tanagra , R et SAS

Composant « CANONICAL DISCRIMINANT ANALYSIS »

Les résultats importants pour l'interprétation sont disponibles.

Il est possible de produire les cartes factorielles et le cercle de corrélation.

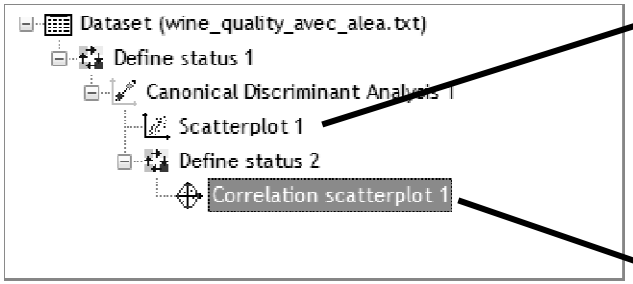
Réf. Françaises. Utilise (1/n) pour l'estimation des covariances.

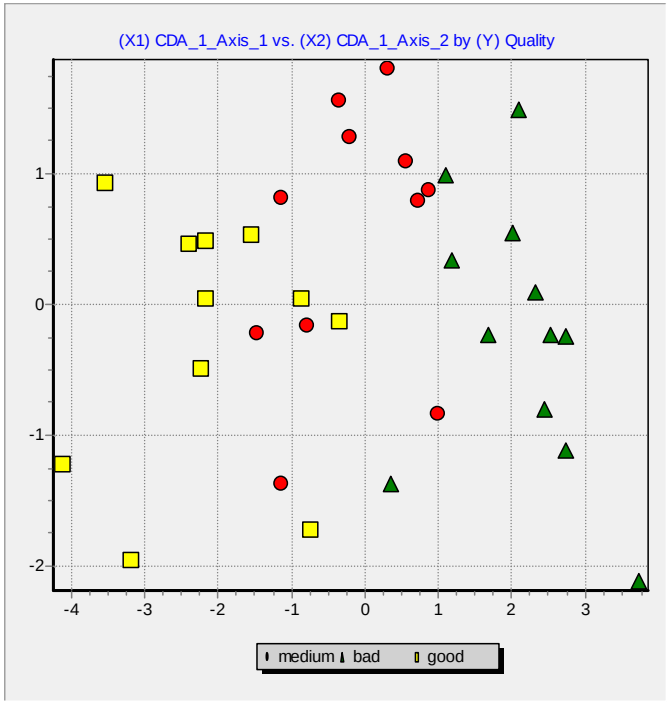


AFD avec TANAGRA

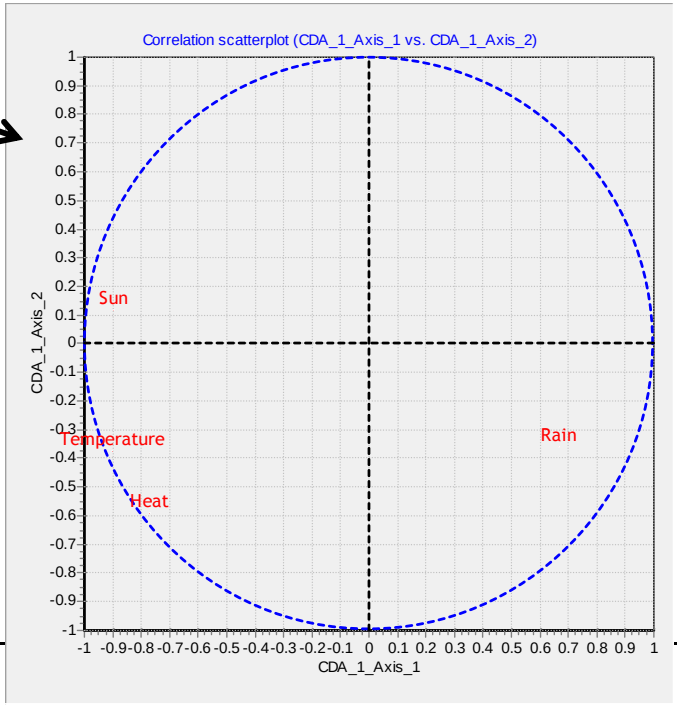
Représentations graphiques

Projection dans le 1er plan factoriel





Cercle des corrélations



AFD avec R

La fonction « lda » du package MASS

```
rm (list=ls())

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

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

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

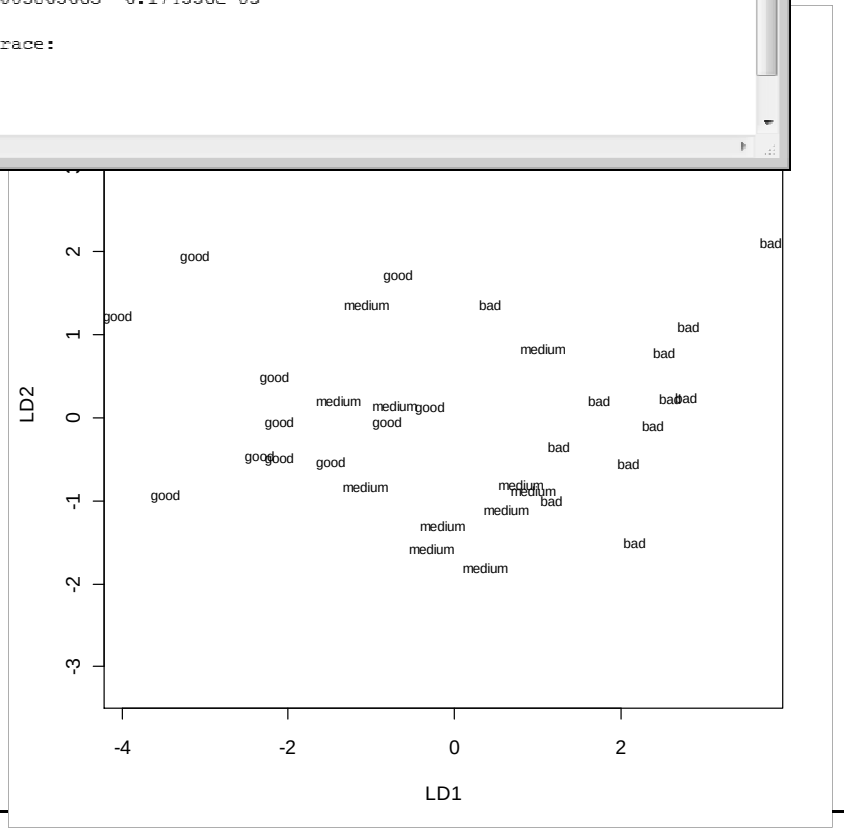
```
R Console
> print(wine.lda)
Call:
lda(Quality ~ ., data = wine)

Prior probabilities of groups:
      bad      good    medium
0.3529412 0.3235294 0.3235294

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

Coefficients of linear discriminants:
                    LD1           LD2
Temperature -0.008566046 -4.625059e-05
Sun         -0.006773869 -5.329293e-03
Heat         0.027054492  1.276362e-01
Rain         0.005865665  6.171556e-03

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



Résultats initiaux un peu succincts.

Mais en programmant un peu, on peut obtenir tout ce que l'on veut !!!
C'est l'avantage de R.

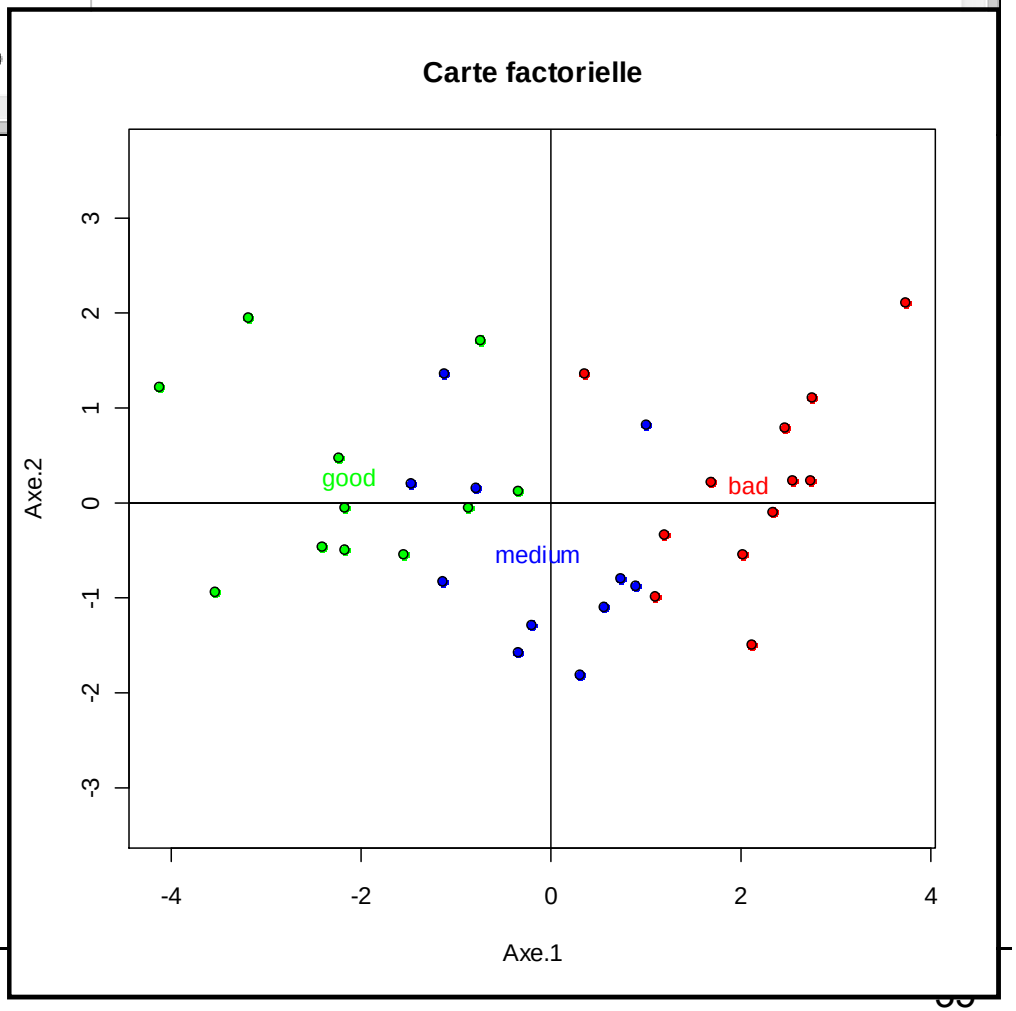
Anglo-saxon. Utilise [1/(n-1)] pour l'estimation des covariances.

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

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

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

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



AFD avec R

Un peu plus loin avec R

Un peu de programmation et le résultat en vaut la peine…

AFD avec SAS
La procédure CANDISC

wine sas code.sas

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

Résultats très complets.

Avec l'option « ALL » on peut obtenir tous les résultats intermédiaires (matrices V, W, B ; etc.).

Anglo-saxon. Utilise [1/(n-1)] pour l'estimation des covariances (comme R).

SAS

Fichier Édition Affichage Outils Solutions Fenêtre Aide

Résultats

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

Résultats | Explorateur

Sortie - (Sans titre)

Le Système SAS 06:38 F

The CANDISC Procedure

Coefficients canoniques normalisés sur la totalité de l'échantillon

Variable	Libellé	Can1	Can2
Temperature	Temperature	1.209391427	-0.006529859
Sun	Sun	0.857727422	-0.674810955
Heat	Heat	-0.270999045	1.278475787
Rain	Rain	-0.536131215	0.564364371

Coefficients canoniques normalisés et groupés intra-classe

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

Coefficients canoniques bruts

Variable	Libellé	Can1	Can2
Temperature	Temperature	0.0085660155	-.0000462506
Sun	Sun	0.0067738690	-.0053292933
Heat	Heat	-.0270544919	0.1276361644
Rain	Rain	-.0058656650	0.0061745562

Moyennes de classes sur les variables canoniques

Quality	Can1	Can2
bad	-2.079247035	0.221183875
good	2.121902956	0.271012010
medium	0.146306536	-0.513103518

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

C:\Users\Maison

Conclusion

AFD : méthode de description de groupes

Outils pour l'évaluation et l'interprétation des résultats (significativité des axes, coefficients canoniques standardisés, structures canoniques…)

Outils de visualisation (projections dans lesplans factoriels, cercle des corrélations)

Connexions avec d'autres techniques factorielles (ACP, Analyse Canonique)
(cf. SAPORTA, 2006 ; pages 444 et 445)

Connexions avec les méthodes prédictives, en particulier l'Analyse discriminante linéaire prédictive

Permet de fournir une explication aux prédictions

Bibliographie

L. Lebart, A. Morineau, M. Piron, « Statistique exploratoire multidimensionnelle », Dunod, 2000.
Chapitre 3, Section 3.3, pages 251 à 283.
Ouvrage de référence pour les calculs de TANAGRA (réf. formules dans le code source)

M. Tenenhaus, « Statistique – Méthodes pour décrire, expliquer et prévoir », Dunod, 2007.
Chapitre 10, pages 351 à 386.
Pratique, lecture des résultats. Sorties de SAS. Base de ce support de cours.

G. Saporta, « Probabilités, Analyse de données et Statistique », Technip, 2006.
Chapitre 18, pages 439 à 485.
Théorique et pratique, inclut l'AD pour prédicteurs qualitatifs.

Wikipédia, « Analyse discriminante »
http://fr.wikipedia.org/wiki/Analyse_discriminante
Avec le fameux exemple des IRIS.